Identifying characteristic scales in the human genome

P. Carpena,[1] P. Bernaola-Galván,[1] A. V. Coronado,[1] M. Hackenberg,[2] and J. L. Oliver[2]
[1]*Departamento de Física Aplicada II, Universidad de Málaga, 29071 Málaga, Spain*
[2]*Departamento de Genética, Universidad de Granada, 18071 Granada, Spain*
(Received 1 June 2006; published 16 March 2007)

The scale-free, long-range correlations detected in DNA sequences contrast with characteristic lengths of genomic elements, being particularly incompatible with the isochores (long, homogeneous DNA segments). By computing the local behavior of the scaling exponent α of detrended fluctuation analysis (DFA), we discriminate between sequences with and without true scaling, and we find that no single scaling exists in the human genome. Instead, human chromosomes show a common compositional structure with two characteristic scales, the large one corresponding to the isochores and the other to small and medium scale genomic elements.

DOI: 10.1103/PhysRevE.75.032903 PACS number(s): 87.15.Cc, 87.14.Gg, 87.15.Aa

The long-range, power-law fractal correlations detected in human DNA sequences [1] imply that compositional segments should appear at all scales, showing a power-law distribution of sizes [2]. This view contrasts with the well-known characteristic lengths shown by most genomic elements (genes, exons, introns, transposable elements or TEs, etc.), being also particularly incompatible with the view of the genome as a mosaic of long homogeneous segments or "isochores" [3,4]. The strong controversy generated by these conflicting views has challenged the existence of either the correlations themselves [5–8] or the isochore genome structure [9]. By developing a method able to disclose the different scales potentially built into the genome, we show here that DNA correlations are much more complex than power-laws with a single scaling exponent: actually the exponents of such "power laws" are different for different scales, thus not existing a clear scaling at all. The deviations from uniform power laws are known to imply heterogeneities of several characteristic sizes [10], or (less likely in DNA) the existence of segments with different type of correlations [12]. We find two characteristic scales in human chromosomes: one corresponding to the size of isochores and the other to the small and medium scale genome elements.

To detect long-range correlations in DNA we use DFA, a scaling analysis method that can deal with seemingly nonstationary time series [11,12], and which provides a simple quantitative parameter—the scaling exponent α—to represent the correlation properties of a signal (see Ref. [11] for details). DFA provides a relationship between the root mean square fluctuation $F(l)$ and the window size l. True scaling appears when

$$F(l) \sim l^{\alpha}. \tag{1}$$

Thus, α can be obtained by fitting $F(l)$ vs l to a straight line in a log-log plot, the slope being α. If $\alpha=0.5$, there is no correlation and the signal behaves as a random series (white noise), $\alpha<0.5$ indicates anticorrelations, and $\alpha>0.5$ positive correlations.

Usually, one gets a good correlation coefficient in the fitting since $F(l)$ can be very smooth [13]. However, this procedure can mask information present in the signal. In Fig. 1 we show an example of how $F(l)$ seems to scale properly as in Eq. (1) for both a real scale-free fractal sequence and a large contig human sequence which shows clearly nonfractal behavior. Thus, to better discriminate between both types of signals, following the work by Viswanathan *et al.* [10] we propose to obtain α not as the slope of a linear fitting of $\log_{10} F(l)$ vs $\log_{10} l$ but as

$$\alpha(l) = \frac{d \log_{10}[F(l)]}{d \log_{10}(l)}, \tag{2}$$

where we write explicitly $\alpha(l)$ since α is not necessarily constant. $\alpha(l)$ is then the local α value at scale l. Indeed, when α is constant, the signal fluctuates in a similar way at

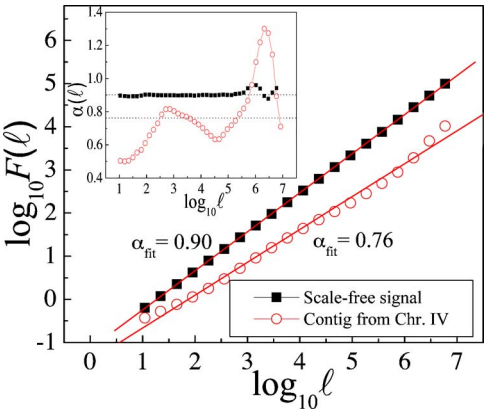

FIG. 1. (Color online) The fluctuation function $F(l)$ vs l in a log-log plot obtained for two time series. The solid lines correspond to linear fits (both with $R>0.99$), with scaling exponents α_{fit} 0.90 and 0.76, respectively. The signal shown in (■) was generated using a standard method to create long-range correlated time series [14] by imposing that $\alpha=0.90$, and the DFA recovers correctly this scaling exponent. However, (○) corresponds to a large human DNA sequence from chromosome (chr.) IV mapped into a binary sequence using the SW (strong-weak) mapping rule: C or $G\rightarrow 1$, A or $T\rightarrow 0$. In the inset we represent $\alpha(l)$ obtained using Eq. (2) for the two cases. The signal with real scaling presents a practically constant value of $\alpha(l)$ with small fluctuations around the linear fitting value—$\alpha_{\text{fit}}=0.90$ (dotted line). However, in human DNA (○), $\alpha(l)$ is far from being a constant value: the fluctuations around the value obtained in the linear fitting ($\alpha_{\text{fit}}=0.76$, dotted line) are not only very large but also following a particular pattern, indicating that this α_{fit} value is meaningless, and that there is no real scaling in spite of the good linear fitting.

 ©2007 The American Physical Society

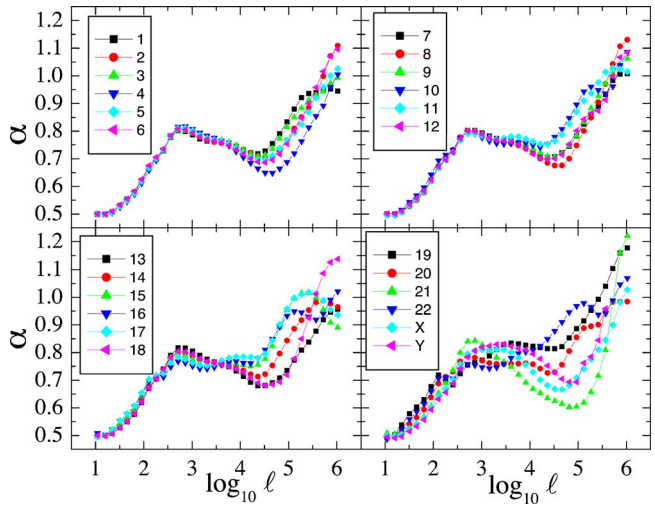

FIG. 2. (Color online) The function $\alpha(l)$ vs the scale l [Eq. (2)] for the 24 human chromosome sequences from the National Center for Biotechnology Information (NCBI) build 35 human genome assembly. In every chromosome, we considered all the possible contigs of $N=11$ Mb, and calculated $\alpha(l)$ in each one of them, reaching a maximum scale $l_{\max}$ in the DFA procedure of $l_{\max}=N/10=1.1$ Mb (except in chromosome Y, where we considered a smaller $l_{\max}$ since no contigs of $N=11$ Mb exist). Finally, we averaged all the $\alpha(l)$ values obtained from the individual contigs to obtain the corresponding $\alpha(l)$ function for the chromosome.

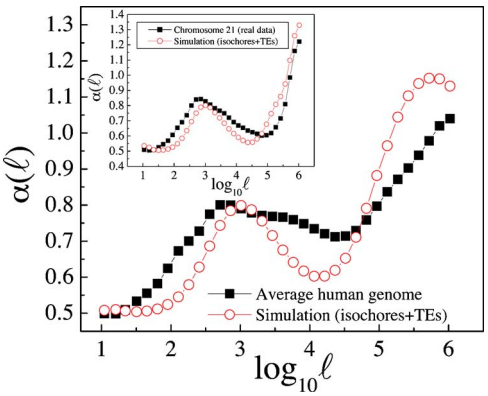

FIG. 3. (Color online) Behavior of $\alpha(l)$ obtained by a weighted average of the 24 chromosomes shown in Fig. 2 using as weights the relative length of each chromosome (■), and also obtained from simulations using a two-scales model (see text for details) shown in (□). Inset: the same for chromosome 21.

all scales, indicating self-similarity characterized by the single value of α—(■) in Fig. 1. However, α could change as a function of l, and one could not say that the signal presents scaling since $F(l)$ is not a power-law—as the human case shown in (○) in Fig. 1.

The variation of $\alpha(l)$ reveals useful to analyze the structure at different scales [10]. First, as shown above, $\alpha(l)$ is able to discriminate between apparently fractal long-range correlated signals, helping to distinguish real scaling behavior from nonfractal signals. And second, when the signal does not present real scaling, the behavior of $\alpha(l)$ is a powerful tool to detect crossovers between different regimes when changing the scale l at which the signal is studied: the pattern shown by $\alpha(l)$ indicates the characteristic scales present in the signal. Thus, the signal corresponding to human chromosome IV with the SW mapping rule (inset of Fig. 1) used from now on, exhibits clearly nonfractal behavior and suggests the presence of two main characteristic scales [the two major peaks in $\alpha(l)$] at intermediate and large l values. Similarly, a genome-wide analysis of the behavior of $\alpha(l)$ (Fig. 2) can help to elucidate whether the human genome presents good scaling properties and therefore fractal behavior (as widely thought), or, in contrast, if it has characteristic scales. Two main conclusions can be drawn: (i) human DNA cannot be considered in general as a fractal signal with a single type of scale invariance, since $\alpha(l)$ is neither a constant nor consists of small fluctuations around a constant, and therefore there are no good scaling properties of the type (1) and (ii) this pattern is shared by most of the chromosomes, thus reflecting the same characteristic scales in the whole genome. The structure of long-range correlations in the human genome has been studied previously [7] by using mainly Fourier analysis, finding crossovers at different scales which are consistent with our results presented here.

Figure 3 shows the "universal" profile of $\alpha(l)$ obtained by averaging the 24 curves shown in Fig. 2. At small scales ($\log_{10} l<1.5, l<30$ bp), human DNA behaves practically as a random sequence—$\alpha(l)\simeq 0.5$, indicating that there is no patterns of such small scales in the genome, and therefore no patchiness is detected in this range. Afterwards, $\alpha(l)$ starts to increase up to reach a maximum value of about $\alpha=0.8$ at $\log_{10} l\simeq 2.8, l\simeq 630$ bp, after which $\alpha(l)$ starts to decrease (except in chrs. 19 and 22). This behavior indicates patchiness at these intermediate scales, probably due to the presence of genomic entities of the appropriate size (see below). The slow decrease of $\alpha(l)$ after the maximum continues up to a scale of about $\log_{10} l\simeq 4.5, l\simeq 30$ kb, where a local minimum is found. This decrease indicates that the patches of intermediate scales are disappearing and the sequence looks more random, the extreme case being chr. 21, where $\alpha(l)\lesssim 0.6$ at $\log_{10} l\simeq 5$. Beyond $\log_{10} l\simeq 4.5$, $\alpha(l)$ increases noticeably again when increasing l, revealing the existence of large scale patchiness or heterogeneities (for which $\log_{10} l\simeq 6$) present in all the chromosomes. To our knowledge, these large-scale heterogeneities appearing with ubiquity throughout the genome can very likely [15,16] be due to the isochore structure, indicating the posibility of *in silico* isochore detection using $\alpha(l)$. Noteworthy, the possibility of a statistically rigorous detection of isochores, and even the proper isochore concept, has been recently questioned (see Cohen *et al.* in Ref. [9].) Nevertheless, isochores appears naturally via the properties of $\alpha(l)$.

Some chromosomes do not follow strictly this general picture: after reaching the local maximum at intermediate scales (for $\log_{10} l\simeq 3$), $\alpha(l)$ does not decrease clearly, but remains in a kind of plateau for about two orders of magnitude (chrs. 15, 16, 17, 19) or even continues increasing (chr. 22) prior to the final large-scale peak. Given the constancy of α along two orders of magnitude, these chromosomes are more fractal and scale-invariant than the others [5].

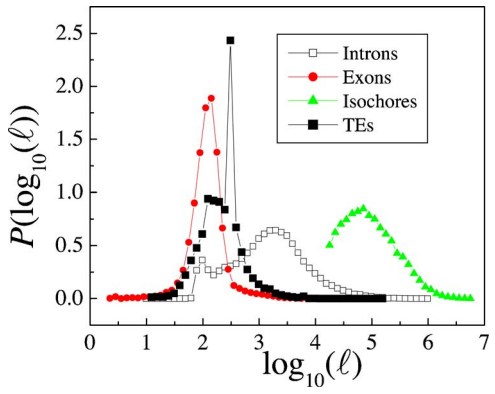

FIG. 4. (Color online) The probability distribution of the logarithms of the sizes of several genomic entities. The sizes of exons and introns are obtained from annotated genes with perfect alignment with the coding sequences (CDS) of their corresponding messenger RNA (mRNA) sequences from the RefSeq database [17]. The lengths of TEs were obtained using RepeatMasker [18]. The isochore sizes comes from the algorithm IsoFinder [4] for segments with $l > 20000$ bp.

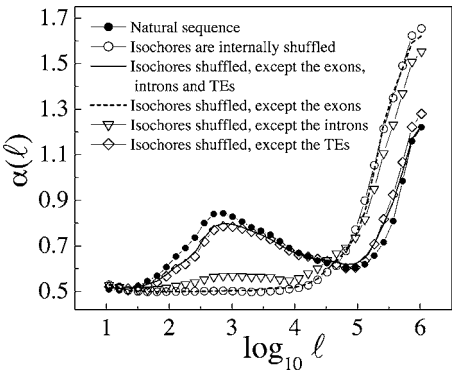

FIG. 5. Behavior of $\alpha(l)$ in different sequence-shuffling experiments carried out in human chromosome 21.

Figure 4 shows the probability distributions of the sizes of different genomic elements, which allow us to elucidate the origin of the intermediate and the large scale patchiness. All the distributions are essentially of log-normal nature, being quite symmetric around a large central peak. We consider that exons, introns and transposable elements (TEs), as well as other genomic entities of similar scale, are the source of heterogeneities in the intermediate scale (in the range 10^2-10^4), which coincide (see Fig. 2) with the first wide peak in the profile of $\alpha(l)$. In Ref. [6], the influence in the correlations of interspersed repeats is studied, finding that they affect mainly short-range correlations, in agreement with the behavior of TEs we report here.

In addition, the second peak at large scale coincides with the scales corresponding to the isochore structure, as revealed by the isochore size distribution, since no other genomic entity can produce heterogeneities of such a large scale. The isochore peak is well apart from the intermediate-scale elements, thus explaining the general decrease of $\alpha(l)$ in the range 10^3-10^4 bp (see Fig. 3): this is the range of the left tail of the isochore distribution and the right tail of the intermediate scale elements. In the case of the chromosomes 15, 16, 17, 19, and 22 referred above, their isochores are smaller on average (247.2±9.0 kb) than in the remaining chromosomes (368.0±6.3 kb), thus explaining why $\alpha(l)$ does not decrease as in the other chromosomes.

Therefore, we propose that the two clear characteristic scales (intermediate and large) observed in human DNA (Fig. 2) come from two different sources: (i) in the intermediate case, the patchiness and correlations come from the alternation of several genomic elements (exons, introns, TEs, etc.). Note that only one of these elements cannot justify the first wide peak in $\alpha(l)$, and is the combined effect of all of them the reason for the wide peak and the relatively high α value in the interval $10^2 < l < 10^4$. For example, introns can account for high α values for $l > 10^3$ (see Fig. 2), but cannot explain these high α values for smaller sizes. (ii) In the large scale case, the peak observed in $\alpha(l)$ can only be produced by large patches or heterogeneities, which can be identified with the isochores. Quantitatively speaking, the isochore-type segments identified by IsoFinder agrees well with the large-scale peak appearing in all the human chromosomes.

To show that our proposal for the source of the two characteristic scales works, we performed systematic sequence-shuffling experiments. Several conclusions can be drawn from the results, (shown in Fig. 5 only for chr. 21 for simplicity): (i) when the isochores are internally shuffled, the small scale properties disappear, and the $\alpha(l)$ profile remains flat with a constant value of 0.5, as expected in a random sequence without structure. This is also important because as the large scale is practically unaffected by this shuffling and the large scale structure considered comes from the isochores provided by IsoFinder, we can also conclude that the large scale properties observed in the $\alpha(l)$ profile are unambiguously due to the isochores. (ii) The heterogeneities at intermediate scale consists mainly of TEs, introns and exons, since when the isochores are shuffled except these three types of structures, the $\alpha(l)$ profile is almost identical to the profile without shuffling. (iii) The major contribution to the $\alpha(l)$ profile comes from the TEs, while a minor contribution is due to introns and (even less) to exons.

Finally, to show that the structure of $\alpha(l)$ for the human chromosomes is mainly produced by this two-scale picture, we propose a simple procedure to generate artificial two-scale DNA sequences. As the sizes of genomic elements responsible for the patchiness observed at different scales seem to be distributed in a log-normal fashion (Fig. 4), we design a two-step generation procedure in which the characteristic scales are introduced also in this way. The two steps generate, respectively, the large and small scale of the sequence: (i) We produce segments lengths $\{l_i\}$ following a log-normal distribution generating the numbers $\{\log_{10} l_i\}$ from a normal distribution with mean μ_{large} and standard deviation σ_{large} and assign to any segment a random composition x_i (fraction of 1's) picked alternatively from two Gaussian distributions: one of mean μ_{odd} (for i odd) and the other one with mean μ_{even} (for i even), both with the same standard deviation σ_{all}. We finish this step when the sum of the segment sizes reaches a desired size N, $\Sigma_i l_i = N$, thus having generated the structure of the signal *at large scale*. (ii) We proceed simi-

larly, but now working within any of the segments of size l_i and composition x_i generated in the first step. Now, we produce subsegment lengths $\{l_{ij}\}$ log-normally distributed by generating the numbers $\{\log_{10} l_{ij}\}$ from a normal distribution with mean $\mu_{\text{small}} < \mu_{\text{large}}$ and standard deviation σ_{small} to cover the ith segment and finish when $\Sigma_j l_{ij} = l_i$. Now, we generate each subsegment ij of size l_{ij} with a composition x_{ij} (1's with probability x_{ij} and 0's with $1-x_{ij}$) fluctuating around x_i, the composition of the subjacent large scale segment. We choose randomly x_{ij} alternating between even and odd subsegments in the following way: $x_{ij} = x_i + (-1)^j a + c$, where c is a Gaussian random number with 0 mean and σ_c standard deviation, and a is a fixed number establishing the average difference in composition between even and odd subsegments: $2a$. We repeat a similar procedure in any of the large scale segments. To use the model, we calculated the parameters of the two scales for the human DNA case: for the isochores scale, we used the ISOFINDER results, from where $\mu_{\text{large}} = 4.96$, $\sigma_{\text{large}} = 0.6$, $\mu_{\text{odd}} = 45.68\%$, $\mu_{\text{even}} = 41.33\%$, and $\sigma_{\text{all}} = 1.91\%$. For the small scale, for simplicity we consider only a single type of heterogeneity (TEs, the one with major contribution) for which $\mu_{\text{small}} = 2.3$, $\sigma_{\text{small}} = 0.33$, $a = 0$ and $\sigma_c = 5.6\%$. The results of simulations produced with this model are shown in Fig. 3 for the whole genome and for chromosome 21, and compared to the corresponding real sequences. Note the good agreement of $\alpha(l)$ for the real sequences and the simulations, especially in the case of chromosome 21, where TEs are more dominant than in the rest of the genome, and therefore closer to the conditions imposed to the model.

We are aware that real DNA is not as simple as the model proposed here, which is probably the simplest way to preserve not only long-range correlations (since $\alpha > 0.5$ at most of the scales) but also the existence of characteristic scales. We simply want to show that the real DNA organization can appear following similar mechanisms. Roughly speaking, any characteristic scale introduced in a log-normal way leads to a peak in the $\alpha(l)$ curve, and conversely: when a Gaussian-like peak is observed in the $\alpha(l)$ profile, it indicates the presence of a log-normal (or similar) distribution of patches centered at (or close to) the l value of the peak. Finally, we want to point out that our results make compatible the two opposite views described in the Introduction—the existence of isochores and the presence of long-range correlations and segments of many scales: both are present in the human genome structure.

This work was supported by the Spanish Government (Grant No. BIO2005-09116-C03-01) and Plan Andaluz de Investigación (Grant No. P06-FQM-01858).

[1] C.-K. Peng *et al.*, Nature (London) **356**, 168 (1992); W. Li and K. Kaneko, Europhys. Lett. **17**, 555 (1992); R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[2] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996).
[3] J. Filipski, J. P. Thiery, and G. Bernardi, J. Mol. Biol. **80**, 177 (1973); G. Macaya, J. P. Thiery, and G. Bernardi, *ibid.* **108**, 237 (1976); J.-P. Thiery, G. Macaya, and G. Bernardi, *ibid.* **108**, 219 (1976); G. Bernardi *et al.*, Science **228**, 953 (1985); G. Bernardi, Annu. Rev. Genet. **29**, 445 (1995); G. Bernardi, *Structural and Evolutionary Genomics. Natural Selection in Genome Evolution* (Elsevier, Amsterdam, 2004).
[4] J. L. Oliver *et al.*, Nucleic Acids Res. **32**, W287 (2004).
[5] P. Bernaola-Galván *et al.*, Gene **300**, 105 (2002).
[6] D. Holste *et al.*, Phys. Rev. E **67**, 061913 (2003).
[7] W. Li and D. Holste, Phys. Rev. E **71**, 041910 (2005).
[8] B. Borstnik, D. Pumpernik, and D. Lukman, Europhys. Lett. **23**, 389 (1993); S. V. Buldyrev *et al.*, Phys. Rev. Lett. **71**, 1776 (1993); S. Karlin and V. Brendel, Science **259**, 677 (1993); S. V. Buldyrev *et al.*, Phys. Rev. Lett. **71**, 1776 (1993).
[9] E. S. Lander *et al.* (International Human Genome Sequencing Consortium), Nature (London) **409**, 860 (2001); W. Li *et al.*, Comp. Biochem. Physiol. **27**, 5 (2003); J. L. Oliver *et al.*, Gene **276**, 47 (2001); J. L. Oliver *et al.*, *ibid.* **300**, 117 (2002); N. T. Cohen, L. Dagan, and D. Graur, Mol. Biol. Evol. **22**, 1260 (2005); I. Grosse *et al.*, Phys. Rev. E **65**, 041905 (2002); O. Clay and G. Bernardi, Mol. Biol. Evol. **22**, 2315 (2005); M. Costantini *et al.*, Genome Res. **16**, 536 (2006).
[10] G. M. Viswanathan *et al.*, Biophys. J. **72**, 866 (1997).
[11] C. K. Peng *et al.*, Phys. Rev. E **49**, 1685 (1994).
[12] K. Hu *et al.*, Phys. Rev. E **64**, 011114 (2001); Z. Chen *et al.*, *ibid.* **71**, 011104 (2005); A. V. Coronado and P. Carpena, J. Biol. Phys. **31**, 121 (2005).
[13] This is in contrast with other methods such as the power spectrum of the signal, which usually is a very noisy function and therefore more difficult to be fitted.
[14] H. A. Makse *et al.*, Phys. Rev. E **53**, 5445 (1996).
[15] Other authors [see, for example, S. Nicolay *et al.*, Phys. Rev. Lett. **93**, 108101 (2004)] have detected large scale periodicities, and atributed their existence to the chromatin structure or to the existence of replicons. We are confident that the large scale heterogeneities we detect are mainly due to the isochore structure, since our results of shuffled sequences (see Fig. 5) are consistent with the isochore-like segments provided by IsoFinder (see Ref. [4]). Nevertheless, both views are not incompatible, since the replicons or the large-scale chromatin structure must be closely related to the isochores.
[16] C. K. Peng *et al.*, Phys. Rev. E **47**, 3730 (1993).
[17] http://www.ncbi.nih.gov/RefSeq
[18] A. F. A. Smit, R. Hubley, and P. Green, RepeatMasker Open-3.0, 1996–2004, http://www.repeatmasker.org