ELSEVIER

Gene 300 (2002) 105–115

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Study of statistical correlations in DNA sequences

P. Bernaola-Galván[a,*], P. Carpena[a], R. Román-Roldán[b], J.L. Oliver[c]

[a]*Departamento de FÍsica Aplicada II, E.T.S.I. de Telecomunicación, Universidad de Málaga, 29071 Málaga, Spain*
[b]*Departamento de FÍsica Aplicada, Universidad de Granada, Granada, Spain*
[c]*Departamento de Genética and Instituto de BiotecnologÍa. Universidad de Granada, Granada, Spain*

Received 21 December 2001; received in revised form 6 July 2002; accepted 18 September 2002

Abstract

Here we present a study of statistical correlations among different positions in DNA sequences and their implications by directly using the autocorrelation function. Such an analysis is possible now because of the availability of large sequences or even complete genomes of many organisms. After describing the way in which the autocorrelation function can be applied to DNA-sequence analysis, we show that long-range correlations, implying scale independence, appear in several bacterial genomes as well as in long human chromosome contigs. The source for such correlations in bacteria, which may extend up to 60 kb in *Bacillus subtilis*, may be related to massive lateral transfer of compositionally biased genes from other genomes. In the human genome, correlations extend for more than five decades and may be related to the evolution of the 'neogenome', a modern evolutionary acquisition composed by GC-rich isochores displaying long-range correlations and scale invariance.

Keywords: DNA sequence analysis; Statistical correlation; Autocorrelation function; Long range correlation

1. Introduction

The autocorrelation function has been widely used in Physics and Signal Theory as a measure of linear dependence and periodicity. The application of this measure to DNA-sequence analysis, although used previously, became 'popular' in 1992 with the finding of power-law correlations in DNA sequences implying the presence of a high complexity and scale invariance in the heterogeneity of those sequences (i.e. fractal properties: Peng et al., 1992; Li and Kaneko, 1992; Voss, 1992). Since then, a great controversy over the biological implications of this finding has arisen, even questioning the existence of the correlations themselves (Borštnik et al., 1993; Buldyrev et al., 1993b; Karlin and Brendel, 1993; Voss, 1993). The lack of long enough sequences in those days led to the use of indirect methods of estimating the correlations and, frequently, each method studied the problem from a different viewpoint, leading to results not comparable to each other, being this fact a plausible source of the above-mentioned controversy. Now, however, the availability of large sequences or even complete genomes of many organisms makes the direct calculation of the autocorrelation functions possible. Therefore, the existence of scale-independent correlations in DNA sequences can be directly tested, a possibility which may have profound implications for understanding genome organization and evolution.

The present work is organized as follows. In Section 2.1 we define the autocorrelation function, $C(\ell)$, describe the way in which it can be applied to DNA analysis, comment the effects on $C(\ell)$ of the finite size of the analysed sequence and briefly describe other indirect methods to calculate the autocorrelations. In Section 2.2 we show several examples of $C(\ell)$ plots for computer-generated sequences having different kinds of heterogeneity. The DNA sequences used are summarized in Section 2.3. In Section 3 we present and discuss the results of the analysis of prokaryotic complete genomes (Section 3.1) and the longest human contigs available in July 2001 (Section 3.2). Finally, Section 4 concludes the paper.

2. Data and methods

2.1. The autocorrelation function

Given the numerical sequence $S = \{x_1, x_2, \ldots, x_N\}$ with

* Corresponding author. Tel.: + 34-952-132748; fax: + 34-952-131450.
E-mail address: rick@uma.es (P. Bernaola-Galván).

0141-933/02/$ - see front matter
PII: S0378-1119(02)01037-5

Table 1
Mapping rules used to convert DNA sequences into binary numerical sequences

Rule	Assignment	
SW	C or G = 1	A or T = 0
RY	A or G = 1	C or T = 0
KM	G or T = 1	A or C = 0
A	A = 1	T, C or G = 0
T	T = 1	A, C or G = 0
C	C = 1	A, T or G = 0
G	G = 1	A, T or C = 0

variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2 \quad (1)$$

the autocorrelation of S at distance ℓ is defined by:

$$C(\ell) = \frac{1}{\sigma^2} \times \left[\frac{1}{N-\ell}\sum_{i=1}^{N-\ell} x_i x_{i+\ell} - \frac{1}{(N-\ell)^2}\sum_{i=1}^{N-\ell} x_i \sum_{i=1}^{N-\ell} x_{i+\ell}\right] \quad (2)$$

Basically, the autocorrelation function measures the deviation of $\langle x_i x_{i+\ell}\rangle$ from $\langle x_i\rangle\langle x_{i+\ell}\rangle$ (where the brackets denote the average along the sequence). If this deviation is zero, there are no linear correlations between the values of the sequence at position i (x_i) and the values at distance ℓ ($x_{i+}\ell$). The greater this deviation the stronger the linear correlations between values separated a distance ℓ. $C(\ell) > 0$ means that the probability that x_i and $x_{i+}\ell$ reach the same or similar values is higher than in a random sequence whereas for $C(\ell) < 0$ this probability is smaller than that expected in a random sequence (anti-correlations). Note that, by definition, $C(\ell = 0) \equiv 1$ and thus carries no information about the correlations in the sequence.

The definition of $C(\ell)$ (Eq. (2)) makes the implicit assumption of stationarity since it includes the variance of the whole sequence (Eq. (1)). Nevertheless, as we will show in Section 2.2, we can compute $C(\ell)$ for non-stationary sequences and extract from it relevant information about the heterogeneities present in the sequence. This is the reason why we do not use several well-known results on autocorrelation which are valid only under the assumption of stationarity of the sequence analysed, e.g. the relationship between $C(\ell)$ and the power spectrum, Wiener–Khinchin theorem.

It is important to note that $C(\ell)$ measures only linear correlations. To take into account non-linear correlations, we would need to include higher-order terms in Eq. (2), i.e. products in which higher powers of x_i and $x_{i+}\ell$ appear. For this reason $C(\ell)$ is sometimes called second-order autocorrelation function (Herzel and Grosse, 1995; Teitelman and Eeckman, 1996). Nevertheless, previous studies (Arneodo et al., 1995; Yu et al., 2001a) seem to indicate that correlations in DNA are essentially linear and thus $C(\ell)$ turns out to be an appropriate tool for DNA analysis.

As follows directly from its definition (Eq. (2)), the autocorrelation function can be applied only to a numerical sequence. Thus, to calculate the autocorrelation function of a DNA sequence, the symbols (A,T,C,G) have to be converted into numerical quantities. At this point, a caveat should be mentioned: if we simply assign a numerical value to each nucleotide the resulting correlations will depend on the particular assignment, i.e. the mapping itself can cause spurious results. It can be proven directly from Eq. (2) that this problem does not appear when the sequence is binary, that is, the autocorrelation function is the same for all possible numerical assignments[1] Thus, this problem can be overcome by converting the DNA sequence into a binary sequence, grouping the four nucleotides into two groups and assigning 1 to one group and 0 to the other. There are seven possible groupings, usually called mapping rules (Buldyrev et al., 1995), as shown in Table 1. In addition, we can define the cross-correlations by making a different numerical assignment to x_i and $x_{i+}\ell$, e.g. $x_i = 1$ when an A is found at position i and 0 for anything else and $x_{i+}\ell = 1$ when a T is found at position $i + \ell$ and 0 for anything else. Such an autocorrelation function would measure the statistical properties of the pairs (A,T) separated by a distance ℓ. All combinations of these mapping rules and cross-correlations are also possible but only nine of these autocorrelation functions are independent while the rest can be expressed as linear combinations of them (Herzel and Grosse, 1995; Teitelman and Eeckman, 1996).

In principle, the results obtained with each of these mapping rules are independent of each other because they refer to different aspects of the DNA chain, all of them containing relevant information. For example, the RY rule describes how purines and pyrimidines are distributed along the sequence whereas the A-rule deals with the distribution of nucleotide A along the sequence. In previous works on statistical correlations in DNA, the rules mainly used were the RY rule (Arneodo et al., 1995; Bernaola-Galván et al., 1996; Buldyrev et al., 1993a, 1995; Mohanty and Narayana Rao, 2000; Peng et al., 1992, 1994) as well as the single-letter rules (A,T,C and G rules in Table 1) (Voss, 1992, 1994; Audit et al., 2001; de Sousa Vieira, 1999) but also the SW rule (Arneodo et al., 1998). Also, the average of the four correlation functions obtained with the four single-letter rules (Li and Kaneko, 1992; Li et al., 1994, 1998) were used. The SW mapping rule is particularly appropriate to analyse genome-wide correlations; this rule corresponds to the most fundamental partitioning of the four bases into their natural pairs in the double helix (G + C, A + T). The composition

[1] Another definition of the autocorrelation function without the subtracting term has been used to analyse DNA sequences (de Sousa Vieira, 1999). For such definition, spurious results can appear in the autocorrelation function even for binary sequences since $C(\ell)$ depends on the particular numerical assignment.

of base pairs, or GC level, is thus a strand-independent property of a DNA molecule and is related to important physico-chemical properties of the chain such as the transport of electrons (Dandliker et al., 1997; Carpena et al., 2002b) or mechanical waves (Rief et al., 1999) along the sequence. One of its best-known correlates is resistance to denaturation at high temperatures.

Li and Kaneko (1992) proposed the use of the mutual information to measure the correlations in DNA sequences. This function is a direct measure of all correlations (not only linear) and is especially suitable for analysing of symbolic sequences (Herzel and Grosse, 1995, 1997; Li, 1990). In a first-order approximation, taking into account that correlations in DNA are essentially linear, it can be shown that mutual information is a linear combination of the squares of autocorrelation and cross-correlation functions. The only drawback of this measure would appear in the study of the individual contribution of each mapping rule to the correlations in the sequence. This is the issue we address here.

One of the main drawbacks of $C(\ell)$ is the need of long sequences to ensure good results free from statistical fluctuations. The error in the determination $C(\ell)$, due to statistical fluctuations, can be easily estimated to be (Weiss and Herzel, 1998):

$$\Delta C = \frac{1}{\sqrt{N}} \tag{3}$$

where N is the size of the sequence analysed. Thus, the shorter the sequence, the larger the fluctuations of $C(\ell)$. This is a serious disadvantage, especially when the correlations we seek to measure are weak. Note that ΔC is a threshold below which the correlations can be considered to be due to statistical fluctuations.

In 1992, when the first results on long-range fractal correlations were presented, the longest available DNA sequences were around 100 kb in length, making it difficult to draw clear plots of $C(\ell)$ vs. ℓ for long distances. This fact motivated the use of indirect measures of $C(\ell)$. These include the analysis of variance of base composition[2] (Peng et al., 1992), power-spectrum analysis (Li and Kaneko, 1992; Voss, 1992) and wavelet analysis (Arneodo et al., 1995, 1998; Audit et al., 2001).

In short, the variance of base composition can be analysed as follows: first, consider a window of length ℓ located at the beginning of the sequence and compute the sum of the sequence inside the window, s_1. Then, move the window one position (overlapping windows) or ℓ positions (non-overlapping windows), and compute s_2, and so on. Finally, calculate the variance of these numbers $\{s_i\}$, $\sigma^2(\ell)$ and repeat the procedure for each window length, ℓ. If the sequence is stationary (or loosely speaking, homogeneous, Clay, 2001), we can obtain $C(\ell)$ from $\sigma^2(\ell)$. This requirement raised the main objections to the results obtained with this method (Karlin and Brendel, 1993). In 1994, an improvement to the $\sigma^2(\ell)$ method was introduced to eliminate the effect of the heterogeneity of the sequence analysed: the Detrended Fluctuation Analysis (DFA) (Peng et al., 1994). Although DFA considerably improves the results obtained with $\sigma^2(\ell)$, a recent study shows that the DFA can still be affected by the trends present in the sequence (Hu et al., 2001). Here it is worth mentioning that most of the criticisms against Peng and co-workers' results predate the DFA (Borštnik et al., 1993; Buldyrev et al., 1993b; Karlin and Brendel, 1993; Voss, 1993).

The power spectrum gives another indirect way of computing the autocorrelation function but, again, the sequence analysed has to be stationary (Wiener–Khinchin theorem). This drawback is solved by wavelet analysis, which can be considered as a 'local Fourier transform' and is able to eliminate the effects of the non-stationarity of the sequence.

On the contrary, the direct measure of $C(\ell)$ using Eq. (2) requires neither homogeneity of the sequence nor an additional procedure or improvement to eliminate the heterogeneity of the sequence. In addition, several authors have proposed that the complex organization of such heterogeneity is responsible for the fractal correlations observed in human DNA (Bernaola-Galván et al., 1996; Buldyrev et al., 1993a; Li, 1997; Román-Roldán et al., 1998). Therefore it is not clear whether the elimination of the heterogeneity is a good strategy.

2.2. Examples

In this section we show several examples of $C(\ell)$ plots for computer-generated sequences with different kinds of heterogeneity which will be useful in interpreting the results found for real DNA sequences.

2.2.1. Random homogeneous sequence

Let us consider a sequence obtained by randomly generating 1's (with probability p) and 0's (with probability $1-p$). In such a sequence the value at position i (x_i) is independent of the value at position $i+\ell$ ($x_i+\ell$) for all $\ell \geq 1$. Thus in the limit of a sequence of infinite length $\langle x_i x_{i+\ell}\rangle = \langle x_i\rangle\langle x_{i+\ell}\rangle$ and the correlation function will vanish for all $\ell \geq 1$. For a sequence of finite length, N, $C(\ell)$ will not exactly vanish but oscillate around 0, the amplitude of this oscillation being of the order of $1/\sqrt{N}$ (Eq. (3)). This result is also applicable to a sequence with random heterogeneities having sizes smaller than a given size ℓ^*. For such a sequence $C(\ell)$ behaves as the autocorrelation function of a random sequence i.e. $C(\ell) = 0 \pm O(\Delta C)$, when $\ell > \ell^*$ The case of a sequence with a periodic pattern of heterogeneity is discussed in the next paragraph.

[2] Here, the term analysis of variance of base compositon does not refer to ANOVA but to a different analysis used in Solid State Physics, also known as σ^2 analysis.

2.2.2. Patchy sequence

Now, let us consider a sequence built by alternating two types of subsequences, S_1 and S_2, both of length ℓ_0 and let p_1 and p_2 be the relative proportions of 1's in S_1 and S_2, respectively. It can be shown (Bernaola-Galván et al., in preparation) that the autocorrelation function is given by:

$$C(\ell) = \frac{(p_1 - p_2)^2}{2(p_1 + p_2) - (p_1 + p_2)^2}\left[1 - \frac{2\ell}{\ell_0}\right] \quad \text{for} \quad 1 \leq \ell \leq \ell_0 \tag{4}$$

Thus, $C(\ell)$ is a straight line with negative slope which intercepts the x-axis at $\ell = \ell_0/2$ (Fig. 1a).

In a double-logarithmic plot $C(\ell)$ shows an almost flat plateau with a sharp bend around $\ell = \ell_0/2$. The height of the plateau is given by (Fig. 1b):

$$C_{\text{Plateau}} = \frac{(p_1 - p_2)^2}{2(p_1 + p_2) - (p_1 + p_2)^2} \tag{5}$$

Although a perfect alternating sequence such as the one described here is not likely to occur in a natural DNA sequence, a sequence with an almost alternating structure and a distribution of differences in composition centred at $p_1 - p_2$ will give a profile similar to that shown in Fig. 1.

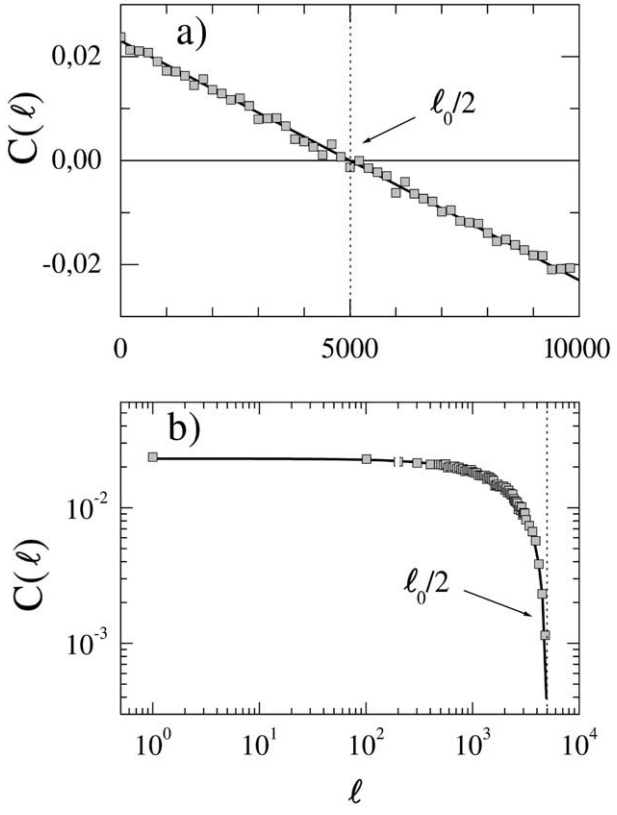

Fig. 1. Autocorrelation function vs. ℓ for an artificial sequence obtained by alternating two types of sequences S_1 and S_2 both of length $\ell_0 = 10,000$ bp and proportions of 1's $p_1 = 0.5$ and $p_2 = 0.65$, respectively; (a) linear scale; (b) double-logarithmic scale. According to Eq. (5), $C_{\text{Plateau}} = 2.31 \times 10^{-2}$.

2.2.3. Distribution of patches with a characteristic length

If, instead of having a sequence composed of patches of the same size, we have different sizes but following a distribution with a characteristic length ℓ_0 (Gaussian, exponential, Poisson, etc.), the shape of $C(\ell)$ is quite similar to the one shown in Fig. 1, although not exactly the same. In general, the wider the distribution of sizes the slower the decrease of $C(\ell)$ (i.e. the bend becomes smoother). In particular, for a sequence of alternating patches with a distribution of sizes following an exponential distribution with mean ℓ_0:

$$p(\ell) = \frac{1}{\ell_0} e^{-\frac{\ell}{\ell_0}} \tag{6}$$

$C(\ell)$ is also an exponential function with characteristic length $\ell_0/2$:

$$C(\ell) = C_0 e^{-\frac{2\ell}{\ell_0}} \tag{7}$$

where C_0 is a constant depending on the strength of the correlations at short scale and, for this model, C_0 is also given by Eq. (5). An example of such an autocorrelation function is shown in Fig. 2. The shape of the autocorrelation function in double logarithmic scale is quite similar to the one corresponding to the patchy sequence, although in the former case the decrease is much more abrupt. In many cases, the only way to distinguish between the two types of behaviour is to examine the linear–linear plot.

2.2.4. Power-law distribution of patches

The first three examples of this section have something in common: in both cases there exists a finite length, ℓ_0, related to the size of the building block or the characteristic length of the distribution of patches. Now, however, let us consider a distribution of lengths following a power law, i.e. the probability of finding a patch with length ℓ is given by

$$p(\ell) \propto \frac{1}{\ell^{\mu}} \tag{8}$$

This distribution is scale-invariant; this means that the heterogeneities of the sequences appear at all scales and are also statistically self-similar.[3] It can be shown (Buldyrev et al., 1993a) that such a distribution of patches results in an autocorrelation function which is also a power law:

$$C(\ell) = \frac{C_0}{\ell^{\gamma}} \tag{9}$$

where, again, C_0 is a constant which depends on the strength of the correlations at short scales. A derivation of the relationship between the exponents μ and γ can be found in Buldyrev et al., 1993a. Due to the fact that Eq. (9) decays slower than does an exponential function (Fig. 3), power-law correlations are usually called long-range correlations, meaning that the range in which $C(\ell)$ is substantially larger

[3] To be precise, the distribution of patches is statistically self-similar only if $\mu = 1$. If $\mu \neq 1$ the distribution is said to be statistically *self-affine*.

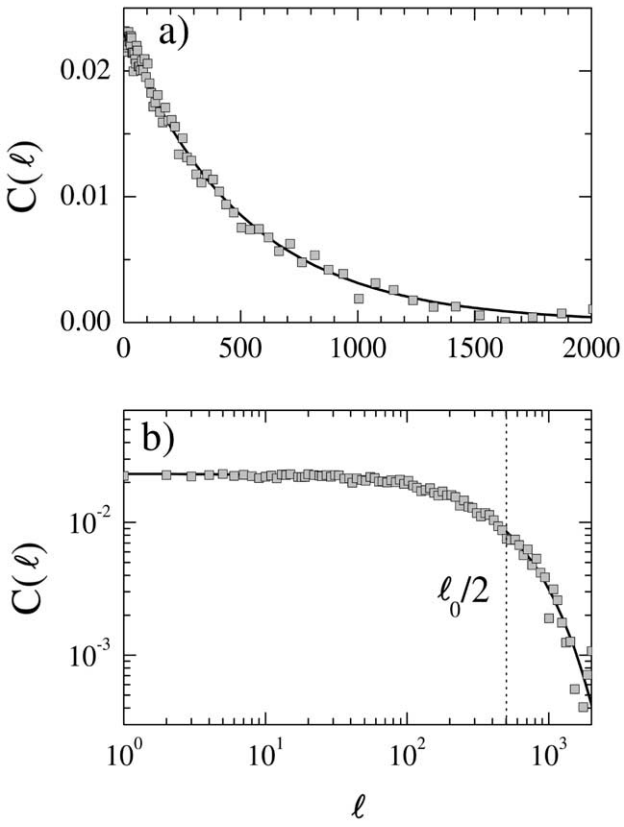

Fig. 2. Autocorrelation function vs. ℓ for an artificial sequence obtained by alternating two types of sequences, S_1 and S_2, with proportions of 1's $p_1 = 0.5$ and $p_2 = 0.65$, respectively, and sizes randomly drawn from an exponential distribution

$$p(\ell) = \frac{1}{\ell_0} e^{-\frac{\ell}{\ell_0}}$$

(mean value $\ell_0 = 1000$). The solid line is the fit of $C(\ell)$ to

$$C(\ell) = C_0 e^{-\frac{2\ell}{\ell_0}}$$

with $C_0 = 2.33 \times 10^{-2}$, very close to the prediction of Eq. (5); (a) linear scale; (b) double-logarithmic scale.

than 0 is very broad. Nevertheless, this term is not completely accurate because an exponential function with large enough ℓ_0 also produces long-range correlations. At this point, it is worth mentioning that such exponential correlations can be wrongly interpreted as power-law correlations when the range over which correlations are observed is not large enough.

The length distributions shown in the first three examples of this section can be produced by very simple mechanisms: repetition of a given pattern, short-range memory (as in a Markov Chain), etc. On the other hand, the presence of power-law correlations or size distributions are usually related to complex interactions in which many effects contribute to the final result. For example, random mechanisms can lead to a power-law distribution, but these mechanisms must act at all scales and in a complex way (Hausdorff and Peng, 1996). Nevertheless, there are several examples of simple procedures based on mutations, deletions and/or substitutions which may lead to power-law correlations (Buldyrev et al., 1993c; Li, 1989, 1991).

2.3. Analysed DNA sequences

We analysed sequences of prokaryotic complete genomes retrieved from the EBI website: http://www.ebi.ac.uk/cgi?bin/genomes/genomes.cgi?genomes = Bacteria. The longest contigs of human chromosomes were retrieved from the July 2001 freeze of the publicly available human-genome draft sequence: ftp://ncbi.nlm.nih.gov/genomes/H sapiens

3. Results and discussion

3.1. Prokaryotic genomes

A common feature, found in all the bacterial and archaebacterial genomes analysed, is the presence of two branches in the autocorrelation function (Fig. 4). On one hand, we observe that, for distances which are multiples of 3 and below a critical length $\ell^* \approx 1000-2000$ bp, $C(\ell)$

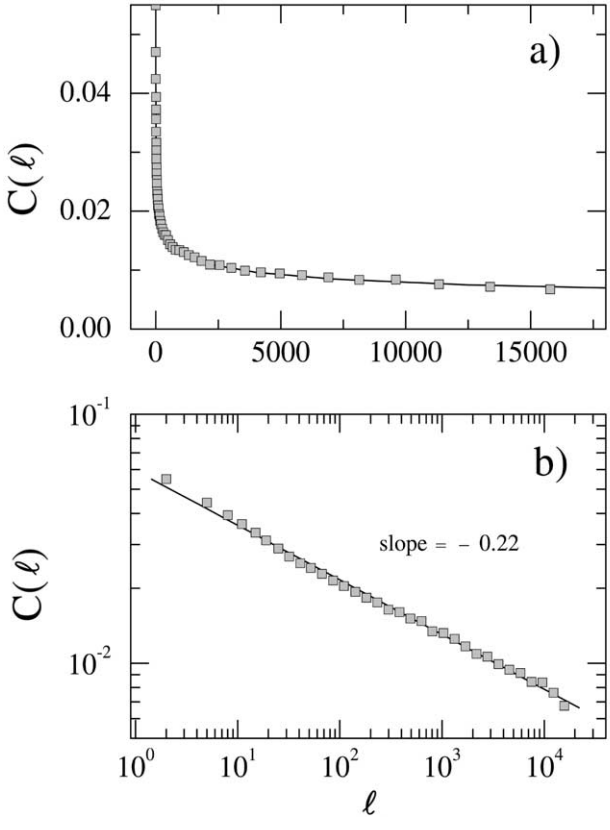

Fig. 3. Autocorrelation function vs. ℓ for an artificial sequence obtained by alternating two types of sequences, S_1 and S_2, with proportions of 1's $p_1 = 0.5$ and $p_2 = 0.65$, respectively, and sizes obtained randomly from an power-law distribution $p(\ell) \propto \ell^{-\mu}$ (Generalized Levy-walk model, Buldyrev et al., 1993a). The solid line is the fit of $C(\ell)$ to $C_0 \ell^{-\gamma}$ with $\gamma = 0.22$; (a) linear scale; (b) double-logarithmic scale.

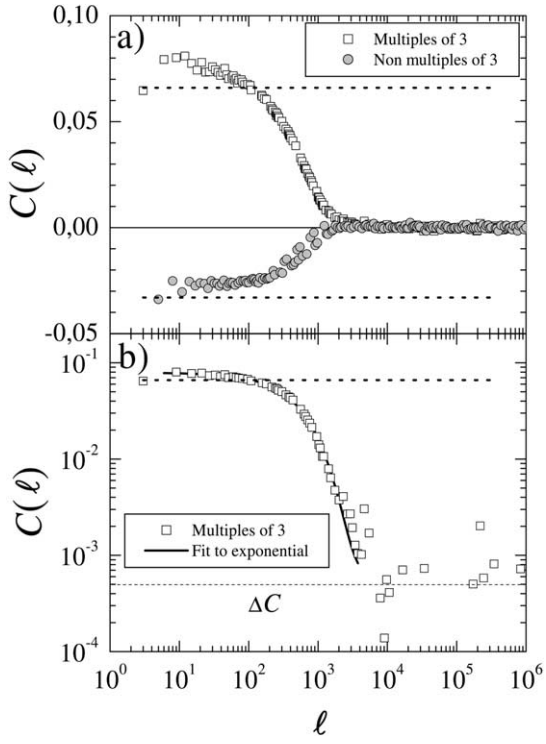

Fig. 4. Autocorrelation function vs. ℓ using the SW mapping rule for the complete genome of *Mycobacterium tuberculosis* (4,411,529 bp). Full squares correspond to distances which are multiples of 3, and open circles to the rest of distances. (a) Log-linear scale. The upper dotted line corresponds to the correlations due to the non-uniform distribution of base composition at the three codon positions at distances which are multiples of 3, the lower one at distances which are not multiples of 3. (b) Log-log scale. The dotted line is again the value of the correlations due to the non-uniform distribution of base composition at the three codon positions for distances which are multiples of 3. The thick solid line corresponds to the fit of $C(\ell)$ for distances which are multiples of 3 to an exponential decay (Eq. (7)). The value of the statistical fluctuations, ΔC, is marked by the dashed line at bottom. Below this value, the correlations can be considered to be due to statistical fluctuations.

reaches high values whereas for $\ell > \ell^*$, a fast decrease of $C(\ell)$ is observed. On the other hand, for distances which are not multiples of 3, we find anticorrelations ($C(\ell) < 0$) or very weak correlations. For lengths larger than ℓ^*, both branches collapse on top of each other. This effect, previously observed using other direct measures of correlation (Herzel and Grosse, 1997),[4] can be explained by the well-known non-uniform distribution of base composition at the codon positions inside coding regions. This effect causes the concentration of each nucleotide to be different in all three positions of the reading frame. To obtain a first estimate of the effect of this position asymmetry in the correlation function, let us consider a DNA coding region as the concatenation of randomly chosen codons (Herzel and Grosse, 1997), using the empirically observed frequency of bases at the three codon positions. For such an artificial sequence, for all distances which are multiples of 3 the autocorrelation function is almost constant, $C(\ell) \approx C_3$, and is also almost constant for all distances which are not multiples of 3, $C(\ell) \approx C_{1,2}$ with $C_3 \neq C_{1,2}$. In Fig. 4, the dotted lines represent the values of C_3 and $C_{1,2}$ obtained generating an artificial sequence of random codons using the probability of bases at the three codon positions of *Mycobacterium tuberculosis* (Table 2). Almost all correlations at short scales can be explained by this different base frequency at each codon position; nevertheless, there is not a perfect fit, this means that there are correlations at short scales that cannot be explained by this effect. The fact that for distances greater than $\ell^* \approx 1000-2000$ both curves decay to zero is a consequence of the finite size of the coding regions: in our simple model we consider a concatenation of codons all in the *same phase*, but, in a real sequence, the presence of non-coding regions may cause the successive coding regions to be out of phase. Here, we can consider the sequence to be a concatenation of patches (coding plus non-coding). As the distribution of sizes of these patches follow an exponential-like distribution we might expect that the decay of $C(\ell)$ could be fitted by an exponential function. In Fig. 4b we show the fit of $C(\ell)$ to an exponential decay for distances that are multiples of 3 and obtain a characteristic length $\ell_0/2 = 639$ bp, which gives a mean patch size of $\ell_0 = 1278$ bp, in agreement with the mean value of the sizes of coding regions plus the size of its adjacent non-coding region (1192 bp). The results obtained with the RY mapping rule are quite similar.

In this example (*M. tuberculosis*), beyond ℓ^* the correlations practically vanish (in Fig. 4b, $C(\ell)$ reaches the noise level around $\ell = 6000$ bp). This means that almost all the correlations present in this genome are due only to the different base frequency at the three codon positions. Also, the fact that above ℓ^*, $C(\ell)$ is practically zero means that above the characteristic size of the genes, the sequence is homogeneous, in agreement with the commonly accepted idea that bacterial genomes are homogeneous (Rolfe and Meselson, 1959; Sueoka, 1959). However, as we show in Fig. 5 for *Bacillus subtilis*, this behaviour is not common to all bacteria analysed. In this case, we again observe two branches in the autocorrelation function which collapse on top of each other around 1000–2000 bp, but now $C(\ell)$ significantly differs from zero for lengths of up to 60,000 bp, indicating the presence of heterogeneities up to this size.

Table 2
Probabilities of finding each nucleotide in the three codon positions for the complete genome of *Mycobacterium tuberculosis*

Codon position	A	T	C	G
1	0.18807	0.13256	0.25864	0.42072
2	0.22508	0.27418	0.28695	0.21379
3	0.09149	0.11318	0.42255	0.37278

[4] Here it bears mentioning that these branches in the autocorrelation function cannot be detected with indirect measures of correlation because, implicit or explicitly, they make an integration which smoothes out the alternating behaviour of $C(\ell)$.

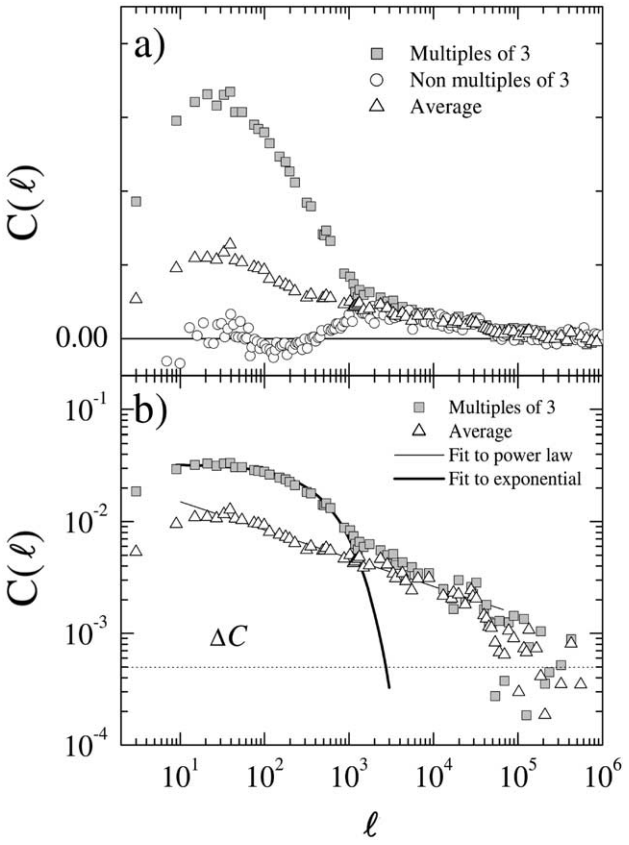

Fig. 5. Autocorrelation function vs. ℓ using the SW mapping rule for the complete genome of *Bacillus subtilis* (4,214,814 bp). Full squares correspond to distances which are multiples of 3, open circles the rest of distances, and open triangles the average between multiples and non-multiples of 3. (a) Log-linear scale. (b) Log-log scale. The thick solid line corresponds to the fit of $C(\ell)$ for distances which are multiples of 3 to an exponential decay (Eq. (7)) and the thin solid line to the fit of the average (over multiples and non-multiples of 3) to a power law. The value of the statistical fluctuations, ΔC, is marked with a dotted line. Below this value, the correlations can be considered to be due to statistical fluctuations.

This result, found in many other bacteria (not shown), clearly challenges the validity of the assumption of homogeneity in bacteria beyond the size of genes. Actually the value of $C(\ell)$ at the collapsing length, and therefore where the influence of genes disappear, can be considered to be a measure of the heterogeneity in a given bacterial genome.

Another question, widely debated in recent years, is whether long-range power-law correlations exist in prokaryotic genomes. The literature is inconclusive since, depending on the method considered, all possible results can be found: the lack of power-law correlations in prokaryotes (Bernaola-Galván et al., 1996; Buldyrev et al., 1995; Peng et al., 1992, 1994), the existence of power-law correlations covering all sizes in all prokaryotes (Yu et al., 2001b) and the existence of intervals in which the correlations follow a power-law only in certain intervals of lengths and in certain organisms (de Sousa Vieira, 1999; Mohanty and Narayana Rao, 2000; Audit et al., 2001; Yu et al., 2001a).

Here, we find that, in general, the correlations are long-ranged in the sense that $C(\ell) \neq 0$ for large values of ℓ but, in terms of the raw data, $C(\ell)$ does not seem to follow a power-law but an exponential decay. Nevertheless studying the correlations averaging the behaviour for distances which are multiples of 3 and non-multiples of 3 in some cases, we find that $C(\ell)$ can be well fitted by a power-law decay (see Fig. 5b). The sequences for which we find a power-law behaviour (not shown) agree well with the results presented in de de Sousa Vieira (1999), although the distance ranges are not exactly the same.

Note in Fig. 5b that throughout most of the region in which the average of $C(\ell)$ can be fitted by a power law, the values of the average differ from the values at distances which are multiple of 3. This fact makes the explanation of these correlations plausible in terms of non-uniform distribution of base composition at the codon positions. In fact, *B. subtilis* has coding regions with sizes of up to 15,000 bp. Nevertheless, for ℓ between 10^3 and 10^5 this difference is not very large and, in any case, much smaller than the value of the correlation itself. In this range, additionally to the influence of non-uniform distribution of base composition at the codon positions, the 537 genes clustered in several A + T-rich islands scattered along the chromosome, coming from lateral transfer, may also help explain the long-range correlations found in this genome. Nevertheless, the question which arises is whether or not this behaviour results from a true scale invariance in the sequence or is simply an artefact of the averaging procedure.

With respect to the RY mapping rule, the results are qualitatively similar, although the autocorrelation function significantly differs from zero up to lengths close to 1 Mb. In this case, there is a clear explanation: the great difference in the asymmetry and coding content between the two

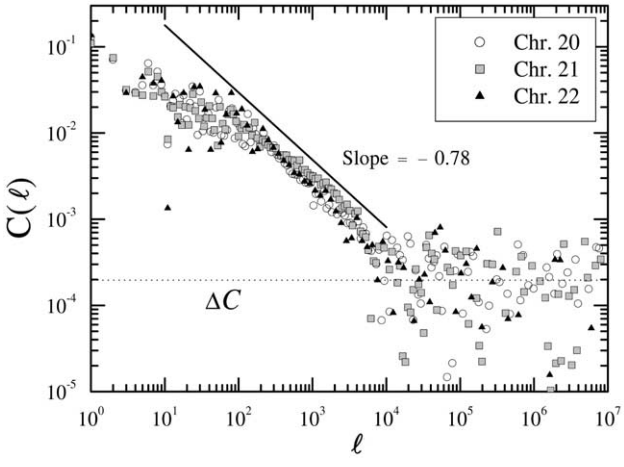

Fig. 6. Autocorrelation function vs. ℓ using the RY mapping rule for the longest contigs of human chromosomes 20 (NT_011362.4, 24,982,240 bp), 21 (NT_011512.3, 28,515,322 bp) and 22 (NT_011520.6, 22,963,592 bp). The solid line corresponds to a power-law decay with the exponent observed in (Peng et al., 1994) using DFA. The value of the statistical fluctuations, ΔC, is marked with a dotted line. Below this value, the correlations can be considered to be due to statistical fluctuations.

replichores of *B. subtilis* (Carpena et al., 2002; Li et al., 2002).

3.2. Human DNA

One of the most striking results presented in the same year by several authors (Li and Kaneko, 1992; Peng et al., 1992; Voss, 1992) was the existence of long-range power-law correlations in human DNA sequences, indicating the presence of scale-invariant structure covering lengths close to the whole sequence length. It is important to note that the longest sequences available at that time were around 100 kb in length and the observed correlations spread over distances of tens of kb (Peng et al., 1994). Despite the controversy concerning the presence of these correlations in other organisms and their presence only in non-coding DNA or in all human DNA (Arneodo et al., 1998; Buldyrev et al., 1995; Voss, 1994), now such correlations in human DNA are commonly-accepted.

Using the RY mapping rule to compute $C(\ell)$, we find similar behaviour for all human contigs analysed (Fig. 6): (a) the absence of two branches for multiples and non-multiples of 3, due to the small proportion of coding regions (Lander et al., 2001; Venter et al., 2001); (b) a region that can be fitted to a power-law decay in the range $\ell < \ell < 10^4$ with similar exponents in all contigs analysed; and (c) after this region the correlations reach noise level.

The finding of power-law correlations in the range $\ell < \ell < 10^4$ (also found by Holste et al., 2001, for chromosome 22) agrees with the results previously reported using DFA, as shown in Fig. 6, where we show a power law with the exponent found in Peng et al. (1994) for the human sequence HUMTCRADCV (human T-cell receptor alpha/delta locus, 97,634 bp).[5] Nevertheless, these correlations do not extend to sizes comparable to the whole size of the chromosome as was believed when the longest human sequences available were only a few tens of kb in length.

With respect to the correlations found using the SW mapping rule, the main features are: (a) again, the absence of two branches due to the small proportion of coding DNA in these sequences (Lander et al., 2001; Venter et al., 2001); (b) noisy behaviour at short scales ($\ell < 200-300$ bp); and (c) the persistence of correlations for lengths which in many cases are greater than 1 Mb.

This fact implies the existence of inhomogeneities of sizes which can surpass even 1 Mbp. As a first approximation, we can estimate the differences in G + C content between such DNA fragments using Eq. (5). For example, for the longest contig of human chromosome 22 (Fig. 7), we find a characteristic difference in G + C content of around 6% between fragments of sizes around 300 kb. Both the characteristic size of the greater fragments as well as their difference in composition agree with the values commonly accepted for isochores (Bernardi et al., 1985). It is noteworthy that two conflicting views currently exist on isochores. Although originally described as 'fairly homogeneous regions' (Bernardi, 2001; Cuny et al., 1981), isochores have recently been identified with random uncorrelated sequence regions (Häring and Kypr, 2001; Lander et al., 2001; Nekrutenko and Li, 2000). While a certain level of internal heterogeneity is accepted under the first definition, only statistical fluctuations below the standard deviation are permitted under the second, i.e. strict concept of isochores. As shown below, our results support the first view, in agreement with other authors using different techniques (Bernardi, 2001; Clay and Bernardi, 2001; Li, 2001).

Fig. 7 shows that $C(\ell)$ can be well fitted by a power-law decay for more than five decades and thus the sequence presents scale invariance, at least for this range of sizes: we can find heterogeneity at all scales and we have no 'special' or characteristic scale. This picture does not seem to agree with the presence of isochores, nor with either of the two above-mentioned definitions for isochores. The presence of power-law correlations for more than five decades in the sequence of human chromosome 22 has also been found recently using the mutual information function as the measure of the correlations (Holste et al., 2001).

Nevertheless, among all human contigs analysed the behaviour in human chromosome 22 is a clear exception. A fairly typical plot of $C(\ell)$ vs. ℓ is in Fig. 8 for the longest contig of human chromosome 21 and in Fig. 9 for the longest contig of human chromosome 20. In general, we find two regions in $C(\ell)$: a power-law decay in the range $1 \leq \ell \leq \ell_0 \approx 1000$ bp and a second region for $\ell_0 < \ell < \ell_{max}$, ℓ_{max} being different for each contig. This second region can also be fitted to a power law with an exponent usually smaller than that in the first region, although the possibility

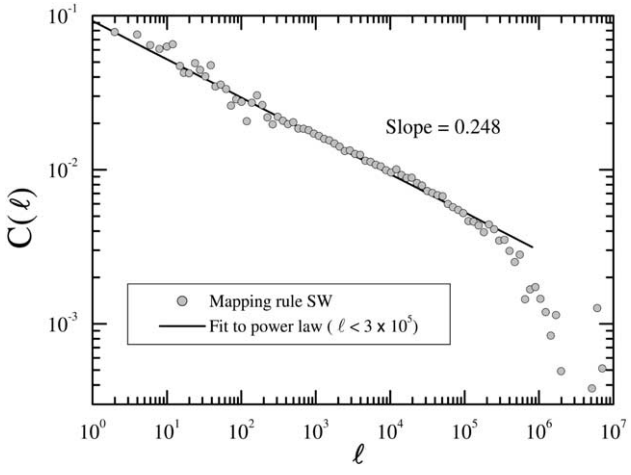

Fig. 7. Autocorrelation function vs. ℓ using the SW mapping rule for the longest contig of human chromosome 22 (NT_011520.6, 22,963,592 bp). Solid line: fit to a power law of the data in the range $1 \leq \ell \leq 3 \times 10^5$ (slope = −0.248).

[5] The exponent used by Peng et al. (1994) to charcterize the correlations, α, differs from the exponent of the power-law decay used here, γ, but they are related to the formula $\gamma = 2(1 - \alpha)$.

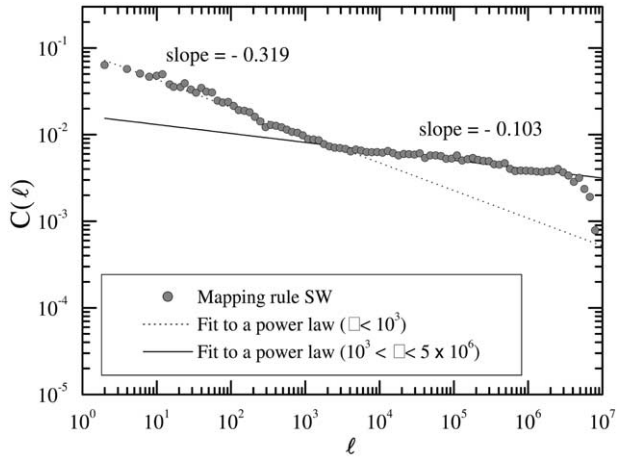

Fig. 8. Autocorrelation function vs. ℓ using the SW mapping rule for the longest contig of human chromosome 21 (NT_011512.3, 28,515,322 bp). Solid line: fit to a power law of the data in the range $10^3 < \ell < 5 \times 10^6$ (slope $= -0.103$). Dotted line: fit to a power law of the data in the range $1 < \ell < 10^3$ (slope $= -0.319$)

of fitting this second region to stretched exponentials or even to linear functions (see Eq. (4)) must be taken into account. In fact, in several cases (e.g. Fig. 8), this second regime is an almost flat plateau followed by a sharp bend, which clearly resembles the one in Fig. 1b.

This picture is more likely to correspond to the presence of long regions with different composition, relatively homogeneous above a certain scale (ℓ_0), i.e. fairly homogeneous isochores. If we had long perfectly homogeneous regions (as the strict isochore model proposes), as those discussed in Section 2.2, we would have a plot similar to that of Figs. 1 or 2, i.e. a flat plateau starting from $\ell = 1$. However, the steeper slope found for $\ell < \ell_0$ indicates a relative increase of heterogeneities at short distances that has been explained in terms of repetitive DNA (Holste et al., 2001). In addition, the length up to which the correlations extend in each contig, ℓ_{max}, is related to the size of the largest isochore found in each contig (Oliver et al., 2002). The fact that the second regime can be considered to be a power law, instead of a flat plateau, can be attributed to the presence of a broad distribution of long homogeneous regions: lognormal or stretched exponential distribution (Oliver et al., 2001, 2002).

Now, coming back to Fig. 7, the single power-law covering the whole range implies that, in this case, there is not a clear distinction between both, long and short-range regimes making even more difficult the identification of strict isochores. The longest, GC-rich contig of chromosome 22 is probably unrepresentative of the GC distribution of the human genome, as it lacks GC-poor DNA and shows a strikingly variegated compositional structure mainly composed of the heaviest (H2, H3) isochores (Bernardi, 2001; Oliver et al., 2001; Pavlicek et al., 2002). Although is doubtful that this contig can be viewed as characteristic of the neogenome (GC rich DNA) which is usually alternating with GC-poorer DNA from the paleogenome along chromosomes (Bernardi, 1989, 2000), its GC richness indicates that it probably belong to the more modern part of the genome appeared along the evolutionary process. If so, scale independence may be a recent acquisition in evolution. The human genome may be thus viewed as composed by an ancestral, short-range correlated part (the paleogenome), derived from cold-blooded vertebrates, and a more recent part, typical of warm-blooded vertebrates, displaying long-range correlations and scale independence (the neogenome). Probably it will be difficult to find many regions with the scale-invariance properties of chromosome 22 because, although GC-rich regions show this property, it is very unlikely to find long GC regions uninterrupted by GC-poor regions.

Using the single-letter rules (A, T, C and G), we found $C(\ell)$ profiles practically identical for A and T on the one hand and also practically identical for C and G on the other. However, in all four cases, $C(\ell)$ resembles what we find with the SW mapping rule (not shown) and, therefore clearly differs from what we find with the RY mapping rule. This could be the origin of the controversy between Peng and co-workers and Voss (Buldyrev et al., 1993a; Voss, 1993): that is, whereas the former use the analysis of variance with the RY mapping rule, the latter uses the power-spectrum analysis with the four single letter rules.

4. Conclusion

We have shown that the study of correlations in DNA sequences using direct measures is now possible thanks to the availability of large sequences of many organisms and thus, many of the drawbacks of previous indirect measures can be overcome. The autocorrelation function, $C(\ell)$, does not require homogeneity for its applicability to DNA sequences and, in addition, $C(\ell)$ can be used as a measure of compositional heterogeneity which is considered to be a

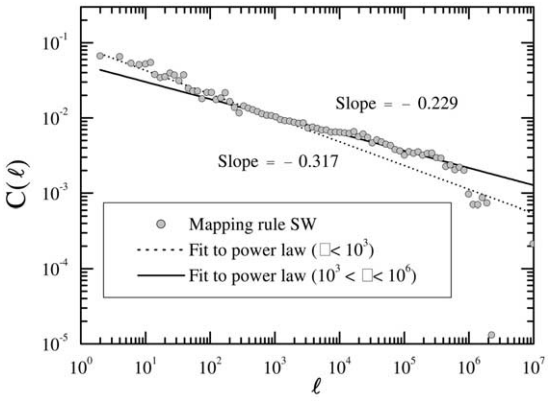

Fig. 9. Autocorrelation function vs. ℓ using the SW mapping rule for the longest contig of human chromosome 20 (NT_011362.4, 24,982,240 bp). Solid line: fit to a power law of the data in the range $10^3 < \ell < 10^6$ (slope $= -0.229$). Dotted line: fit to a power law of the data in the range $1 < \ell < 10^3$ (slope $= -0:317$).

relevant feature of DNA sequences and is closely related to the existence of power-law correlations (Bernaola-Galván et al., 1996; Buldyrev et al., 1993a; Li, 1997; Román-Roldán et al., 1998).

The plots of $C(\ell)$ for prokaryotic genomes show that, at short scales (below the characteristic size of genes) correlations are dominated by the non-uniform base composition in the three codon positions. At larger scales we observe both behaviours, genomes for which $C(\ell)$ almost vanishes (e.g. *M. tuberculosis*) and genomes for which $C(\ell)$ is significantly different from zero in a broad range of sizes (e.g. *B. subtilis*). In the former, the behaviour beyond the characteristic size of genes is similar to what could be observed in a random sequence, thus implying that these genomes are essentially homogeneous at large scales (a commonly accepted idea, Rolfe and Meselson, 1959; Sueoka, 1959). Nevertheless, the latter class of prokaryotic genomes presents correlations implying the presence of heterogeneities that cannot be explained in terms of non-uniform base composition in the three codon positions and could be related to a massive lateral transfer of compositionally biased genes from other genomes or even to natural selection. In addition, we observe power-law correlations in these genomes which, in some cases extend to more than four orders of magnitude, in agreement with previous results (de Sousa Vieira, 1999). Thus, the results obtained for such genomes clearly questions the assumption of homogeneity in prokaryotic DNA.

For human DNA, we have observed that $C(\ell)$, when computed with the RY mapping rule, shows power-law correlations over four orders of magnitude with exponents which are consistent with previous results obtained analysing short sequences (Peng et al., 1994). The size of the analysed sequences (more than 20 Mb) ensures that the fact that these correlations do not extend beyond 10 4 bp is not due to finite size effects.

With respect to $C(\ell)$ computed with the SW mapping rule, we show that the finding of power-law correlations in more than five orders of magnitude in the sequence of chromosome 22 (Holste et al., 2001) is not the common behaviour among the human contigs analysed. The presence of a scale-invariant structure in this chromosome indicated that it probably belongs to the 'neogenome' (Bernardi, 1989, 2000), the more modern part of the genome appearing along the evolutionary process. The behaviour observed in the rest of contigs analysed is compatible with the presence of isochores, long regions with different composition, and 'fairly homogeneous', in the sense that these regions are homogeneous only above a certain scale (Bernardi, 2001; Cuny et al., 1981).

Acknowledgements

We would like to thank Oliver Clay and Wentian Li for useful discussions. We also thank Giorgio Bernardi for his kind invitation to attend the 5th Anton Dohrn Workshop held in Ischia. This work is partially supported by Grant BIO99-0651-CO2-01 from the Spanish Government.

References

Arneodo, A., Bacry, E., Graves, P.V., Muzy, J.F., 1995. Characterizing long-range correlations in DNA sequences from wavelets analysis. Phys. Rev. Lett. 74, 3293–3296.

Arneodo, A., d'Aubenton-Carafa, Y., Audit, B., Bacry, E., Muzy, J.F., Thermes, C., 1998. Nucleotide composition effects on the long-range correlations in human genes. Eur. Phys. J. B 1, 259–263.

Audit, B., Thermes, C., Vaillant, C., d'Aubenton-Carafa, Y., Muzy, J.F., Arneodo, A., 2001. Long-range correlations in genomic DNA: a signature of the nucleosomal structure. Phys. Rev. Lett. 86, 2471–2474.

Bernaola-Galván, P., Román-Roldán, R., Oliver, J.L., 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. Phys. Rev. E 53, 5181–5189.

Bernardi, G., 1989. The isochore organization of the human genome. Annu. Rev. Genet. 23, 637–661.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241, 3–17.

Bernardi, G., 2001. Misunderstandings about isochores. Part 1. Gene 276, 3–13.

Bernardi, G., Olofsson, B., Filipsky, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of warm blooded vertebrates. Science 228, 953–958.

Borštnik, B., Pumpernik, D., Lukman, D., 1993. Analysis of apparent $1/f^{\alpha}$ spectrum in DNA sequences. Europhys. Lett. 23, 389–394.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simons, M., Stanley, H.E., 1993a. Generalized Levy-walk model for DNA nucleotides. Phys. Rev. E 47, 4514–4523.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Simons, M., Sciortino, F., Stanley, H.E., 1993b. Long-range fractal correlations in DNA. Phys. Rev. Lett. 71, 1776.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.-K., Stanley, H.E., Stanley, M.H.R., Simons, M., 1993c. Fractal landscapes and molecular evolution: modeling the myosin heavy chain gene family. Biophys. J. 65, 2673–2679.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.S., Matsa, M.E., Peng, C.-K., Simons, M., Stanley, H.E., 1995. Long-range correlations properties of coding and noncoding DNA sequences: GenBank analysis. Phys. Rev. E 51, 5084–5091.

Carpena, P., Bernaola-Galván, P., Román-Roldán, R., Oliver, J.L., 2002a. A simple and species-independent coding measure. Gene 300, 97–104.

Carpena, P., Bernaola-Galván, P., Ivanov, P.C.h., Stanley, H.E., (2002b) Metal–insulator transition in one-dimensional solids with correlated disorder. Nature 418, 955–959.

Clay, O., 2001. Standard deviations and correlations of CG levels in DNA sequences. Gene 276, 33–38.

Clay, O., Bernardi, G., 2001. Compositional heterogeneity within and among isochores in mammalian genomes. II. some general comments. Gene 276, 25–31.

Cuny, G., Soriano, P., Macaya, G., Bernardi, G., 1981. The major components of the mouse and human genomes: preparation, basic properties and compositional heterogeneity. Eur. J. Biochem. 115, 227–233.

Dandliker, P.J., Holmlin, R.E., Barton, J.K., 1997. Oxidative thymine dimer repair in the DNA helix. Science 275, 1465–1468.

de Sousa Vieira, M., 1999. Statistics of DNA sequences: A low-frequency analysis. Phys. Rev. E 60, 5932–5937.

Häring, D., Kypr, J., 2001. No isochores in the human chromosomes 21 and 22? Biochem. Biophys. Res. Commun. 280, 567–573.

Hausdorff, J.M., Peng, C.-K., 1996. Multiscaled randomness: A possible source of 1/*f* noise in biology. Phys. Rev. E 54, 2154–2157.

Herzel, H., Grosse, I., 1995. Measuring correlations in symbol sequences. Physica A 216, 518–542.

Herzel, H., Grosse, I., 1997. Correlations in DNA sequences: the role of protein coding segments. Phys. Rev. E 55, 800–810.

Holste, D., Grosse, I., Herzel, H., 2001. Statistical analysis of the DNA sequence of human chromosome 22. Phys. Rev. E 64, 041917/1.

Hu, K., Ivanov, P.Ch., Chen, Z., Carpena, P., Stanley, H.E., 2001. Effect of trends on detrended fluctuation analysis. Phys. Rev. E 64, 011114.

Karlin, S., Brendel, V., 1993. Patchiness and correlations in DNA sequences. Science 259, 677–680.

International human genome sequencing consortium, Lander, E.S., Waterston, R.H., Sulston, J., Collins, F.S., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Li, W., 1989. Spatial 1/*f* spectra in open dynamical systems. Europhys. Lett. 10, 395–400.

Li, W., 1990. Mutual information functions versus correlation function. J. Stat. Phys. 60, 823–837.

Li, W., 1991. Expansion-modification systems. A model for spatial 1/*f* spectra. Phys. Rev. A 43, 5240–5260.

Li, W., 1997. The complexity of DNA: the measure of compositional heterogeneity in DNA sequences and measures of complexity. Complexity 3, 33–37.

Li, W., 2001. Delineating relative homogeneous G + C domains in DNA sequences. Gene 276, 57–72.

Li, W., Kaneko, K., 1992. Long-range correlations and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. Europhys. Lett. 17, 555–660.

Li, W., Marr, T.G., Kaneko, K., 1994. Understanding long-range correlations in DNA sequences. Physica D 75, 392–416.

Li, W., Stolovitzky, G., Bernaola-Galván, P., Oliver, J.L., 1998. Compositional heterogeneity within, and heterogeneity between, DNA sequences of yeast chromosomes. Genome Res. 8, 916–928.

Li, W., Bernaola-Galván, P., Haghighi, F., Grosse, I., 2002. Applications of recursive segmentation to the analysis of DNA sequences. Comput. Chem., in press.

Mohanty, A.K., Narayana Rao, A.V.S.S., 2000. Factorial moments analyses show a characteristic length scale in DNA sequences. Phys. Rev. Lett. 84, 1832–1835.

Nekrutenko, A., Li, W.H., 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res. 10, 1986–1995.

Oliver, J.L., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., 2001. Isochore chromosome maps of eukaryotic genomes. Gene 276, 47–56.

Oliver, J.L., Carpena, P., Román-Roldán, R., Mata-Balaguer, T., Mejías-Romero, A., Hackenberg, M., Bernaola-Galván, P., 2002. Isochore chromosome maps of long human contigs. Gene 300, 117–127.

Pavlicek, A., Paces, J., Clay, O., Bernardi, G., 2002. A compact view of isochores in the draft human genome sequence. FEBS Lett. 511, 165–169.

Peng, C.-K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., Stanley, H.E., 1992. Long-range correlations in nucleotide sequences. Nature 356, 168–170.

Peng, C.-K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E., Goldberger, A.L., 1994. Mosaic organization of DNA nucleotides. Phys. Rev. E 49, 1685–1689.

Rief, M., Clausen-Schaumann, H., Gaub, H.E., 1999. Sequence-dependent mechanics of single DNA molecules. Nat. Struct. Biol. 6, 346–349.

Rolfe, R., Meselson, M., 1959. The relative homogeneity of microbial DNA. Proc. Natl. Acad. Sci. USA 45, 1039–1042.

Román-Roldán, R., Bernaola-Galván, P., Oliver, J.L., 1998. Sequence compositional complexity of DNA through an entropic segmentation method. Phys. Rev. Lett. 80, 1344–1347.

Sueoka, N., 1959. A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. Proc. Natl. Acad. Sci. USA 45, 1480–1490.

Teitelman, M., Eeckman, F.H., 1996. Principal component analysis and large-scale correlations in noncoding sequences of human DNA. J. Comput. Biol. 3, 573–576.

Venter, J.C., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.

Voss, R.F., 1992. Evolution of long-range fractal correlations and 1/*f* noise in DNA base sequences. Phys. Rev. Lett. 68, 3805–3808.

Voss, R.F., 1993. Voss replies. Phys. Rev. Lett. 71, 1776.

Voss, R.F., 1994. Long-range fractal correlations in DNA introns and exons. Fractals 2, 1–6.

Weiss, O., Herzel, H., 1998. Correlations in protein sequences and property codes. J. Theor. Biol. 190, 341–353.

Yu, Z.-G., Anh, V.V., Wang, B., 2001a. Correlation property of length sequences based on global structure of the complete genome. Phys. Rev. E 63, 011903.

Yu, Z.-G., Anh, V., Lau, K.-S., 2001b. Measure representation and multifractal analysis of complete genomes. Phys. Rev. E 64, 031903.