Eur. Phys. J. B (2012) 85: 211
DOI: 10.1140/epjb/e2012-20969-5

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Segmentation of time series with long-range fractal correlations

P. Bernaola-Galván[1,a], J.L. Oliver[2], M. Hackenberg[2], A.V. Coronado[1], P.Ch. Ivanov[3,4,5], and P. Carpena[1]

[1] Dpto. de Física Aplicada II, Universidad de Málaga, 29071 Málaga, Spain
[2] Dpto. de Genética, Inst. de Biotecnología, Universidad de Granada, 18071 Granada, Spain
[3] Harvard Medical School, Division of Sleep Medicine, Brigham & Women's Hospital, 02115 Boston, MA, USA
[4] Department of Physics and Center for Polymer Studies, Boston University, 2215 Boston, MA, USA
[5] Institute of Solid State Physics, Bulgarian Academy of Sciences, 1784 Sofia, Bulgaria

Received 28 November 2011 / Received in final form 9 April 2012
Published online 25 June 2012 –

Abstract. Segmentation is a standard method of data analysis to identify change-points dividing a nonstationary time series into homogeneous segments. However, for long-range fractal correlated series, most of the segmentation techniques detect spurious change-points which are simply due to the heterogeneities induced by the correlations and not to real nonstationarities. To avoid this oversegmentation, we present a segmentation algorithm which takes as a reference for homogeneity, instead of a random i.i.d. series, a correlated series modeled by a fractional noise with the same degree of correlations as the series to be segmented. We apply our algorithm to artificial series with long-range correlations and show that it systematically detects only the change-points produced by real nonstationarities and not those created by the correlations of the signal. Further, we apply the method to the sequence of the long arm of human chromosome 21, which is known to have long-range fractal correlations. We obtain only three segments that clearly correspond to the three regions of different G + C composition revealed by means of a multi-scale wavelet plot. Similar results have been obtained when segmenting all human chromosome sequences, showing the existence of previously unknown huge compositional superstructures in the human genome.

1 Introduction

Many phenomena in different fields generate nonstationary time series of data with statistical properties that change under time translations. As a result, the series of data is heterogeneous in the sense that one can obtain different values of the mean, standard deviation or higher moments depending on the time interval where they are calculated. This makes the analysis of such time series more complicated because the validity of many statistical techniques relies on the assumption of stationarity of the analyzed data. In addition, for correlated time series, the presence of such correlations can be easily misidentified as nonstationarities [1]. This situation is specially dramatic for series with long-range fractal correlations, also known as $1/f$-type correlations, where heterogeneities appear at all scales in a self-similar fashion[1] [2–4].

One of the standard methods of analysis of such nonstationary series is the segmentation: given a heterogeneous input series the segmentation procedure divides it into a certain number of non overlapping contiguous pieces called *segments* in such a way that these segments are homogeneous or, at least, more homogeneous than the original data. This problem has been widely studied in Mathematics where it is known as the *change-point problem* [5].

On the one hand, the segmentation can be viewed as a "detrending procedure" in the sense that it can be used to filter out the effects of nonstationarity (e.g. daily periodicities in solar irradiation, sleep-wake differences in heart rate, etc.) and study the more subtle fluctuations that may reveal intrinsic correlation properties of the dynamics of the system under study [6–10]. But, on the other hand, the nonstationarity itself can be also an important feature of the phenomena. For example, it is known that nonstationarity properties of physiological time series can change from healthy to pathological conditions [11–14] and under different physiological states [15–21], in DNA the regions with higher concentrations or densities of a certain dinucleotide (CpG islands) are related to the presence of genes [22–24], the greater the inhomogeneity in the distribution of a word along a text the higher its relevance to the text [25,26], different volatility periods of stock market records are related to

[a] e-mail: rick@uma.es

[1] During last twenty years $1/f$ correlations have been found in practically all fields of Science. Visit http://www.nslij-genetics.org/wli/1fnoise for an udpated bibliographical review.

the expansion-contraction of the economy [27], the distribution of periods with different Internet activity are closely related to the congestion state of the net [28], etc. Usually, the segmentation procedure consists in the partition of a nonstationary series into segments with different mean [14,29–32], although it can be designed also to find regions with different variance [33–36], different correlation properties [37,38] or even with different probability distributions of data [39,40]. Here, we concentrate ourselves on the segmentation based on the mean.

In 2001 some of us proposed a heuristic segmentation algorithm [14] designed to study the distribution of periods with constant heart-rate which has been also applied to the detection of climate changes [41,42], to study the large scale structure of DNA sequences [43–46] and to search for periods with different Internet activity [28]. This algorithm (see later for a complete description) iteratively divides the series into segments with mean values that are *significatively* different from the mean values of adjacent segments. The iteration ends when none of the segments can be further divided into subsegments with significatively different means. In [14] we considered that the difference between the means of two adjacent segments (i.e. the change-point) is statistically significative if the probability of obtaining such a difference just by chance in a random i.i.d. series is less than a given value – typically 5%. This criterion, implicitly or explicitly, has been widely used and is equivalent to consider the random i.i.d. series as the reference for homogeneity. The problem is that, for "real-life data" a random i.i.d. series is a too restrictive model for homogeneity. Several approaches have been proposed to overcome this problem: Oliver et al. proposed to filter out the short scale heterogeneities [43], Thakur et al. [47] have developed an algorithm aimed at segmenting symbolic sequences using Markov models as reference for homogeneity and, more recently, Toth et al. [48] proposed the segmentation of certain economic time series using compound Poisson processes to model the homogeneous segments.

However, as commented above, many physical phenomena can be described in terms of self-similar models: fractional Gaussian noise (fGn), fractional Brownian motion (fBm), ARFIMA, etc. [49,50], which are, at first sight, much more heterogeneous than a random i.i.d. series and, in addition, show heterogeneities at all scales (see Fig. 1). Thus, the application of the above referred segmentation algorithm [14] to series with such correlations will lead to the detection of change-points in the mean which are simply due to the presence of correlations in the data. Although these change-points are indeed present in the series and could be interesting for the description of the phenomena [51], the ability to discern between them and those which cannot be attributed to correlations would be of valuable help to obtain insights into the dynamics responsible for the generation of the observed data. In fact, the detection of change-points in the presence of long-range correlations has been studied in several fields [29–32].

Here, we use these ideas to design a modified version of the segmentation algorithm introduced in [14] that takes as the reference for homogeneity, instead of a random i.i.d. series, a correlated series modeled by a fractional noise with the same degree of correlations as the series to be segmented.

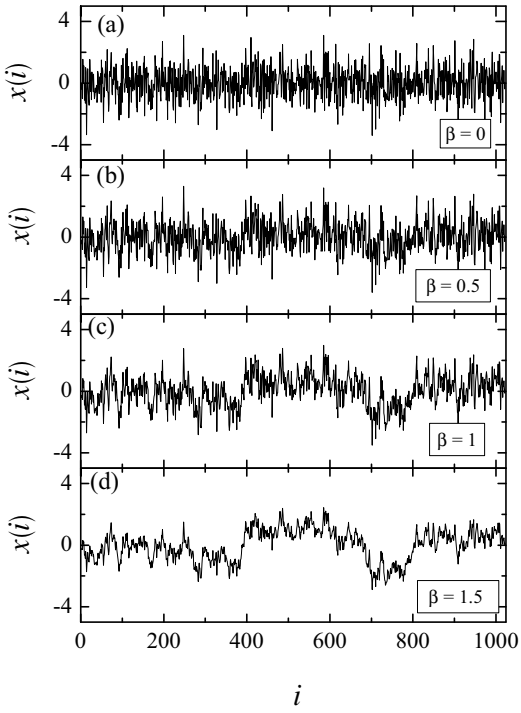

Fig. 1. Several examples of fractional noise generated using the inverse Fourier filtering method [54]. The method is a direct application of fractional noise definition in equation (1): first, we generate random uncorrelated and Gaussian distributed series $\eta(i)$ and calculate its Fourier transform coefficients $\hat{\eta}(q)$ using FFT. To obtain the series with the desired power-law exponent β in the power spectrum, we simply divide $\hat{\eta}(q)$ in the Fourier space by $q^{\lambda} = q^{\beta/2}$ and transform it back using the inverse FFT. Taking into account that both direct and inverse Fourier transforms are computed using the FFT algorithm, the sizes of the generated signals should be always integer powers of 2. The signals plotted here have 1024 data points each, and all of them have been generated starting from the same series of Gaussian noise. (a) White noise ($\beta = 0$), (b) fractional Gaussian noise with $\beta = 0.5$, (c) $1/f$ noise ($\beta = 1$) and (d) fractional Brownian motion with $\beta = 1.5$.

This article is structured as follows: in Section 2 we introduce the fractional noise which is the model that we adopt for long-range correlated series. In Section 3 we describe two segmentation algorithms that use as a reference for homogeneity the fractional noise: the heuristic segmentation (Sect. 3.1) which is the direct generalization of the algorithm proposed in [14] and a new version of the "optimal segmentation" that includes a criterion to decide the correct number of segments (Sect. 3.2). Section 4 describes the calculation of the significance level using the fractional noise as the reference for homogeneity. In Section 5 we present examples of the segmentation of artificial series as well as of DNA sequences. Finally, the conclusions section ends the article.

2 Reference for homogeneity

As we stated in the introduction, when segmenting a time series with long-range correlations, we have to use as a reference for homogeneity a long-range correlated random series with the same degree of correlations as the target time series. The model for long-range correlated random series that we adopt here is the fractional noise: a family of models that can be obtained through fractional integrations or derivations [52,53] of a white noise – a random uncorrelated Gaussian series.

Consider a white noise $\eta(i)$, i.e. a series of uncorrelated Gaussian-distributed numbers all with the same mean and variance and let $\hat{\eta}(q)$ be its Fourier transform. Given a real number λ the fractional noise of order λ, $\eta_\lambda(i)$ is defined, up to a multiplicative constant, as:

$$\eta_\lambda(i) \equiv \mathcal{F}^{-1}\left[\frac{\hat{\eta}(q)}{q^\lambda}\right], \tag{1}$$

where $\mathcal{F}^{-1}[\cdot]$ denotes the inverse Fourier transform.

Having in mind the properties of the Fourier transform, positive values of λ in equation (1) could be interpreted as "integrations" of order λ (not necessarily integer) of $\eta(i)$, and negative values of λ as "derivatives" of order $-\lambda$.

In particular, the fractional Gaussian noise (fGn) corresponds to $\lambda < 0.5$ and the fractional Brownian motion (fBm) to $\lambda \geq 0.5$, including the $1/f$ noise ($\lambda = 0.5$) [50].

From the definition (1) it is clear that a fractional noise shows an inverse power-law behavior in its power-spectrum with exponent $\beta = 2\lambda$ and, for this reason, they are also known as $1/f^\beta$ noises.

In Figure 1 we plot several examples of fractional noise generated using the inverse Fourier filtering method [54] in order to show how correlations affect the heterogeneity of the signals: the greater the exponent the more heterogeneous the signal appears. In fact, signals with $\beta \geq 1$ are *not* stationary because they do not have well defined mean [49]. In this particular example, specially for the signals corresponding to $\beta = 1$ and $\beta = 1.5$, one can easily identify three or four clear regions with different mean values. Nevertheless, if we consider the fractional noise as our reference for homogeneity, these regions should not be identified as segments by the segmentation algorithm because they are due to the normal fluctuations present in the signal as a direct consequence of its correlations.

Having in mind the reference for homogeneity that we propose here, the first step in our segmentation algorithm will be the measurement of the correlations in the original series we want to segment, in order to select the model of fractional noise that best fits these correlations. The direct way of doing this is to compute the power spectrum of the signal and to fit it to a straight line in a double logarithmic plot. The slope of the line gives the exponent β. Nevertheless, in practice, the plots of the power spectrum are quite noisy and the estimation of the exponent β is not easy.

Thus, instead of the power spectrum procedure to estimate the correlation exponent, here we use the detrended fluctuation analysis (DFA), a scaling analysis developed by Peng et al. [55]. This method was specially designed to work with nonstationary series and also provides a single quantitative parameter – the scaling exponent α – to represent the correlation properties of a long-range correlated series (see Appendix A). The performance of the DFA method has been extensively studied for time series with different trends and nonstationarities [56,57], under conditions of dataloss [58], and after application of nonlinear filters and coarse-graining of the data [59,60], and has been favorably compared to other detrended moving average techniques [61].

3 Segmentation algorithms

3.1 Heuristic search for the change-points

First, we describe the heuristic procedure proposed in [14] to find segments with different mean in numerical series. This algorithm is a modified version of a previous one designed to segment symbolic sequences into regions of different composition [40,62].

To divide a nonstationary series $\mathcal{S} = \{x_1, x_2, \ldots, x_N\}$ of length N into stationary segments of constant mean we start moving a sliding pointer from left to right along the series and, at each position j we consider the two subseries $\mathcal{S}_1 = \{x_1, x_2, \ldots, x_j\}$ to the left of the pointer and $\mathcal{S}_2 = \{x_{j+1}, x_{j+2}, \ldots, x_N\}$ to the right. Their means are given by:

$$\mu_1 = \frac{1}{N_1}\sum_{x_i \in \mathcal{S}_1} x_i, \quad \mu_2 = \frac{1}{N_2}\sum_{x_i \in \mathcal{S}_2} x_i \tag{2}$$

where $N_1 = j, N_2 = N - j$.

As a measure of the difference between both means we use the Student's t-statistics:

$$t(\mathcal{S}_1, \mathcal{S}_2) \equiv \left|\frac{\mu_1 - \mu_2}{\sqrt{\sigma_P}}\right|, \tag{3}$$

where σ_P is the pooled variance [63]:

$$\sigma_P = \frac{N\left[V(\mathcal{S}_1) + V(\mathcal{S}_2)\right]}{(N-2)N_1N_2}, \tag{4}$$

and $V(\mathcal{S})$ is the sum of squared deviations of the data in $\mathcal{S}$:

$$V(\mathcal{S}) = \sum_{x_i \in \mathcal{S}} (x_i - \mu)^2. \tag{5}$$

In this way, we obtain t as a function of the position in the time series, j, and select as a candidate for the change-point the position $j_{\max}$ where $t(j)$ reaches its maximum value $t_{\max}$.

Next, we determine the statistical significance of $t_{\max}$. To this end we consider the following probability distribution:

$$\mathcal{P}_{\beta,N}(\tau) = \text{Prob}\left\{\max[t(j)] \leq \tau \mid \mathcal{S}_0(\beta, N)\right\}, \tag{6}$$

i.e. the probability of obtaining values of the maximum value of the t-Student's statistics smaller or equal than

τ when trying to segment a series $\mathcal{S}_0$ of size N generated with a fractional noise model with exponent β. In the next sections we will discuss on how to obtain $\mathcal{P}_{\beta,N}(\tau)$.

Larger values of $\mathcal{P}_{\beta,N}(t_{\max})$ imply that it is less likely to obtain high $t_{\max}$ values just due to chance alone. In Mathematics,

$$p(t_{\max}) \equiv 1 - \mathcal{P}(t_{\max}) \tag{7}$$

is called a p-value. It can be interpreted as the probability that the null hypothesis (H_0) is true. In our case, H_0 is that the observed $t_{\max}$ value can be obtained in a series S_0 of fractional noise. We reject H_0 if the p-value is smaller than a given threshold p_0 (usually 0.05) accepting thus the alternative hypothesis H_1 that the observed $t_{\max}$ is higher than it could be expected to occur within a random series of fractional noise. The acceptance of the alternative hypothesis H_1 entails the acceptance of $j_{\max}$ as a change point, i.e. the series is cut at position $j_{\max}$ into two segments. If H_0 is not rejected the series remains uncut. If the series is cut, the procedure continues recursively inside each of the two resulting subseries created by each cut.

Before a new cut is accepted, we also compute t between the right-hand new segment and its right neighbor (obtained by a previous cut) and the t between the left-hand new segment and its left neighbor (also obtained by a previous cut) and check if both values of t have p-values smaller than p_0. If so, we proceed with the new cut; otherwise we do not cut. This ensures that all resulting segments have a statistically significant difference in their means. The process stops when none of the possible change-points verify $p(t_{\max}) \leq p_0$, and we say that the series has been segmented at the "significance level p_0" (see Fig. 2 of [14] for an illustrative example of the procedure).

Note that the distribution used in previous versions of the algorithm [14,28,41–46,51] to compute the p-value was $\mathcal{P}_{0,N}(\tau)$, i.e. the particular case for $\beta = 0$ which corresponds to use a random i.i.d. series as the reference for homogeneity.

The strategy described above to decide whether a new cut is accepted or not is known as *hypothesis testing*. Although this strategy is the most widely used in segmentation problems, it is not unique. An alternative way to address this problem is the *model selection* strategy, where segmentation is viewed as the selection between two models describing the target sequence: with and without the cut [64,65]. Although both strategies look different, they are quite similar and it has been shown that, in some cases, they are strictly equivalent [40,66].

To demonstrate the effect of correlations on the segmentation algorithm, in Figure 2 we plot Student's t-statistics as a function of the position of the pointer j ($t(j)$) for the same series of fractional noise shown in Figure 1. The qualitative behavior of the profiles is similar for all of them because all series have been generated starting from the same series of Gaussian white noise. In fact, all profiles reach their maxima at the same values of j, around $j = 400$. In all cases this maximum appears as a consequence of the statistical fluctuations (note how it is present even in the white noise signal $\beta = 0$) but the correlations, in some sense, amplify this effect.

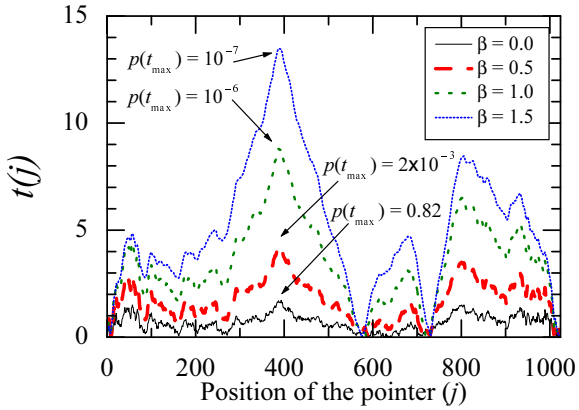

Fig. 2. (Color online) Student's t-statistics vs. the position of the pointer (see text) for the series of fractional noise shown in Figure 1. The p-values of each maximum computed using the random i.i.d. ($p(t_{\max}) = 1 - \mathcal{P}_{0,N}(t_{\max})$) as the reference for homogeneity [14], are specified close to each curve. Note that, despite of their stationarity, the original algorithm would cut the signals with $\beta = 0.5, 1$ and 1.5 at practically any significance level p_0. Nevertheless, if we compute the p-value using the fractional noise as the reference for homogeneity, i.e. using the correct value of β, we obtain $p(t_{\max}) = 0.82$ for all four maxima in this example. This means that none of the series will be segmented at the usual significance levels, namely 0.01, 0.05 or 0.1.

3.2 The optimal segmentation

The procedure described in the previous section is fast and it performs in time proportional to $\mathcal{O}(N \log k)$ where N is the length of the series and k the number of cuts, and also gives good results as compared to other segmentation methods [67].

Nevertheless it has certain limitations. For example, as we already pointed out in [14], in the case where a long homogeneous segment is interrupted by a short segment with a different mean, the heuristic algorithm could fail to detect it since when trying to cut at the beginning or the end of the small segment there is no much difference in the mean at both sides of the pointer, since the mean is mainly controlled by the two large flanking segments which have the same mean. Moreover, when segmenting a series composed of segments of similar size and alternating mean values, the algorithm could fail if the number of alternating segments is high even if the difference between the means of adjacent segments is significant.

In order to overcome these problems, we will adopt a different approach: first, we decide the number of change points k we are looking for (see later for a discussion on this issue) and then we check all their possible positions and look for the set of positions maximizing a certain objective function. This procedure is usually called *the optimal segmentation*. In principle, the computation time of this algorithm seems to scale as $\mathcal{O}(N^k)$ which would

make it unfeasible in practice. However, using dynamic programming it is possible to obtain algorithms with running times proportional to $\mathcal{O}(N^2)$ [68] which could be reasonable, at least for series of no more than few hundred thousands data points. Note that the search for a change-point described in the previous section is a particular case of the optimal segmentation with $k = 1$ and with the Student's t-statistics being the objective function.

Given a series S, consider k change points which divide it into $k+1$ segments $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{k+1}$. As we want to find segments with different means, we propose to find the k change-points that maximize the following objective function:

$$\Delta V = V(\mathcal{S}) - \sum_{j=1}^{k+1} V(\mathcal{S}_j), \tag{8}$$

where $V(\cdot)$ is defined in (5).

The function ΔV measures the reduction of squared deviations when the series is divided into $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{k+1}$. ΔV is always positive and increases monotonously with the number of change-points k. In some sense, it measures how well the series is described as set of $k+1$ segments with different means $\mu_1, \mu_2, \ldots, \mu_{k+1}$ as compared to the series described as a single segment of mean μ. A similar measure based on Shannon entropy was used for the segmentation of symbolic sequences [40,62] and to describe their complexity [69–71].

Although this approach seems to be quite different to the heuristic segmentation described above, it can be seen (Appendix B) that for $k = 1$ both the search for a change-point with Student's t-statistics and with ΔV are almost equivalent, i.e. the search for the point which divides the series into two subseries with mean values as much different as possible is equivalent to search for the point which divides the series into two segments with the smallest total variance.

Here is important to recall that the number of segments in the optimal segmentation is an input parameter of the algorithm. Given the number of change-points, k, the optimal segmentation obtains the best possible segmentation of the series into $k+1$ segments. However, in many cases we do not have prior information about the number of change-points we are looking for[2]. Thus, we should perform the optimal segmentation for any possible values of k. However, the problem arises when trying to decide which one of the optimal segmentations, each with a different number of segments, is the most appropriate. Note that ΔV does not provide an objective criterion to select the best segmentation since ΔV is always an increasing function of k.

We propose here to select among all possible optimal segmentations the one with the maximum number of change-points, $k_{\max}$, provided that all of them have a p-value below a certain threshold p_0. Using the notation introduced in (3) and (7) this means that:

$$p\left[t(\mathcal{S}_j, \mathcal{S}_{j+1})\right] = 1 - \mathcal{P}_{\beta,N}\left[t(\mathcal{S}_j, \mathcal{S}_{j+1})\right] \leq p_0 \quad \text{for} \quad j = 1, 2, \ldots, k_{\max}, \tag{9}$$

where β is the power spectrum exponent of the series and $N = N_j + N_{j+1}$ the size of the series obtained by linking together $\mathcal{S}_j$ and $\mathcal{S}_{j+1}$.

Nevertheless, this approach present a potential drawback: it may happen that for a given number of change-points k_1 not all of them fulfill (9) but if we continue with larger values of k we could obtain that for some $k_2 > k_1$ all change-points are statistically significant. Thus, to be sure that we obtain the maximum number of statistically significant change-points we should check for all possible values of $k = 1, 2, \ldots, N$. In practice, this segmentation is almost unavoidable because it will lead to computing times proportional to $\mathcal{O}(N^3)$. In general, although it is not possible to ensure whether or not this "return to significance" occurs, it may happen in several cases of interest after a few new cuts. In particular we have observed this behavior in the two examples referred at the beginning of Section 3.2: (i) a long homogeneous segment interrupted by a short segment with a significantly different mean. Here the first cut tries to divide the sequence at the beginning or the end of the short segment and, in many cases, this cut is not significant since the left and right subsequences present a very similar mean. The return to significance takes place when the algorithm tries to give two cuts which would appear at the beginning and the end of the small segment, both of them statistically significant. (ii) A series composed of n segments of similar size and alternating mean values. Here the cuts become significant, in the worst case, when $k = n - 1$. Although we have not studied systematically this problem, these preliminary results lead us to suggest the following strategy: compute the optimal segmentations for $k = 1, 2, 3, \ldots$ and when we find that not all change-points are statistically significant, instead of stopping we continue for several more cuts (Δk) to give a chance to recover the significance. The greater Δk values the more chances to obtain the correct $k_{\max}$ but with a subsequent increase of the computing time.

4 Significance level of the change-points

In both algorithms described above, we need to calculate the p-value of the change-points, in the former to decide when to stop the recursive segmentation and in the latter to decide whether or not all the change-points of an optimal segmentation are statistically significative.

As the probability $\mathcal{P}_{\beta,N}(\tau)$ (6) does not seem to admit a closed analytical form even for the simplest case of uncorrelated noise ($\beta = 0$)[3], we have obtained it by means

[2] In reference [67] the authors show an example in which the heuristic segmentation achieves better results than the optimal segmentation because they make a wrong a priori choice of the number of change-points. Note that the heuristic one does not require any assumption about the number of change-points.

[3] Note that even for an uncorrelated Gaussian noise ($\beta = 0$) $\mathcal{P}_{\beta,N}(\tau)$ is not the Student's t-distribution because the value $t_{\max}$ was not obtained by comparing to independent samples of Gaussian noise but maximizing the difference along a set of non independent samples – all possible left and right subseries.

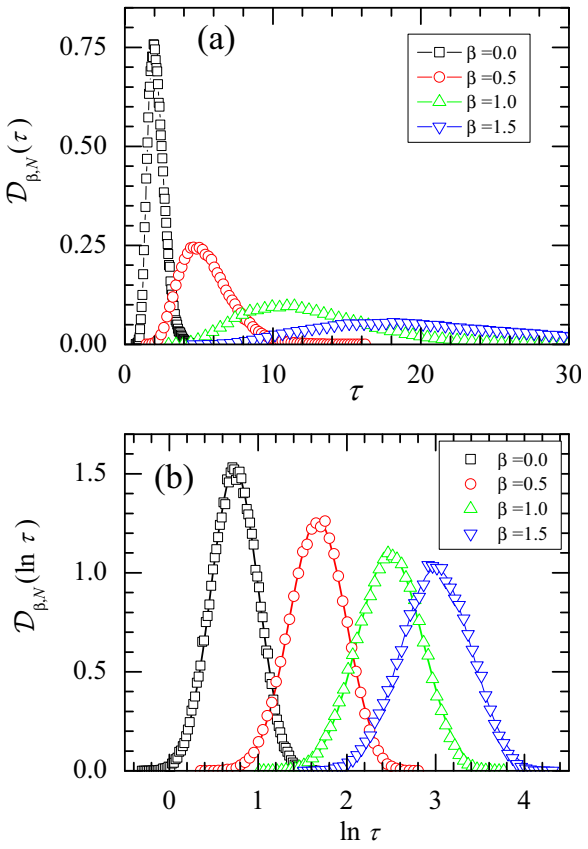

Fig. 3. (Color online) (a) Density histograms of τ obtained by means of numerical simulations for $N = 1024$ and $\beta = 0\,(\square)$, $\beta = 0.5\,(\bigcirc)$, $\beta = 1\,(\triangle)$ and $\beta = 1.5\,(\triangledown)$. The solid lines correspond to log-normal distributions with the same mean and standard deviation as the normalized histograms of τ. (b) Histograms of $\ln\tau$ for the same simulations. Now the solid lines correspond to Gaussian distributions with the same mean and standard deviation as the normalized histograms of $\ln\tau$.

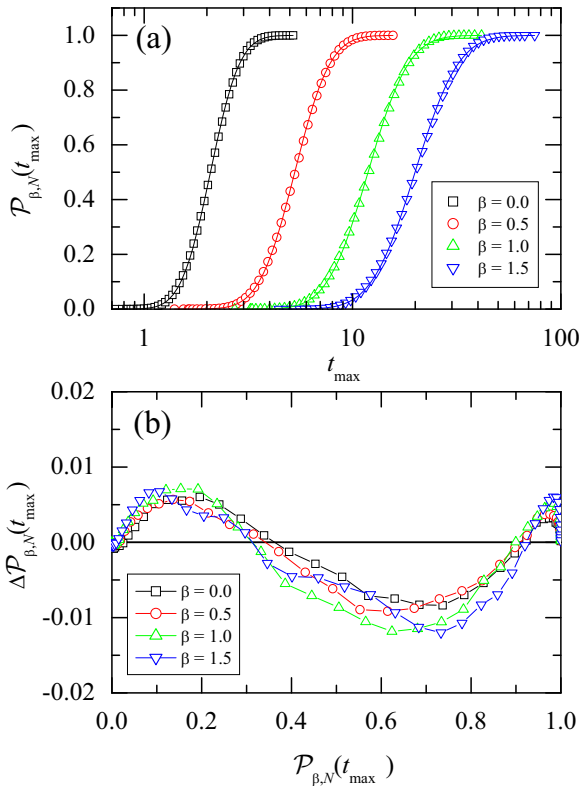

Fig. 4. (Color online) (a) Cumulative histograms of τ in log-linear scale for size $N = 1024$ and $\beta = 0\,(\square)$, $\beta = 0.5\,(\bigcirc)$, $\beta = 1\,(\triangle)$ and $\beta = 1.5\,(\triangledown)$, obtained from numerical simulations. The solid lines are the corresponding log-normal distributions with the same mean and standard deviation of the experimental data. (b) Difference between the log-normal fit and the real histogram obtained from the simulations. Note that, in the worst cases, the error is around 1%.

of numerical simulations. For a given size N and a given value of the correlation exponent β, we generate an ensemble of 10^5 series of fractional noise using the inverse Fourier filtering method (Fig. 1). For each series, we move a pointer along it and obtain $t_{\max}$ (see Sect. 3.1). Finally, for each ensemble of 10^5 series we obtain the histogram $\mathcal{P}_{\beta,N}(\tau)$.

Figure 3a shows the density histograms $\mathcal{D}_{\beta,N}(\tau)$ for $N = 1024$ and $\beta = 0$, 0.5, 1 and 1.5. Note that the histograms are shifted to greater values of τ as β increases in agreement with the fact that correlations increase the heterogeneity of the series. Figure 3b shows the density histograms of $\ln\tau$ for the same experiments.

We observe that the histograms of $\ln\tau$ can be well fitted by normal distributions (Fig. 3b). This means that the original density histogram of τ can be well fitted by a log-normal distribution (Fig. 4a):

$$\mathcal{D}_{\beta,N}(\tau) \simeq \frac{1}{\tau\sqrt{2\pi}\sigma_{\ln\tau}} \exp\left[-\frac{(\ln\tau - \mu_{\ln\tau})^2}{2\sigma^2_{\ln\tau}}\right] \qquad (10)$$

where $\mu_{\ln\tau}$ and $\sigma_{\ln\tau}$ are the mean and standard deviation of $\ln\tau$ respectively which, in general, will depend on N and β. For these examples, the differences between $\mathcal{P}_{\beta,N}(\ln\tau)$ and the corresponding normal distributions with the same mean and standard deviation are around 0.01 (1%) in the worst cases (Fig. 4b). This agreement between $\mathcal{P}_{\beta,N}(\ln\tau)$ and the normal distribution has been systematically observed for series lengths ranging form $N = 256$ to 524 288 (2^8 and 2^{19} respectively) and correlation exponents from $\beta = 0$ to 1.6 (Fig. 5). Tables with values of $\mu_{\ln\tau}$ and $\sigma_{\ln\tau}$ for several values of N and β obtained by simulating series of fractional noise are available in[4].

According to these results, we could characterize approximately $\mathcal{P}_{\beta,N}(\tau)$ by log-normal distributions and estimate the p-values with an error which, even in the worst case, is well below 2% (Fig. 5). To do this, given the value of $t_{\max}$ obtained when trying to segment a series of length N with a correlation exponent β, first we interpolate in the tables[4] to obtain the values of $\mu_{\ln\tau}(\beta, N)$ and $\sigma_{\ln\tau}(\beta, N)$ and then, evaluate the p-value by integrating

[4] http://jander.ctima.uma.es/fractalseg, or http://bioinfo2.ugr.es/segmentLRC/.

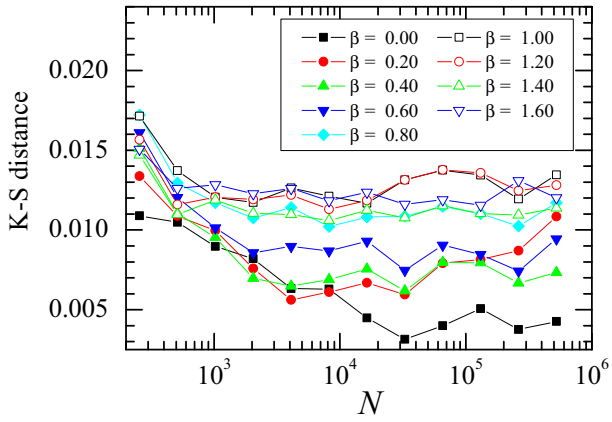

Fig. 5. (Color online) Kolmogorov-Smirnov (K-S) distance between $\mathcal{P}_{\beta,N}(\ln\tau)$ and the normal distribution with the same mean and standard deviation as a function of N for different correlation exponents, β. The K-S distance between two probability distributions $\mathcal{P}$ and $\mathcal{Q}$ is defined as the maximum of the absolute value of the differences between them, $\max|\mathcal{P}(x) - \mathcal{Q}(x)|\ \forall x$. Thus, the K-S distance can be interpreted as the maximum error committed in the evaluation of the probability of a variable following the probability distribution $\mathcal{P}$ if we compute such probability using a wrong distribution $\mathcal{Q}$, or vice-versa. In our case, the K-S distance, gives the maximum error in the evaluation of the p-values when using the log-normal approximation. As can be seen, in all cases this error is well below 0.02 (i.e. 2%).

equation (10):

$$p(t_{\max}) = 1 - \mathcal{P}_{\beta,N}(t_{\max}) = \int_{t_{\max}}^{\infty} \mathcal{D}_{\beta,N}(\tau)d\tau$$

$$\simeq \int_{\ln t_{\max}}^{\infty} \frac{1}{\tau\sqrt{2\pi}\sigma_{\ln\tau}} \exp\left[-\frac{(\ln\tau - \mu_{\ln\tau})^2}{2\sigma^2_{\ln\tau}}\right] d\tau. \quad (11)$$

Nevertheless, in practical applications of the segmentation algorithm we are not interested to know the exact p-value of a given $t_{\max}$ but simply to know whether it lies below a certain threshold p_0 or not. This means that we do not need the full distribution $\mathcal{P}_{\beta,N}(\tau)$ but only its percentile corresponding to the selected p_0 value, i.e. the value t_0 for which the following equation holds:

$$p_0 = p(t_0) = 1 - \mathcal{P}_{\beta,N}(t_0). \quad (12)$$

Indeed, taking into account that $\mathcal{P}_{\beta,N}(t_0)$ is a monotonously increasing function of t_0, checking that the p-value of $t_{\max}$ is below p_0 is equivalent to checking that $t_{\max} \geq t_0$.

For this reason, we have obtained directly from the simulated data (without any approximation) the percentiles for the most usual values of p_0, namely 0.1, 0.05 and 0.01 which correspond to the percentiles 90, 95 and 99% respectively. In Figure 6 we show the 95th percentile of $\mathcal{P}_{\beta,N}(\tau)$ as a function of N for several values of β.

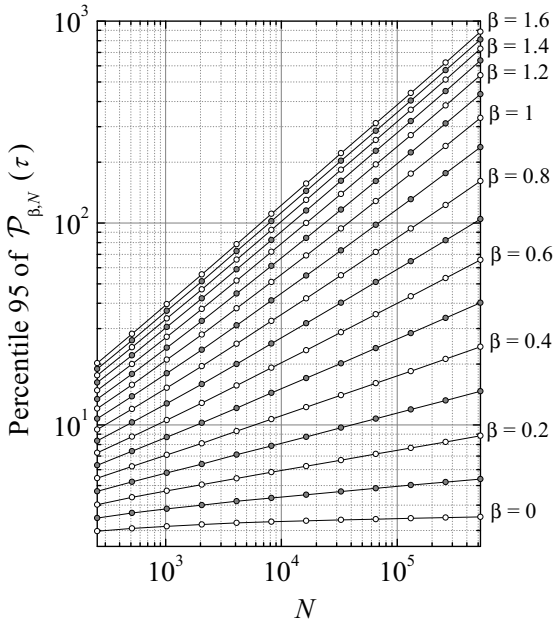

Fig. 6. Percentile 95 of $\mathcal{P}_{\beta,N}(\tau)$ for different sizes N and different values of the correlation coefficient β. In[4] can be found the data displayed here as well as the data corresponding to percentiles 90 and 99%. Note that, for each β, the percentiles as a function of N can be well fitted by power laws with exponents increasing with β.

5 Results

In this section we show two examples of the application of our segmentation algorithms: an artificial series with long-range correlations and a human DNA sequence which is known to have also long-range fractal correlations.

5.1 Artificial series

In order to demonstrate the advantages of the method presented here, we show the segmentation of an artificial series generated by linking together two series of fractional noise with $\beta = 0.6$ and different means (Fig. 7). In this example, the heterogeneity due to a real nonstationarity (two subseries with different means) is blended together with those heterogeneities which are simply due to the correlations introduced in the series.

If we segment the series using the random i.i.d. as the reference for homogeneity, i.e. compute the p-value with $\mathcal{P}_{0,N}(\tau)$, (Fig. 7a) we obtain tens of segments with both the heuristic (Sect. 3.1) and the optimal segmentation (Sect. 3.2). Note that this series is made up of only two stationary patches and thus, the segments with different means found by the algorithm are simply due to the presence of correlations. At this point, note also that the optimal segmentation (dotted line in Fig. 7a) detects several segments that are not unveiled by the heuristic one. As we already commented in Section 3.2, in most cases these segments are small pieces located in the middle of

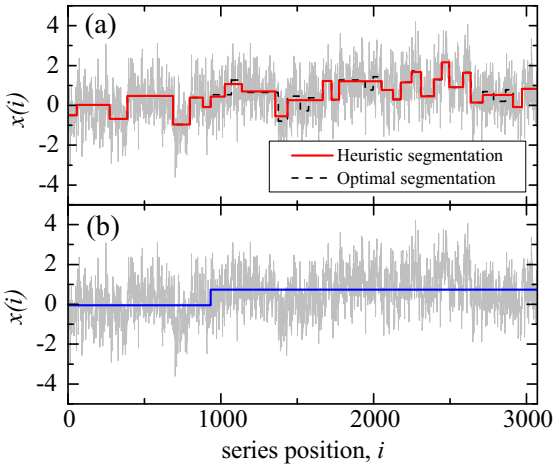

Fig. 7. (Color online) Example of the segmentation of an artificial series generated by linking together two series of fractional noise with $\beta = 0.6$ and sizes $N_1 = 1024$ and $N_2 = 2048$ respectively, both with unit standard deviation but with different means, $\mu_1 = 0$ and $\mu_2 = 0.75$. (a) Segmentation at $p_0 = 0.05$ significance level using the random i.i.d. as reference for homogeneity (i.e. $p(t_{\max}) = 1 - \mathcal{P}_{0,N}(t_{\max})$). The solid line correspond to the segments obtained with the optimal segmentation and the dotted line to the segments obtained with the heuristic segmentation. (b) The same segmentation but using the fractional noise as a reference for homogeneity. Note that, while in (a) all significant change-points in the series are detected, in (b) only the one which cannot be attributed to the correlations is accepted as a valid change-point.

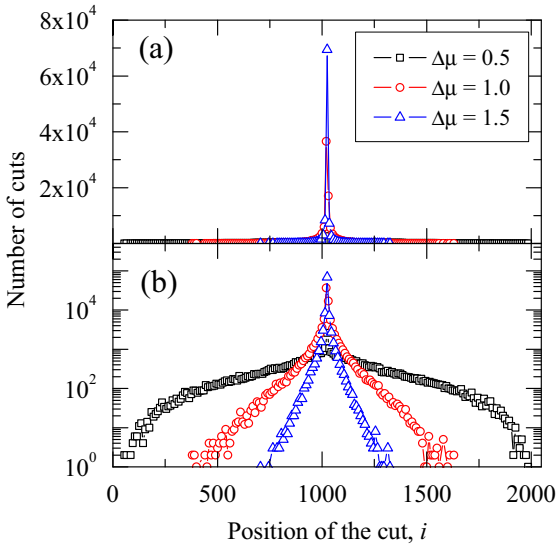

Fig. 8. (Color online) Distribution of the positions of the change-points detected by the segmentation algorithm ($p_0 = 0.05$) in a set of 10^5 artificial series obtained each by linking together two series of fractional noise with $\beta = 0.6$ and sizes $N_1 = N_2 = 1024$, both with unit standard deviation but with different means $\mu_1 \neq \mu_2$. We repeat the experiment for different values of $\Delta\mu = \mu_1 - \mu_2$. If the algorithm gives more than one cut we consider the closest to the real change-point – i.e. $i = 1024$. (a) Linear-linear scale (b) Linear-log scale.

larger ones or series of segments with alternating mean values.

On the contrary, if we take into account the correlations present in the series and use as reference for homogeneity the fractional noise, then we have to compute the p-values using $\mathcal{P}_{\beta,N}(\tau)$ with $\beta = 0.6$. In this case (Fig. 7b) we obtain only a single change-point which divides the series into two segments which approximately correspond to those used to create the artificial series, i.e. our algorithm detects only the change-point produced by a real nonstationarity and not those created by the correlations of the signal. In this case, both the optimal and the heuristic algorithms give the same result. Note also that the algorithms work quite well taking into account the small difference in mean between both segments, as compared to the fluctuations due to the correlations. Actually, in this example, it is hard to identify by eye the location of the change point.

We have systematically repeated this experiment in order to check both the precision of the segmentation algorithms to locate the correct positions of the real change-points as well as their ability to avoid the detection of spurious change-points which are simply due to the correlations. We generate artificial series by linking together two series of fractional noise, both with unit standard deviation but with different means, $\mu_1 \neq \mu_2$, and segment them using the heuristic algorithm. We generate 10^5 time series for each $\Delta\mu \equiv \mu_1 - \mu_2$ value.

We observe that the majority of the cuts detected by the algorithm are close to the real change-point – i.e. close to the midpoint of the series. Indeed, the plots of the histograms of cut positions always show a clear peak centered at the correct position for different values of $\Delta\mu$ (Fig. 8). This peak appears even for $\Delta\mu = 0.5$, a value which is considerably smaller than the standard deviations of the adjacent segments, i.e. considerably smaller than the internal fluctuations inside them. Note that in the example of Figure 7 it was hard to identify by eye the position of the change point although in that case $\Delta\mu = 0.75$.

For the same set of 10^5 artificial series, we also consider the number of cuts detected in the segmentation of each of them (Fig. 9). We observe that: (i) in most cases, the segmentation gives only one cut which is very close to the correct position (Fig. 8). (ii) Only for small values $\Delta\mu$ the fraction of series that remains undivided becomes relevant and the change-point can disappear into the fluctuations produced by the correlations (e.g. $\Delta\mu = 0.5$ in Fig. 9). Nevertheless, as $\Delta\mu$ increases, this fraction drops very fast even for $\Delta\mu < 1$. (iii) There is a small fraction of segmentations with more than one cut due to the statistical nature of the algorithm. This fraction will depend on the significance level p_0 chosen for the segmentation: The higher p_0 the greater the fraction of additional cuts.

Up to now we have considered the simplest case in which the standard deviation and the correlation exponent β remain constant along the segmented signal. To check the performance of the method when dealing with more complex time series we carry out an experiment similar to

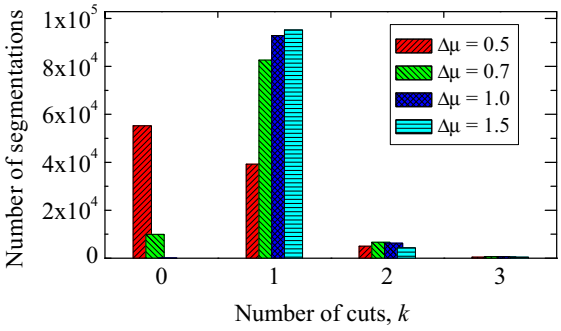

Fig. 9. (Color online) Distribution of the number of cuts (k) detected in each series by the segmentation algorithm ($p_0 = 0.05$) in a set of 10^5 artificial series obtained each by linking together two series of fractional noise with $\beta = 0.6$ and sizes $N_1 = N_2 = 1024$, both with unit standard deviation but with different means $\mu_1 \neq \mu_2$. We repeat the experiment for different values of $\Delta\mu = \mu_1 - \mu_2$.

the previous one, but now we obtain the artificial series by linking together two series of fractional noise with different means and also with: (i) different standard deviation σ; (ii) different correlation exponent β.

We observe in both cases that, again, the majority of the cuts detected by the algorithm are properly located (Fig. 10a): the distribution of the positions of the cuts remains centered at the correct position and its width does not increase with respect to the ideal situation of uniform β and σ. The most remarkable change is the asymmetry of the distribution. This is due to the fact that, in both experiments, the right half of the signal is more heterogeneous than the left one ($\sigma_2 > \sigma_1$ or $\beta_2 > \beta_1$) and thus, spurious cuts are more likely to happen in the right half. We also checked that the fraction of segmentations for which the algorithm detects the correct number of segments is very high and quite similar to that obtained with uniform σ and β (Fig. 10b).

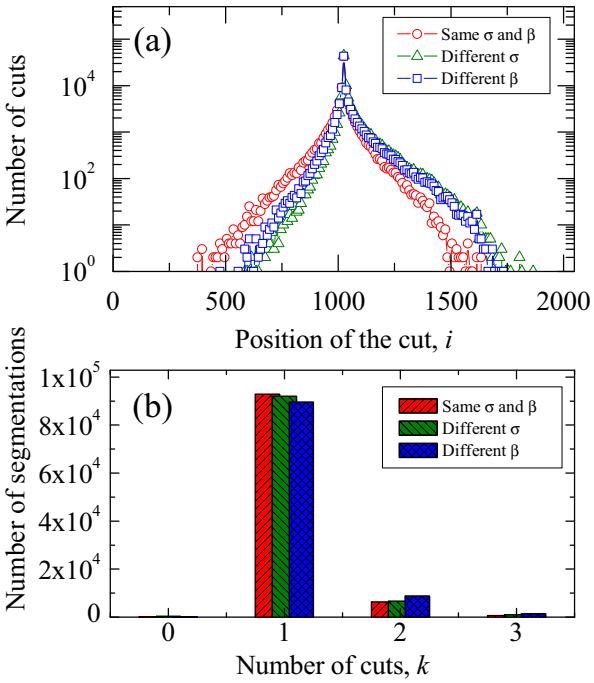

Fig. 10. (Color online) Effects of nonuniform standard deviation (σ) and correlation coefficient (β) on the precision of the segmentation. We generate an ensemble of 10^5 artificial series obtained each by linking together two series of fractional noise with sizes $N_1 = N_2 = 1024$, mean values $\mu_1 = 1$, $\mu_2 = 0$ ($\Delta\mu = 1$) and: (i) same correlation coeficient $\beta_1 = \beta_2 = 0.6$ and different standard deviation $\sigma_1 = 0.75$, $\sigma_2 = 1.25$ (ii) different correlation coeficient $\beta_1 = 0.5$, $\beta_2 = 0.7$ and the same standard deviation $\sigma_1 = \sigma_2 = 1$. We also include the results for $\beta_1 = \beta_2 = 0.6$ and $\sigma_1 = \sigma_2 = 1$ for comparison. (a) Distribution of the positions of the change-points detected by the segmentation algorithm ($p_0 = 0.05$). (b) Distribution of the number of cuts (k) detected in each series by the segmentation algorithm ($p_0 = 0.05$).

5.2 Segmentation of DNA sequences

DNA molecules are basically composed of four different nucleotides: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) and thus, from the point of view of sequence analysis they can be considered as symbolic strings made of four different symbols {A,C,G,T}. Nevertheless, DNA sequences can be studied using methods developed for numerical series provided that the symbols {A,C,G,T} are converted into numerical quantities [72].

In DNA-sequence analysis one of the most frequently used of such conversions the SW mapping rule: given a DNA sequence $B_1, B_2, \ldots, B_N$ with $B_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ we obtain its corresponding numerical series $\{x_i\}$, according to [72]:

$$x_i = \begin{cases} 1 \text{ if } B_i = \mathrm{C} \text{ or } \mathrm{G} \\ 0 \text{ if } B_i = \mathrm{A} \text{ or } \mathrm{T}. \end{cases} \quad (13)$$

This mapping rule is designed to study the G + C content along the sequence and is specially appropriate to analyze genome-wide organization because it corresponds to the most fundamental partitioning of the four bases into their natural pairs in the double helix (G + C, A + T). The proportion of G + C bases (usually called as G + C content or G + C composition), 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 [73,74] or mechanical waves [75] along the sequence, as well as to many biological features.

Another reason to focus our interest on the G + C composition of DNA is because, since the availability of the complete human genome [76], an intense controversy about the large scale organization of G + C composition of human DNA [45,76–78] has arisen. This problem is closely related to the results presented here since most of the above referred controversy is due to the fact that different authors define "homogeneity" in different ways: random uncorrelated sequence [76], homogeneous above a given scale [43,44], etc. Nevertheless, it is known that DNA sequences are far from behaving like random sequences and that they present long-range correlations of

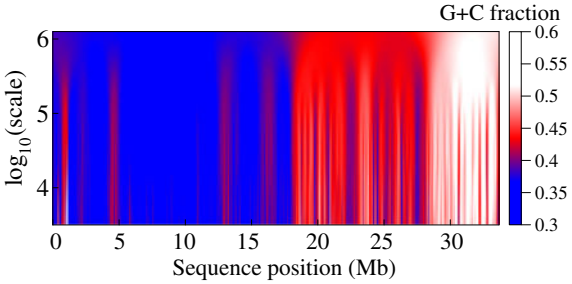

Fig. 11. (Color online) G + C composition of the DNA sequence of the q-arm of human chromosome 21 (33.6 millions bp) at different scales by means of a multi-scale wavelet plot. We have used a Gaussian wavelet with a characteristic scale varying in the range $(3 \times 10^3, 1.25 \times 10^6)$ (bp).

Table 1. Segmentation using the random i.i.d. as reference for homogeneity of the q-arm of the human chromosome 21 at different significance levels.

Significance level (p_0)	Number of segments	Average segment length (bp)
0.10	28836	1168
0.05	20378	1653
0.01	11123	3028

complex nature in their G + C composition, showing heterogeneities at all scales and that these correlations can be modeled as fractional Gaussian noise with $\beta \simeq 0.6$ [79] and thus, our segmentation method seems to be suitable to analyze DNA.

As an example of the application of our segmentation algorithm we use a human sequence, the q-arm of the chromosome 21 with a length of 33.7 millions base pairs (bp). The G + C composition of this sequence is shown in Figure 11 by means of a multi-scale wavelet plot, a useful tool to represent the heterogeneities of a sequence at different scales that was first applied to DNA sequences by Arneodo et al. [80]. The wavelet analysis with different scales of the wavelet function shows a hierarchical cascade from large to small scales (top to bottom in Fig. 11) representing heterogeneities in the concentration of G + C nucleotides along the DNA sequences. As can be seen, the human DNA sequence shows complex heterogeneities at all scales. Indeed, if we segment this sequence using as a reference for homogeneity the random i.i.d. series, i.e. compute the p-value using $\mathcal{P}_{0,N}(\tau)$, we obtain thousands of segments (Tab. 1).

Nevertheless, despite this complex heterogeneity at short scales, it seems clear that, at large scales, three main regions of different G + C composition can be distinguished by eye: one of low G + C in the first 18 million bases (Mb), another of intermediate G + C composition from 18 Mb to roughly 28 Mb and finally from here to the end of the sequence a region of high G + C composition. Taking into account that DNA sequences are far from behaving like random sequences and that they present long-range correlations of complex nature in their G + C composition [79], it seems more appropriate to

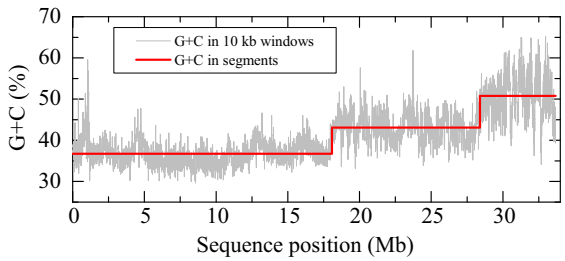

Fig. 12. (Color online) Segmentation of the sequence of the q-arm of human chromosome 21. We compute the significance level using $\mathcal{P}_{\beta,N}(\tau)$ with $\beta = 0.556$ ($\alpha = 0.778$) which is the scaling exponent obtained for this sequence by means of DFA (see Appendix A). The same segments are obtained for significance levels p_0 ranging from 0.1 to 0.01. The G + C composition along the sequence is also shown by averaging the composition in 10 kb non-overlapping windows. This coarse graining of the data has been done only for representation purposes and does not affect to the segmentation procedure.

use a fractional Gaussian noise as the reference for homogeneity. In particular, correlations observed in the q-arm of human chromosome 21 can be modeled by a fGn with exponent $\beta = 0.556$ – this corresponds to a DFA scaling exponent $\alpha = 0.778$, see Appendix A.

In Figure 12 we show the segmentation of the q-arm of human chromosome 21 computing the p-value using $\mathcal{P}_{\beta,N}(\tau)$ with $\beta = 0.556$. We obtain only three segments which clearly correspond to the three regions of different G + C composition described above. It is also worth mentioning that the same segments are obtained for significance levels p_0 ranging from 0.1 to 0.01, thus pointing out the robustness of the method. Similar results have been obtained when segmenting all human chromosome sequences, showing the existence of previously unknown huge compositional structures in human DNA [66].

6 Conclusions

In this paper we present an approach for the segmentation of time series, specially designed to the analysis of series with long-range fractal correlations. The majority of the traditional segmentation techniques consider a time series as homogeneous when its heterogeneities are similar to those found in a random i.i.d. series. Such algorithms, when faced with a time series characterized by long-range fractal correlations (even in the case when the fractal series is stationary) produce a clear oversegmentation, since they detect heterogeneities produced by the correlations, and not by real non-stationarities (such as for example changes in the mean). The new segmentation approach we develope sets as a reference for homogeneity fractional noise, with the same degree of correlations as the original series to be segmented, and therefore it is aimed at detecting real heterogeneities and not those produced merely by the presence of long-range correlations. We provide a procedure to determine the statistical confidence when deciding if a time series has to be split at a given change point, and we incorporate this procedure in

two different segmentation strategies – a heuristic and an optimal strategy. By generating artificial fractal time series with superposed real non-stationarities, we check the good performance of the method to detect change points due to the real non-stationarities introduced in the time series, and to avoid change points that are merely due to the fluctuations caused by presence of fractal correlations. The real non-stationarities we introduce by creating differences in the mean between adjacent segments are detected by the method with high probability even when the difference of mean is smaller than the internal fluctuations of the adjacent segments (Fig. 9). The good results of the method in these control experiments support the validity of the results we obtain in the segmentation of human DNA sequences, where we have revealed the existence of previously unknown compositional structures at large scales [66].

Funding: Spanish Government (Grant BIO2008-01353) and Spanish Junta de Andalucía (Grants P06-FQM1858, P07-FQM3163 and FQM362). Spanish 'Juan de la Cierva' grant to M.H. P.Ch.I. thanks NIH Grant 1RO1-HL098437, ONR Grant 000141010078 and the Brigham and Women's Hospital Biomedical Research Institute Fund for support. P.B. thanks F.M. Guerrero-Cervantes for fruitful discussions at the early stage of this work.

Appendix A: Measure of the correlations. DFA method

Here we describe the detrended fluctuation analysis (DFA) developed by Peng et al. [55]. This method was specially designed to work with nonstationary series and also provides a single quantitative parameter – the scaling exponent α – to represent the correlation properties of a long-range correlated series.

DFA involves the following steps [55]:

(i) Starting with a correlated series $\{x_i\}$ of size N we first integrate the series and obtain

$$y(j) \equiv \sum_{i=1}^{j} [x_i - \mu] \tag{A.1}$$

where μ is the mean value of the entire series.

(ii) The integrated series $y(j)$ is divided into boxes of equal length ℓ.

(iii) In each box of length ℓ, we calculate a linear fit of $y(j)$ which represents the *linear trend* in that box. The y coordinate of the fit line in each box is denoted by $y_\ell(j)$.

(iv) The integrated series $y(j)$ is detrended by subtracting the local trend $y_\ell(j)$ in each box of length ℓ.

(v) For a given box size ℓ, the root mean-square (r.m.s.) fluctuation for this integrated and detrended series is calculated:

$$F(\ell) = \sqrt{\frac{1}{N} \sum_{j=1}^{N} [y(j) - y_\ell(j)]^2}. \tag{A.2}$$

(vi) The above computation is repeated for a broad range of scales (box sizes ℓ) to provide a relationship between $F(\ell)$ and the box size ℓ.

Note that steps (iii)–(iv) eliminate the linear trends present in the integrated signal $y(j)$ or, equivalently, the local trends in the original signal $\{x_i\}$. A fit to a polynomial of higher order ℓ in step (iii) would eliminate higher order polynomial trends. This generic procedure is known as DFA-ℓ. However, we consider here only DFA-1 (or simply DFA) because we are interested on signals made of pieces with different mean values.

A power-law relation between the average root-mean-square fluctuation function $F(\ell)$ and the box size ℓ indicates the presence of scaling: $F(\ell) \sim \ell^\alpha$. The fluctuations can be characterized by a scaling exponent α, a self-similarity parameter which quantifies the long-range power-law correlation properties of the signal. Indeed, the exponent α is related to the exponent β by [55]:

$$\beta = 2\lambda = 2\alpha - 1. \tag{A.3}$$

Thus $\alpha = 0.5$ ($\beta = 0$) corresponds to an uncorrelated signal (white noise). Values of $\alpha > 0.5$ correspond to correlated signals ($\beta > 0$). In particular $\alpha = 1.5$ ($\beta = 2$) is the Brownian Motion and $\alpha = 1$ ($\beta = 1$) is the $1/f$ noise. Values of $\alpha < 0.5$ lead to anti-correlations ($\beta < 0$)

In order to provide a more accurate estimate of $F(\ell)$, the largest box size ℓ we use in our calculations is $N/10$, where N is the total number of points in the signal.

Appendix B: Similarity between t and ΔV for k = 1

If we consider a series $\mathcal{S}$ divided into two subseries $\mathcal{S}_1$ and $\mathcal{S}_2$, the objective function defined in (8) for only one change point ($k = 1$) is given by:

$$\begin{aligned}
\Delta V &= V(\mathcal{S}) - [V(\mathcal{S}_1) + V(\mathcal{S}_2)] \\
&= N_1\mu_1^2 + N_2\mu_2^2 - N\mu^2 \\
&= N_1\mu_1^2 + N_2\mu_2^2 - [N_1\mu_1 + N_2\mu_2]^2 \\
&= N_1\left(1 - \frac{N_1}{N}\right)\mu_1^2 + N_2\left(1 - \frac{N_2}{N}\right)\mu_2^2 - 2\frac{N_1N_2}{N}\mu_1\mu_2 \\
&= \frac{N_1N_2}{N}\left[\mu_1^2 + \mu_2^2 - 2\mu_1\mu_2\right] \\
&= \frac{N_1N_2}{N}\left[\mu_1 - \mu_2\right]^2,
\end{aligned} \tag{B.4}$$

and thus, the maximization of ΔV is equivalent to maximize the difference between μ_1 and μ_2. In fact, for $N \gg 1$ and $V(\mathcal{S}_1) \simeq V(\mathcal{S}_2)$ it is easy to check that:

$$\Delta V \sim t^2. \tag{B.5}$$

References

1. I. Berkes, L. Horvath, P. Kokoszka, Q.M. Shao, Ann. Stat. **34**, 1140 (2006)

2. B.J. West, M.F. Shlesinger, Int. J. Mod. Phys. B **3**, 795 (1989)
3. *Theory and Applications of Long-Range Dependence*, edited by P. Doukhan, G. Oppenheim, M.S. Taqqu (Birkhäuser, Boston, 2002)
4. P.Ch. Ivanov, L.A.N. Amaral, A.L. Goldberger, H.E. Stanley, Europhys. Lett. **43**, 363 (1998)
5. *Change-point Problems. Lecture notes and Monograph series*, edited by E. Carlstein, H.G. Muller, D. Siegmund (Institute of Mathematical Statistics, Hayward, CA, 1994), Vol. 23
6. H. Kantz, T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 1997)
7. T. Schreiber, Phys. Rev. Lett. **78**, 843 (1997)
8. A. Witt, J. Kurths, A. Pikovsky, Phys. Rev. E **58**, 1800 (1998)
9. G. Mayer-Kress, Integr. Physiol. Behav. Sci. **29**, 205 (1994)
10. R. Hegger, H. Kantz, L. Matassini, Phys. Rev. Lett. **84**, 3197 (2000)
11. M.M. Wolf et al., Med. J. Aust. **2**, 52 (1978)
12. C. Guilleminault et al., Lancet **1**, 126 (1984)
13. P.Ch. Ivanov et al., Nature **383**, 323 (1996)
14. P. Bernaola-Galván, P.Ch. Ivanov, L.A.N. Amaral, H.E. Stanley, Phys. Rev. Lett. **87**, 168105 (2001)
15. P.Ch. Ivanov et al., Europhys. Lett. **48**, 594 (1999)
16. J.W. Kantelhardt et al., Phys. Rev. E **65**, 051908 (2002)
17. R. Karasik et al., Phys. Rev. E **66**, 062902 (2002)
18. P.Ch. Ivanov, Z. Chen, K. Hu, H.E. Stanley, Physica A **344**, 685 (2004)
19. P.Ch. Ivanov et al., Proc. Natl. Acad. Sci. USA **104**, 20702 (2007)
20. D.T. Schmitt, P.K. Stein, P.Ch. Ivanov, IEEE Trans. Biomed. Eng. **56**, 1564 (2009)
21. P.Ch. Ivanov, IEEE Eng. Med. Biol. Mag. **26**, 33 (2007)
22. M. Gardiner-Garden, M. Frommer, J. Mol. Biol. **196**, 261 (1987)
23. P.L. Luque-Escamilla et al., Phys. Rev. E **71**, 061925 (2005)
24. M. Hackenberg et al., BMC Bioinformatics **7**, 446 (2006)
25. M. Ortuño et al., Europhys. Lett. **57**, 759 (2002)
26. P. Carpena et al., Phys. Rev. E **79**, 035102 (2009)
27. J.C. Wong, H. Lian, S.A. Cheong, Phys. A **388**, 4635 (2009)
28. K. Fukuda et al., Europhys. Lett. **62**, 189 (2003)
29. L. Horváth, J. Multivar. Anal. **78**, 218 (2001)
30. S. Ben Hariz, J.J. Wylie, C. R. Math. **341**, 765 (2005)
31. L.H. Wang, J. Stat. Comput. Simul. **78**, 653 (2007)
32. L. Horváth, P. Kokoszka, J. Stat. Plann. Inference **64**, 57 (1997)
33. C. Inclán, C. Tiao, J. Am. Stat. Assoc. **89**, 913 (1994)
34. B. Whitcher, P. Guttorp, D.B. Percival, J. Stat. Comput. Simul. **68**, 65 (2000)
35. B. Whitcher, S.D. Byers, P. Guttorp, D.B. Percival, Water Resour. Res. **38**, 1054 (2002)
36. E. Andreou, E. Ghysels, J. Appl. Econ. **17**, 579 (2002)
37. J. Beran, N. Terrin, Biometrika **83**, 627 (1996)
38. L.H. Wang, J.D. Wang, J. Stat. Comput. Simul. **76**, 317 (2006)
39. P. Carpena, P. Bernaola-Galván, Phys. Rev. B **60**, 201 (1999)
40. I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J.L. Oliver, H.E. Stanley, Phys. Rev. E **65**, 041905 (2002)
41. G.L. Feng, Z.Q. Gong, W.J. Dong, J.P. Li, Acta Physica Sinica **54**, 5494 (2005)
42. G.L. Feng, Z.Q. Gong, R. Zhi, D.Q. Zhang, Chin. Phys. B **17**, 2745 (2008)
43. J.L. Oliver et al., Gene **276**, 47 (2001)
44. J.L. Oliver et al., Gene **300**, 117 (2002)
45. W. Li, P. Bernaola-Galván, P. Carpena, J.L. Oliver. Comput. Biol. Chem. **27**, 5 (2003)
46. J.L. Oliver et al., Nucleic Acids Res. **32**, W287 (2004)
47. V. Thakur, R.K. Azad, R. Ramaswamy, Phys. Rev. E **75**, 011915 (2007)
48. B. Toth, F. Lillo, J.D. Farmer, Eur. Phys. J. B **78**, 235 (2010)
49. J. Beran, *Statistics for long memory processes* (Chapman & Wall, 1994)
50. S.B. Lowen, M.C. Teich, *Fractal-Based Point Processes* (Wiley Interscience, 2005), Chap. 6
51. K. Fukuda, H.E. Stanley, L.A.N. Amaral, Phys. Rev. E **69**, 021108 (2004)
52. W. Wyss, Found. Phys. Lett. **4**, 235 (1991)
53. J.R.M. Hosking, Biometrika **68**, 165 (1981)
54. H.A. Makse, S. Havlin, M. Schwartz, H.E. Stanley, Phys. Rev. E **53**, 5445 (1996)
55. C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Phys. Rev. E **49**, 1685 (1994)
56. K. Hu et al., Phys. Rev. E **64**, 011114 (2001)
57. Z. Chen et al., Phys. Rev. E **65**, 041107 (2002)
58. Q.D.Y. Ma et al., Phys. Rev. E **81**, 031101 (2010)
59. Z. Chen et al., Phys. Rev. E **71**, 011104 (2005)
60. Y. Xu et al., Physica A **390**, 4057 (2011)
61. L.M. Xu et al., Phys. Rev. E **71**, 051101 (2005)
62. P. Bernaola-Galván, R. Román-Roldán, J.L. Oliver, Phys. Rev. E **53**, 5181 (1996)
63. W.H. Press et al., *Numerical Recipes in FORTRAN* (Cambridge University Press, Cambridge, 1994)
64. W. Li, Phys. Rev. Lett. **86**, 5815 (2001)
65. W. Li, Gene **276**, 57 (2001)
66. P. Carpena, J.L. Oliver, M. Hackenberg, A.V. Coronado, G. Barturen, P. Bernaola-Galván. Phys. Rev. E **83**, 031908 (2011)
67. N. Haiminen, H. Manila, E. Terzi, BMC Bioinformatics **8**, 171 (2007)
68. R. Bellman, Coummun ACM **4**, 284 (1961)
69. W. Li, Complexity **3**, 33 (1998)
70. R. Román-Roldán, P. Bernaola-Galván, J.L. Oliver, Phys. Rev. Lett. **80**, 1344 (1998)
71. P. Bernaola-Galván, R. Román-Roldán, J.L. Oliver, Phys. Rev. Lett. **83**, 3336 (1999)
72. P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J.L. Oliver, Gene **300**, 105 (2002)
73. P.J. Dandliker, R.E. Holmlin, J.K. Barton, Science **275**, 1465 (1997)
74. P. Carpena, P. Bernaola-Galván, P.Ch. Ivanov, H.E. Stanley, Nature **418**, 955 (2002)
75. M. Rief, H. Clausen-Schaumann, H.E. Gaub, Nat. Struct. Biol. **6**, 346 (1999)
76. J.C. Venter et al., Science **291**, 1304 (2001)
77. N. Cohen, T. Dagan, L. Stone, D. Graur, Mol. Biol. Evol. **22**, 1260 (2005)
78. O. Clay, G. Bernardi, Mol. Biol. Evol. **22**, 2315 (2005)
79. P. Carpena, P. Bernaola-Galván, A.V. Coronado, M. Hackenberg, J.L. Oliver Phys. Rev. E **75**, 032903 (2007)
80. A. Arneodo, E. Bacry, P.V. Graves, J.F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995)