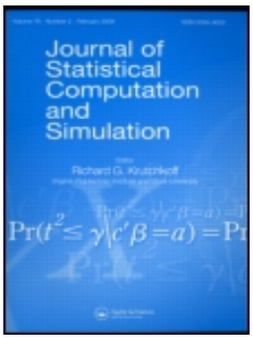

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Statistical properties of detrended fluctuation analysis

N. Crato , R. R. Linhares & S. R.C. Lopes

To cite this article: N. Crato , R. R. Linhares & S. R.C. Lopes (2010) Statistical properties of detrended fluctuation analysis, Journal of Statistical Computation and Simulation, 80:6, 625-641, DOI: 10.1080/00949650902755152

To link to this article: https://doi.org/10.1080/00949650902755152

Journal of Statistical Computation and Simulation
Vol. 80, No. 6, June 2010, 625–641

Taylor & Francis
Taylor & Francis Group

Statistical properties of detrended fluctuation analysis

N. Crato[a], R.R. Linhares[b] and S.R.C. Lopes[b]*

[a]*Cemapre, ISEG, Technical University of Lisbon, Lisbon, Portugal;* [b]*Mathematical Institute, UFRGS, Porto Alegre, RS, Brazil*

(*Received 25 April 2008; final version received 16 January 2009*)

The main goal of this work is to consider the *detrended fluctuation analysis* (DFA), proposed by Peng *et al.* [*Mosaic organization of DNA nucleotides*, Phys. Rev. E. 49(5) (1994), 1685–1689]. This is a well-known method for analysing the long-range dependence in non-stationary time series. Here we describe the DFA method and we prove its consistency and its exact distribution, based on the usual $i.i.d.$ assumption, as an estimator for the fractional parameter d. In the literature it is well established that the nucleotide sequences present long-range dependence property. In this work, we analyse the long dependence property in view of the *autoregressive moving average fractionally integrated* ARFIMA(p, d, q) processes through the analysis of four nucleotide sequences. For estimating the fractional parameter d we consider the semiparametric regression method based on the periodogram function, in both classical and robust versions; the semiparametric R/S(n) method, proposed by Hurst [*Long term storage in reservoirs*, Trans. Am. Soc. Civil Eng. 116 (1986), 770–779] and the maximum likelihood method (see [R. Fox and M.S. Taqqu, *Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series*, Ann. Statist. 14 (1986), 517–532]), by considering the approximation suggested by Whittle [*Hypothesis Testing in Time Series Analysis* (1953), Hafner, New York].

Keywords: long memory; detrended fluctuation analysis; semiparametric estimation; robustness

AMS Classification: Primary: 91B70; Secondary: 00A05

1. Introduction

Persistence or long-range dependence has been observed in time series in different areas of the science such as meteorology, astronomy, hydrology, and economics, as reported in Beran [1]. One of the models that exhibit the long-range dependence is the *autoregressive fractionally integrated moving average*, denoted by ARFIMA(p, d, q) process, where d is the fractional parameter and p and q are, respectively, the degrees of the autoregressive and moving average polynomials. There are several estimation procedures for the ARFIMA parameters, mainly in the semiparametric and parametric classes.

The nucleotide sequences can be represented by a time series (see [2]). To obtain a time series from a nucleotide sequence it is necessary to consider some type of transformation (see [3]).

*Corresponding author. Email: silvia.lopes@ufrgs.br

ISSN 0094-9655 print/ISSN 1563-5163 online

DOI: 10.1080/00949650902755152
http://www.informaworld.com

The statistical properties of DNA genomes are of interest because they reflect biological features (see [4]). For instance, the period-three (P-3) property manifests itself as a repeating unit of three nucleotides appearing in coding regions but absent elsewhere (see [5]). Consequently, this property can be used to help identify coding regions.

Several researchers (see [2,3,6–10] among others) have studied the existence of long-range or power-law correlations in DNA sequences. Peng *et al.* [2], Li and Kaneko [11] and Voss [12] point the existence of long-range to fractal (scale-invariant) structure in DNA sequences. It is known that DNA nucleotides form a mosaic of long homogeneous segments or 'isochores' (see [13–15]). For some authors the existence of long-range power-law correlations seems to be related to such 'isochore' segments (see [16]). Carpena *et al.* [17] argue that the DNA correlations are much more complex than power laws with a single scaling exponent. In fact, these authors propose to analyse different scales for the exponents of such power laws. They show that the sequence corresponding to human chromosome IV, by considering the SW mapping rule, exhibits non-fractal behaviour suggesting the presence of two major peaks in the power-law exponent. So, their conclusion is that no single scaling exists in the human genome. Oliver *et al.* [15] explore the phylogenetic distribution of large-scale genome patchiness by considering the deviations of the power-law behaviour in long-range correlations.

In the literature, the '*detrended fluctuation analysis*' (DFA), proposed by Peng *et al.* [2], has successfully been applied to different fields of interest, such as DNA sequences (see [3,18]), economical time series (see [19]) heart rate variability analysis (see [20]) and long-time weather records (see [21]). The DFA is a well-established method for detecting long-range dependence in non-stationary time series. This method is based on random walk theory; it is similar to the R/S(n) method ('*rescaled range analysis*') (see [22]) and also similar to another method based on wavelet transform (see [21]). The object of this technique is to evaluate the statistical fluctuation $F(l)$ in order to obtain a set of measures, where l represents the window length. By varying the length l, the fluctuation can be characterized by the scaling exponent, that is the slope of the line obtained by regressing $\ln(F(l))$ on $\ln(l)$.

The main goal of this paper is to analyse the statistical properties of the DFA method. We are interested in analysing the *long-range dependence* parameter in four nucleotide sequences. This will be done by considering several estimation methods for the *fractional parameter* d, in the semiparametric and parametric classes.

The paper is organized as follows. In Section 2, we present the autoregressive fractionally integrated moving average process (ARFIMA). In Section 3 we review some estimation methods for the *fractional parameter* d, in the semiparametric and parametric classes. Section 4 describes the DFA method and presents its statistics properties where we prove its consistency and its exact distribution as an estimator of the *fractional parameter* d. In Section 5 we present the analysis of four nucleotide sequences. Section 6 gives the conclusions.

2. ARFIMA(p, d, q) process

In this section we define the ARFIMA process, which exhibits the long memory property.

DEFINITION 2.1 *Let* $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ *be a white noise process with zero mean and variance* $\sigma_\epsilon^2 > 0$, $\mathcal{B}$ *be the backward-shift operator, that is,* $\mathcal{B}^k(X_t) = X_{t-k}$ *and* $\Phi(\cdot)$ *and* $\Theta(\cdot)$ *polynomials of orders* p *and* q, *respectively, given by*

$$\Phi(\mathcal{B}) = 1 - \phi_1\mathcal{B} - \cdots - \phi_p\mathcal{B}^p$$

and

$$\Theta(\mathcal{B}) = 1 - \theta_1\mathcal{B} - \cdots - \theta_q\mathcal{B}^q,$$

where ϕ_i, $1 \leq i \leq p$, and θ_j, $1 \leq j \leq q$, are real constants. If $\{X_t\}_{t\in\mathbb{Z}}$ is a linear process given by

$$\Phi(\mathcal{B})(1-\mathcal{B})^d(X_t-\mu) = \Theta(\mathcal{B})\varepsilon_t, \quad t \in \mathbb{Z}, \tag{1}$$

where μ is the mean of the process, then $\{X_t\}_{t\in\mathbb{Z}}$ is called a general fractionally differenced ARFIMA(p, d, q) *process, where $d \in (-0.5, 0.5)$ is the degree or parameter of differencing.*

Remark 2.1 (a) The process

$$U_t = (1-\mathcal{B})^d X_t, \quad t \in \mathbb{Z},$$

given by

$$\Phi(\mathcal{B})U_t = \Theta(\mathcal{B})\varepsilon_t, \quad t \in \mathbb{Z},$$

is an *autoregressive moving average process* ARMA(p, q).

(b) If $d \in (-0.5, 0.5)$ then the process $\{X_t\}_{t\in\mathbb{Z}}$ is stationary and invertible and its spectral density function is given by

$$f_X(w) = f_U(w)\left[2\sin\left(\frac{w}{2}\right)\right]^{-2d} \quad \text{for } 0 < w \leq \pi, \tag{2}$$

where $f_U(\cdot)$ is the spectral density function of the ARMA(p, q) process. One observes that $f_X(w) \simeq w^{-2d}$, when $w \to 0$.

(c) The term $(1-\mathcal{B})^d$, in the expression (1), is the binomial expansion

$$(1-\mathcal{B})^d = \sum_{k=0}^{\infty}\binom{d}{k}(-\mathcal{B})^k = 1 - d\mathcal{B} - \frac{d}{2!}(1-d)\mathcal{B}^2\cdots \quad \text{for } d \in \mathbb{R}. \tag{3}$$

Persistence or *long memory* property has been observed in time series from different fields such as meteorology, astronomy, hydrology and economy. One can characterize the persistence by two equivalent forms:

- In the time domain, the autocorrelation function $\rho_X(\cdot)$ decays hyperbolically to zero, that is, $\rho_X(k) \simeq k^{2d-1}$, when $k \to \infty$.
- In the frequency domain, the spectral density function $f_X(\cdot)$ is unbounded when the frequency is near zero, that is, $f_X(w) \simeq w^{-2d}$, when $w \to 0$.

Remark 2.2 The ARFIMA(p, d, q) process exhibits the property of long memory when $d \in (0.0, 0.5)$, of intermediate memory when $d \in (-0.5, 0.0)$ and of short memory when $d = 0$.

Important properties for ARFIMA(p, d, q) processes can be found in Hosking [23], Beran [1] and Doukhan *et al.* [24].

3. Estimation methods

To estimate the fractional parameter d we consider semiparametric and parametric estimation classes. We consider the following estimation methods: the semi-parametric regression method based on the periodogram function, both classical and robust versions; the semiparametric R/S(n) method, proposed by Hurst [22] and the maximum likelihood method (see [25]), by considering the approximation suggested by Whittle [26].

3.1. *Semiparametric class*

In the semiparametric class, the parameters are estimated in two steps: only d is estimated in the first step and the others are estimated in the second step.

For the estimation of the fractional differencing parameter d, we now summarize some methods in this class:

- The semiparametric regression method based on the periodogram function, proposed by Geweke and Porter-Hudak [27], both classical and robust versions.
- The semiparametric regression method based on GPH with trimming l and bandwidth $g(n)$, proposed by Robinson [28], both classical and robust versions.
- The semiparametric method based on Hurst [22] estimator. This estimator is largely known as the R/S statistics.

Let $\{X_t\}_{t\in\mathbb{Z}}$ be an ARFIMA(p, d, q), given by Equation (1). Taking the logarithm of the spectral density function $f_X(\cdot)$ given by Equation (2), we have

$$\ln(f_X(w)) = \ln(f_U(w)) - d\ln\left(4\sin^2\left(\frac{w}{2}\right)\right),$$

or writing

$$\ln(f_X(w)) = \ln(f_U(0)) - d\ln\left(4\sin^2\left(\frac{w}{2}\right)\right) + \ln\left(\frac{f_U(w)}{f_U(0)}\right). \tag{4}$$

Substituting w by $w_j = (2\pi j)/n$ and adding $\ln(I_n(w_j))$ to both sides of Equation (4), we obtain

$$\ln(I_n(w_j)) = \ln(f_U(0)) - d\ln\left(4\sin^2\left(\frac{w_j}{2}\right)\right) + \ln\left(\frac{f_U(w_j)}{f_U(0)}\right) + \ln\left(\frac{I_n(w_j)}{f_X(w_j)}\right), \tag{5}$$

where $I_n(\cdot)$ is the periodogram function given by

$$I_n(w) = \frac{1}{2\pi}\left(\widehat{\gamma}_X(0) + 2\sum_{k=1}^{n-1}\widehat{\gamma}_X(k)\cos(wk)\right), \quad w \in (0, \pi], \tag{6}$$

where $\widehat{\gamma}_X(\cdot)$ is the sample autocovariance function of the process $\{X_t\}_{t\in\mathbb{Z}}$.

When considering only the frequencies close to zero, the term $\ln(f_U(w_j)/f_U(0))$ may be discarded (see [27]). Then, we may rewrite Equation (5) in the context of a simple linear regression model

$$y_j = a + bx_j + \epsilon_j, \quad j = 1, \ldots, g(n), \tag{7}$$

where $g(n) = n^\beta$, for $0 < \beta < 1, b = -d, a = \ln(f_U(0)), y_j = \ln(I_n(w_j)), x_j = \ln(4\sin^2(w_j/2))$ and $\epsilon_j = \ln(I_n(w_j)/f_X(w_j))$, for $j \in \{1, \ldots, g(n)\}$.

A semiparametric regression estimator (see [29,30]) may be obtained by minimizing some loss function of the residuals

$$r_j = y_j - a - bx_j \quad \text{for } j = 1, \ldots, g(n). \tag{8}$$

We consider three different loss functions. They give rise to the classical ordinary least squared method (OLS), the least trimmed squared (LTS), proposed by Rousseeuw [31] and the MM method, proposed by Yohai [32].

DEFINITION 3.1 *The OLS Estimators are the values* $(\hat{a}, \hat{b})$ *which minimize the loss function*

$$L_1(g(n)) = \sum_{j=1}^{g(n)} (r_j)^2, \tag{9}$$

where r_j *is given by expression (*8*), for* $j \in \{1, \ldots, g(n)\}$.

DEFINITION 3.2 *The robust estimators LTS* (*see* [31]) *are the values* $(\hat{a}, \hat{b})$ *that minimize the loss function*

$$L_2(g(n)) = \sum_{j=1}^{g^*(n)} (r^2)_{j:g(n)}, \tag{10}$$

where $(r^2)_{j:g(n)}$ *are the squared and ordered residuals, that is,* $(r^2)_{1:g(n)} \leq \cdots \leq (r^2)_{g^*(n):g(n)}$, *and* $g^*(n)$ *is the number of points used in the optimization procedure*.

DEFINITION 3.3 *The robust estimators MM* (*see* [32]) *are the values* $(\hat{a}, \hat{b})$ *that minimize the loss function*

$$L_3(g(n)) = \sum_{j=1}^{g(n)} \rho_2 \left(\frac{r_j}{s}\right)^2, \tag{11}$$

subject to the constraint

$$\frac{1}{g(n)} \sum_{j=1}^{g(n)} \rho_1 \left(\frac{r_j}{s}\right) \leq C, \tag{12}$$

where $\rho_1(\cdot)$ *and* $\rho_2(\cdot)$ *are symmetric, bounded, nondecreasing functions on* $[0, \infty)$ *with* $\rho_j(0) = 0$ *and* $\lim_{u\to\infty} \rho_j(u) = 1$, $j = 1, 2$, s *is a scale parameter and* C *is a tuning constant*.

3.1.1. *GPH, GPH-LTS and GPH-MM estimators*

The first estimation method based on the periodogram function was proposed by Geweke and Porter-Hudak [27]. To obtain an estimate for d, these authors propose to apply the OLS method (see [29]) in Equation (8) based on Equation (6), which we denote it by GPH. The estimator of d is given by

$$\text{GPH} = -\frac{\sum_{j=1}^{g(n)} (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{g(n)} (x_j - \bar{x})^2}, \tag{13}$$

where

$$y_j = \ln(I_n(w_j)), \quad x_j = \ln\left(2 \sin\left(\frac{w_j}{2}\right)\right)^2 \quad \text{and} \quad \bar{x} = \frac{1}{g(n)} \sum_{j=1}^{g(n)} x_j.$$

The variance of the GPH estimator (see [27]) is given by

$$\text{Var(GPH)} = \frac{\pi^2}{6 \sum_{j=1}^{g(n)} (x_j - \bar{x})^2}.$$

To obtain the robust version of the GPH estimator we just apply the least trimmed squared (LTS) and MM methodologies (see [29]), respectively, to the regression model (8). This gives rise to the GPH-LTS and the GPH-MM estimators.

3.1.2. *R, R-LTS and R-MM estimators*

The regression estimator, proposed by Robinson [28], is obtained by applying the OLS method in Equation (8) based on Equation (6), but considering only the frequencies ω_j, for $j \in \{l, l+1, \ldots, g(n)\}$, where $l > 1$ is a trimming value that tends to infinity more slowly than $g(n)$.

The asymptotic variance of the estimator R (see [28]) is given by

$$\mathrm{Var(R)} \sim \frac{\pi^2}{24g(n)}.$$

To obtain the robust version of the R estimator, denoted, respectively, by R-LTS and R-MM, we just apply the least trimmed squared (LTS) and MM methodologies (see [27]) to the regression model (8).

3.1.3. *$R/S(n)$ and $R/S(q)$ estimators*

Here we introduced the *$R/S(n)$ statistic* proposed by Hurst [22] and a modified version of it, denoted by $R/S(q)$ and proposed by Lo [33].

DEFINITION 3.4 *Let $\{X_t\}_{t=1}^n$ be a time series. The rescaled range statistic $R/S(n)$, introduced by Hurst* [22], *is defined by*

$$\mathrm{R/S}(n) = \frac{1}{s_n}\left[\max_{1\le k\le n}\sum_{j=1}^{k}(X_j - \bar{X}) - \min_{1\le k\le n}\sum_{j=1}^{k}(X_j - \bar{X})\right],$$

where $\bar{X} = 1/n\sum_{j=1}^n X_j$ and $s_n^2 = 1/n\sum_{j=1}^n (X_j - \bar{X})^2$ is the sample variance.

For the fractional Gaussian noise process or the ARFIMA process (see [34]),

$$\mathbb{E}[\mathrm{R/S}(n)] \sim C_{\mathrm{H}} n^{\mathrm{H}} \quad \text{as } n \to \infty,$$

where H is the parameter suggested by Harold Edwin Hurst (1880–1978), to estimate long-range dependence, and C_{H} is a positive constant independent of n.

To determine H from the *$R/S(n)$statistic*, one proceeds as follows:

- For each $j \in \{1, \ldots, s\}$, one divides the time series $\{X_t\}_{t=1}^n$ into $[n/k_j]$ blocks, each one of size k_j, where $k_j = \ell k_{j-1}$.
- For each block, one computes the $\mathrm{R/S}(k_j)$ statistic.
- One adjusts a regression line, by regressing $\ln(\mathrm{R/S}(k_j))$ on $\ln(k_j)$, $j = 1, \ldots, s$, to obtain H the *Hurst parameter*, that is, the slope of the adjusted line.

Remark 3.1 The Hurst parameter H is related to the fractional parameter d by the equation (see [35])

$$d = H - \frac{1}{2}. \tag{14}$$

DEFINITION 3.5 *The HAC variance estimator with bandwidth q, is defined as*

$$\hat{\sigma}_n^2(q) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2 + \frac{2}{n}\sum_{j=1}^{q}\omega_j(q)\left(\sum_{l=j+1}^{n}(X_l - \bar{X})(X_{l-j} - \bar{X})\right), \tag{15}$$

where $\bar{X} = 1/n\sum_{j=1}^{n} X_j$ and the weights $\omega_j(q)$ are given by

$$\omega_j(q) = 1 - \frac{j}{q+1} \quad \text{for all } q < n.$$

Remark 3.2 There is no selection rule for choosing the order q. However, q should be related to the sample size n satisfying

$$\frac{1}{q} + \frac{q}{n} \to 0 \quad \text{as } n \to \infty.$$

A standard choice is $q = n^{0.5}$ (see [36]).

DEFINITION 3.6 *The $R/S(n)$ modified statistic, proposed by Lo* [33] *and denoted by $R/S(q)$, is defined as*

$$\mathrm{R/S}(q) = \frac{1}{\hat{\sigma}_n(q)}\left[\max_{1\le k\le n}\sum_{j=1}^{k}(X_j - \bar{X}) - \min_{1\le k\le n}\sum_{j=1}^{k}(X_j - \bar{X})\right],$$

where $\bar{X} = 1/n\sum_{j=1}^{n} X_j$ and $\hat{\sigma}_n(q)$ is defined in (15).

3.2. *Parametric class*

In the parametric class, all parameters (the autoregressive and moving average coefficients and the fractional differencing) can be simultaneously estimated.

In this subsection we present one of the most popular method in the parametric class. We summarize the maximum likelihood method (see [25]), by considering the approximation suggested by Whittle [26].

The estimator for d, by using the maximum likelihood method, denoted by W, is the value of

$$\boldsymbol{\eta} = (\sigma_X^2, d, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q) \tag{16}$$

that minimizes the function

$$\mathcal{Q}(\boldsymbol{\eta}) = \sum_{j=1}^{[(n-1)/2]}\left(\frac{I(w_j)}{f_X(w_j, \boldsymbol{\eta})}\right), \tag{17}$$

where $\boldsymbol{\eta}$ is the vector of unknown parameters given in Equation (16), $f_X(\cdot, \boldsymbol{\eta})$ is the spectral density function of the $\{X_t\}_{t\in\mathbb{Z}}$, $[x]$ is the integer part of x, $w_j = 2\pi j/n$ is the Fourier frequencies, for $j \in \{1, \ldots, [(n-1)/2]\}$, and $I(\cdot)$ is the periodogram function given by Equation (6).

The asymptotic variance of the estimator W (see [25]) is given by

$$\mathrm{Var}(W) \sim \frac{6}{\pi^2 n}.$$

More details on this estimator can be found in Fox and Taqqu [25] and Beran [1].

4. DFA method and some properties

Given a time series $\{X_t\}_{t=1}^n$, the DFA, proposed by Peng *et al.* [2], consists of five steps. In the first one, for each $t \in \{1, 2, \ldots, n\}$, we calculate

$$Y_t = \sum_{j=1}^{t} X_j. \tag{18}$$

Observe that the stochastic process $\{Y_t\}_{t\in\mathbb{Z}}$ is not stationary. In the second step we divide the time series $\{Y_t\}_{t=1}^n$ into $[n/l]$ non-overlapping blocks, each containing l observations. In the third step, for each block, one fits a least-square line to the data (that represents the local trend in the block). In the fourth step, we detrend the time series $\{Y_t\}_{t=1}^n$, that is, in each block we calculate

$$Z_t = Y_t - Y_t^l, \tag{19}$$

where Y_t^l denotes the adjusted fit on each block.

To illustrate the DFA method we show, in Figure 1, a 1,000-nucleotide subsequence of the *Enterobacteria phage lambda* (genbank name: LAMCG, with 48,502 base pair). The DFA is applied to blocks of size $l = 100$.

Finally, in the fifth step, for each $l \in \{4, 5, \ldots, g(n)\}$, we calculate the *root mean square fluctuation* (see Definition 4.1).

DEFINITION 4.1 *The root mean square fluctuation is defined by*

$$F(l) = \sqrt{\frac{1}{\tilde{n}} \sum_{t=1}^{\tilde{n}} Z_t^2}, \tag{20}$$

where Z_t is given by Equation (19) *and $\tilde{n}$ is the maximum multiple of l, smaller or equal to n, that is, $\tilde{n} = [M \cdot l] \leq n$, with $M = [n/l]$.*

Remark 4.1 In the literature an optimal choice of $g(n)$ is $[n/10]$ (see [37]). In Section 5, we consider $g(n) = [n/10]$.

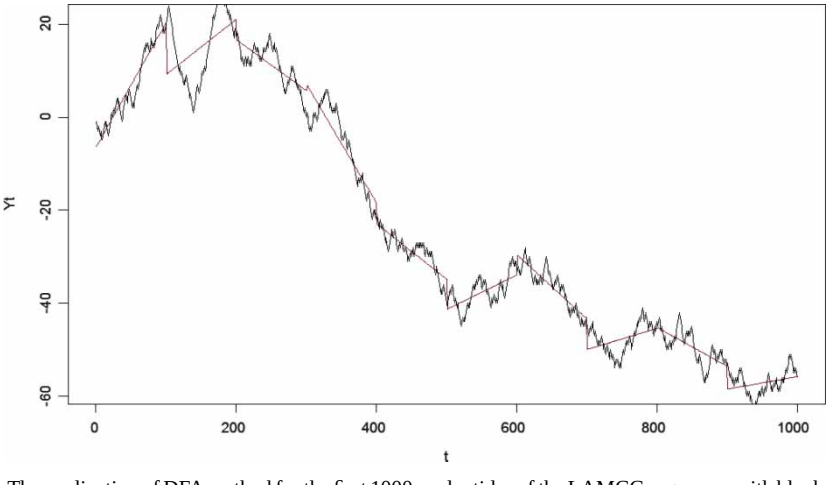

Figure 1. The application of DFA method for the first 1000 nucleotides of the LAMCG sequence, with blocks of $l = 100$ observations.

Observe that F(l), given by Equation (20), will increase with block size l. A linear relationship on a log–log plot indicates the presence of power-law scaling

$$F(l) \sim \varphi l^{\alpha}. \tag{21}$$

Under such conditions, the fluctuations can be characterized by a scaling exponent α, which is the slope of the line when one regresses $\ln(F(l))$ on $\ln(l)$, where

- $0 < \alpha < 0.5$ indicates intermediate memory;
- $\alpha = 0.5$ indicates short memory;
- $0.5 < \alpha < 1$ indicates long memory.

By taking the logarithm of the *root mean square fluctuation value*, given by Equation (21), we obtain

$$\ln(F(l)) \sim \ln(\varphi) + \alpha \ln(l). \tag{22}$$

Then, we may rewrite (22) in the context of a simple linear regression model given by Equation (7), where now

$$y_j = \ln(F(l)), \quad a = \ln(\varphi), \quad b = \alpha, \quad x_j = \ln(l) \quad \text{and} \quad l = j + 3, \tag{23}$$

with $l \in \{4, 5, \ldots, g(n)\}$ and $m = [g(n) - 3]$. Then, we obtain an estimate of α given by

$$\hat{\alpha} = \frac{\sum_{j=1}^{m}(x_j - \bar{x})y_j}{\sum_{j=1}^{m}(x_j - \bar{x})^2} = \frac{\bar{x}(1 - \bar{y})}{\dfrac{1}{m}\sum_{j=1}^{m}(x_j - \bar{x})^2}, \tag{24}$$

where $y_j = \ln(\mathrm{F}(j + 3))$, $x_j = \ln(j + 3)$, $\bar{x} = (1/m)\sum_{j=1}^{m} x_j$ and $m = [g(n) - 3]$.

4.1. *Some properties of the DFA method*

THEOREM 4.1 *If the set* $\{\epsilon_j\}_{j=1}^{m}$ *in the regression model given by the expression* (7), *with* $m = [g(n) - 3]$, *are independent and identically distributed random variables*, *with distribution function* $\mathcal{N}(0, \sigma^2)$, *then* $\hat{\alpha}$, *given by the expression* (24), *is an U.M.V.U. estimator*.

Proof For a proof see Linhares [38, p. 34]. ■

Remark 4.2 If the set $\{\epsilon_j\}_{j=1}^{m}$ in the regression model given by the expression (7), are independent and identically distributed random variables, with distribution function $\mathcal{N}(0, \sigma^2)$, then

(a) the exponent $\hat{\alpha}$, given by expression (24), is an U.M.V.U. estimator (from Theorem 4.1). Therefore $\hat{\alpha}$ is a consistent estimator;
(b) the expected value of $\hat{\alpha}$ is given by

$$\mathbb{E}(\hat{\alpha}) = \alpha;$$

(c) the variance of $\hat{\alpha}$ is given by

$$\mathrm{Var}(\hat{\alpha}) = \frac{\sum_{j=1}^{m}(x_j - \overline{x})^2 \,\mathrm{Var}(y_j)}{\left(\sum_{j=1}^{m}(x_j - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{j=1}^{m}(x_j - \bar{x})^2}.$$

Theorem 4.2 below gives an approximation for the mathematical expectation of the *mean square fluctuation value*.

THEOREM 4.2 (Taqqu *et al.* [35]) *Let $\{X_t\}_{t\in\mathbb{R}^+}$ be a fractional Gaussian noise process and let $\{X_t\}_{t=1}^n$ be a time series from this process. Then,*

$$\mathbb{E}\left(\sum_{t=1}^{l}(Y_t - Y_t^l)^2\right) \sim C_H l^{2H+1} \quad \text{as } l \to \infty, \tag{25}$$

where $Y_t = \sum_{j=1}^{t} X_j$ and

$$C_H = \left(\frac{2}{2H+1} + \frac{1}{H+2} - \frac{2}{H+1}\right). \tag{26}$$

THEOREM 4.3 *Let $\{X_t\}_{t\in\mathbb{R}^+}$ be a fractional Gaussian noise process and let $\{X_t\}_{t=1}^n$ be a time series from this process. Then,*

$$\mathbb{E}(F^2(l)) \sim C_H\, l^{2H} \quad \text{as } l \to \infty, \tag{27}$$

where $F^2(l)$ is the root mean square fluctuation value given by Equation (20) *and C_H is given by Equation* (26)*.*

Proof One observes that

$$\begin{aligned}
\mathbb{E}(\mathrm{F}^2(l)) &= \frac{1}{\tilde{n}}\mathbb{E}\left(\sum_{t=1}^{\tilde{n}} Z_t^2\right) = \frac{1}{\tilde{n}}\mathbb{E}\left(\sum_{t=1}^{\tilde{n}}(Y_t - Y_t^l)^2\right) \\
&= \frac{1}{\tilde{n}}\mathbb{E}\left(\sum_{t=1}^{l}(Y_t - Y_t^l)^2 + \sum_{t=l+1}^{2l}(Y_t - Y_t^l)^2 + \cdots + \sum_{t=[(n/l)-1]l+1}^{\tilde{n}}(Y_t - Y_t^l)^2\right) \\
&= \frac{1}{\tilde{n}}\left[\mathbb{E}\left(\sum_{t=1}^{l}(Y_t - Y_t^l)^2\right) + \mathbb{E}\left(\sum_{t=l+1}^{2l}(Y_t - Y_t^l)^2\right) + \cdots\right. \\
&\quad \left. + \mathbb{E}\left(\sum_{t=[(n/l)-1]l+1}^{\tilde{n}}(Y_t - Y_t^l)^2\right)\right].
\end{aligned} \tag{28}$$

Therefore, from Theorem 4.2 and the expression (28) we obtain

$$\mathbb{E}(F^2(l)) \sim \frac{1}{\tilde{n}}\left(C_H\, l^{2H+1} + \cdots + C_H\, l^{2H+1}\right) = \frac{1}{\tilde{n}}\frac{\tilde{n}}{l} C_H\, l^{2H+1} = C_H\, l^{2H},$$

where $F^2(l)$ is the root mean square fluctuation value given by the expression (20) and C_H is given by the expression (26). ■

Remark 4.3 By the expression (21) we obtain

$$\mathbb{E}(F^2(l)) \sim \varphi^2 l^{2\alpha}. \tag{29}$$

Comparing the expressions (29) and (27), we find $\alpha = H$. Thus, by using Equation (14) we obtain the following relationship

$$\alpha = H = d + \frac{1}{2}. \tag{30}$$

THEOREM 4.4 *Suppose that the random variables $Z_1, Z_2, \ldots, Z_{\tilde{n}}$, given by expression* (19), *are independent and identically distributed random variables with common distribution function $\mathcal{N}(0, \sigma_l^2)$. Then, $F^2(l)$, defined by the expression* (20), *has the exact distribution function $\Gamma(\tilde{n}/2, \tilde{n}/2\sigma_l^2)$.*

Proof Since the random variables $Z_1, Z_2, \ldots, Z_{\tilde{n}}$, given by the expression (19), are independent and identically distributed random variables with distribution function $\mathcal{N}(0, \sigma_l^2)$, then for each $j \in \{1, 2, \ldots, \tilde{n}\}$, the random variable Z_j/σ_l has a standard normal distribution. Therefore, the random variable $\sum_{j=1}^{\tilde{n}} Z_j^2/\sigma_l^2$ has distribution function $\chi^2(\tilde{n}) = \Gamma(\tilde{n}/2, 1/2)$, where $\tilde{n} = [\mathrm{M} \cdot l] \leq n$.

Denote $X \equiv \sum_{j=1}^{\tilde{n}} Z_j^2/\sigma_l^2$ and $Y \equiv (\sigma_l^2/\tilde{n})\, X$. Then, by using expression (20), we obtain

$$F^2(l) = \frac{1}{\tilde{n}} \sum_{j=1}^{\tilde{n}} Z_j^2 = \frac{\sigma_l^2}{\tilde{n}} \sum_{j=1}^{\tilde{n}} \frac{Z_j^2}{\sigma_l^2} = \left(\frac{\sigma_l^2}{\tilde{n}}\right) X = Y. \tag{31}$$

We know that the characteristic function uniquely determines the distribution function of a random variable. The characteristic function of the random variable Y is given by

$$\begin{aligned} \varphi_Y(t) = \mathbb{E}(\mathrm{e}^{\mathrm{i}tY}) = \mathbb{E}(\mathrm{e}^{\mathrm{i}t(\sigma_l^2/\tilde{n})X}) = \varphi_X\left(\frac{t\sigma_l^2}{\tilde{n}}\right) = \left[\frac{1}{1 - 2\mathrm{i}(t\sigma_l^2/\tilde{n})}\right]^{\tilde{n}/2} \\ = \left[\frac{1}{(\tilde{n} - 2\mathrm{i}t\sigma_l^2)/\tilde{n}}\right]^{\tilde{n}/2} = \left[\frac{\tilde{n}/2\sigma_l^2}{\tilde{n}/2\sigma_l^2 - \mathrm{i}t}\right]^{\tilde{n}/2} \quad \text{for all } t < \frac{\tilde{n}}{2\sigma_l^2}, \end{aligned} \tag{32}$$

since the random variable X has distribution $\Gamma(\tilde{n}/2, 1/2)$.

One observes that the characteristic function in expression (32) is one of the random variables with distribution function $\Gamma(\tilde{n}/2, \tilde{n}/2\sigma_l^2)$. From the uniqueness of the characteristic function, it follows that Y has distribution function $\Gamma(\tilde{n}/2, \tilde{n}/2\sigma_l^2)$, that is, $F^2(l)$, given by expression (20), has distribution function $\Gamma(\tilde{n}/2, \tilde{n}/2\sigma_l^2)$. ■

COROLLARY 4.5 *Suppose that the random variables $Z_1, Z_2, \ldots, Z_{\tilde{n}}$, given by expression* (19), *are independent and identically distributed random variables with distribution function $\mathcal{N}(0, \sigma_l^2)$. Then $F^2(l)$, given by expression* (20), *has expected value and variance, respectively, given by*

$$\mathbb{E}(F^2(l)) = \sigma_l^2 \quad \text{and} \quad \mathrm{Var}(F^2(l)) = \frac{2\sigma_l^4}{\tilde{n}}, \tag{33}$$

wherever $0 < {\sigma_l}^4 < \infty$.

Proof For a proof see Linhares [38, p.37]. ■

5. Nucleotide sequences analyses

A DNA sequence is a long polymer of simple units called nucleotides. Each nucleotide has a nitrogenous base, a deoxyribose, and a phosphate group. The denomination of the nucleotide depends on the nitrogenous bases that composes it. A DNA sequence has four nitrogenous bases: adenine (A), thymine (T), cytosine (C), and guanine (G). Adenine and guanine bases are classified as purines, and cytosine and thymine bases are classified as pyrimidines.

A nucleotide sequence $\{n_i\}_{i=1}^n$ of length n is composed of the bases A (adenine), C (cytosine), T (thymine), and G (guanine), that is, $n_i \in \{\text{A,C,T,G}\}$. To apply numerical methods to a nucleotide sequence it is necessary to transform it into a numerical sequence.

Given a nucleotide sequence $\{n_i\}_{i=1}^n \equiv \{n_1, n_2, \ldots, n_n\}$ of length n, we use the following function that transforms the nucleotide sequence $\{n_i\}_{i=1}^n$ into a numerical sequence $\{f(n_i)\}_{i=1}^n$, where $f(n_i) \in \mathbb{R}$ (see [3]).

SW Rule: We define the *transformation* $f : \{n_1, n_2, \ldots, n_n\} \to \mathbb{R}$, considering the following rule:

$$f(n_i) = \begin{cases} 1, & \text{se } n_i \in \{\text{C,G}\}, \\ 0, & \text{se } n_i \in \{\text{A,T}\}. \end{cases} \tag{34}$$

Below, we give the time series definition, representing any nucleotide sequence.

DEFINITION 5.1 *Given a nucleotide sequence* $\{n_i\}_{i=1}^n$, *the time series* $\{X_t\}_{t=1}^n$, *obtained from this sequence*, *is given by*

$$X_t = f(n_t), \tag{35}$$

where $f(\cdot)$ *is given by the expression* (34).

In this section we analyse four nucleotide sequences, available from the European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/), where the goal is to detect long dependence property on them. In order to check the performance of the different methods presented in Section 3, we consider three sequences corresponding to the *Homo sapiens* chromosome 21 (AL163202, AL163203 and AL163204, each one with 340,000 bp) and the complete sequence of the *Leishmania braziliensis* chromosome 1 (AM494938 with 235,333 bp).

Table 1. Estimators for the parameter d, with their respective confidence levels for four nucleotide sequences.

Estimator	AL163202	AL163203	AL163204	AM494938
GPH	0.1624*	0.1661*	0.1746*	0.1071*
GPH-LTS	0.1384*	0.1356*	0.1472*	0.0912*
GPH-MM	0.1740*	0.1525*	0.1673*	0.0961*
R	0.1622*	0.1658*	0.1744*	0.1062*
R-LTS	0.1374*	0.1359*	0.1471*	0.1303*
R-MM	0.1587*	0.1461*	0.1748*	0.0957*
W	0.0320*	0.0479*	0.0527*	0.0335*
R/S(n)	0.1973*	0.2372*	0.2526*	0.1058*
R/S(q)	0.2037*	0.2443*	0.2603*	0.1025*
DFA	0.2647*	0.3383*	0.3523*	0.3605*

Note:* means rejection of H_0 at 5% significance level.

We consider the following estimators for the fractional differencing parameter d: GPH, GPH-LTS, GPH-MM, R, R-LTS, R-MM, R/S(n),W, R/S(q), and DFA.

For estimating the fractional parameter d by the R/S(n) and the DFA methods, we consider the relationship among H, d and α given by the expression (30).

For each sequence and, for all estimators proposed in this work, we test the hypothesis H_0 : $d = 0$ versus $H_1 : d \neq 0$, that is, we test if the nucleotide sequences have or do not have short memory characteristics.

For each sequence, we represent graphically, the 95% confidence intervals for the fractional parameter d, using the estimators proposed in Section 3.

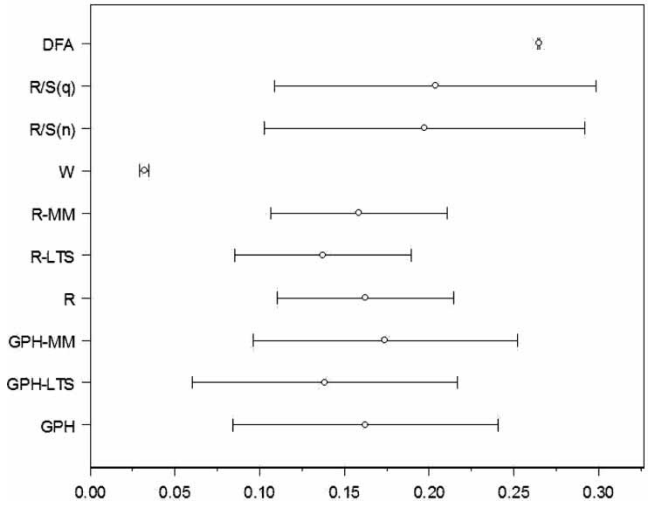

Figure 2. The 95% confidence intervals for the fractional parameter d of the sequence AL163202, based on the considered estimators.

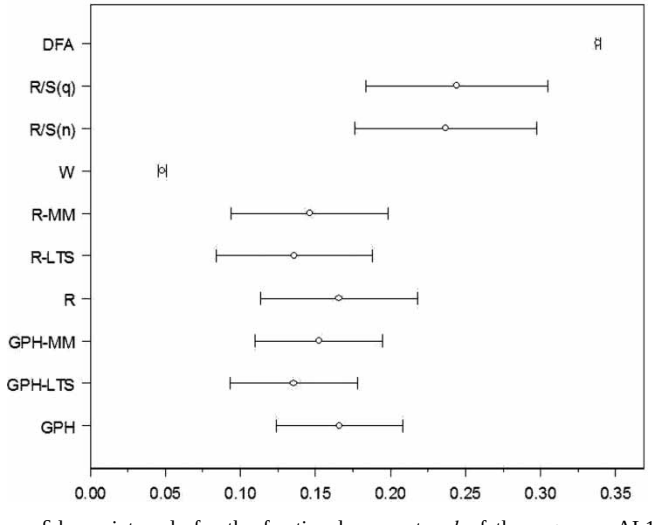

Figure 3. The 95% confidence intervals for the fractional parameter d of the sequence AL163203, based on the considered estimators.

Remark 5.1 (a) For the hypothesis test $H_0 : d = 0$ versus $H_1 : d \neq 0$, the test statistics for any estimator $\hat{d}$ is given by

$$Z = \frac{\hat{d} - d_{H_0}}{\sqrt{\mathrm{Var}(\hat{d})}} = \frac{\hat{d}}{\sqrt{\mathrm{Var}(\hat{d})}},$$

where Z has the standard normal distribution and $\sigma^2_{\hat{d}} \equiv \mathrm{Var}(\hat{d})$ is the variance of the estimator $\hat{d}$ proposed by any estimation method given in Section 3.

(b) For Table 1, we consider the following notation:
*: Rejects H_0 at 5% significance level;

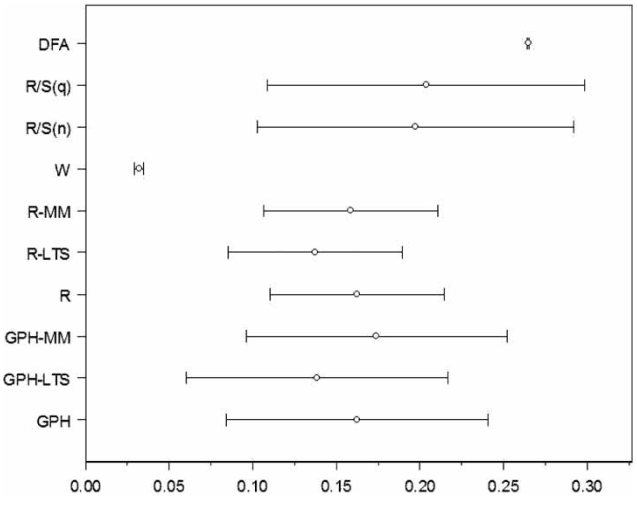

Figure 4. The 95% confidence intervals for the fractional parameter d of the sequence AL163204, based on the considered estimators.

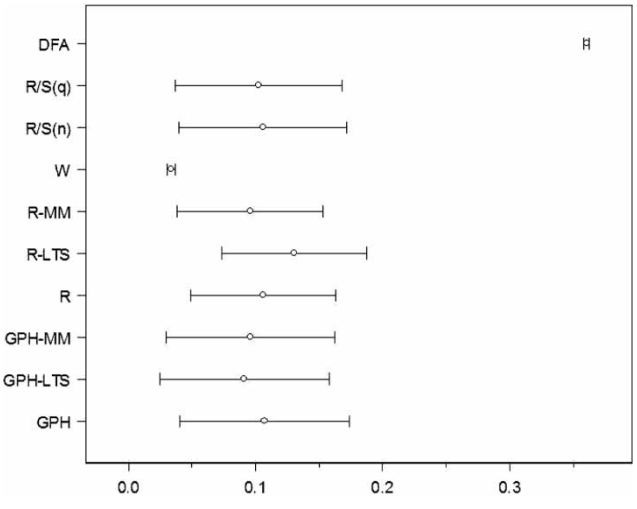

Figure 5. The 95% confidence intervals for the fractional parameter d of the sequence AM494938, based on the considered estimators.

(c) The lower and upper confidence interval values for the parameter d, based on any of the estimation methods proposed here, are given by

$$\text{lower value} = \hat{d} - z_{\frac{\alpha}{2}} \cdot \sigma_{\hat{d}},$$
$$\text{upper value} = \hat{d} + z_{\frac{\alpha}{2}} \cdot \sigma_{\hat{d}},$$

where $z_{\alpha/2} = 1.96$ and $\sigma_{\hat{d}} = \sqrt{\text{Var}(\hat{d})}$.

From Table 1, one observes that the existence of a long-range dependence for the four sequences is statistically significant at 5% level for all estimation methods considered here. Figures 2–5 represent the 95% confidence intervals for the fractional parameter d, respectively, for each sequence in Table 1.

6. Conclusions

We considered here ARFIMA(p, d, q) processes that exhibit the *long memory* property when $d \in (0.0, 0.5)$, the *short memory* property when $d = 0.0$ and the *intermediate memory* one when $d \in (-0.5, 0.0)$.

We studied several estimation methods in both semiparametric and parametric classes to estimate the fractional parameter d.

We considered the R/S(n) method ('*rescaled range*'), proposed by Hurst [22] and the '*detrended fluctuation analysis*' (DFA), proposed by Peng *et al.* [2] to estimate the fractional parameter d, by using the following relationship:

$$\alpha = \text{H} = d + \frac{1}{2},$$

where α is the scale coefficient obtained by the DFA method and H is the Hurst parameter. All three parameters in this relationship measure the long memory property.

We described the '*detrended fluctuation analysis*' and analysed its properties. This has the objective of evaluating the statistical fluctuation $F(l)$, in order to obtain a set of measures, where l represents the window length. By varying the length l, the fluctuation can be characterized by the scaling exponent, that is the slope of the line obtained by regressing $\ln(F(l))$ on $\ln(l)$. We also showed that under some conditions, the slope exponent obtained by the DFA method is a uniformly minimum variance unbiased and consistent estimator for α. To apply the DFA method, one needs to divide the time series $\{X_t\}_{t=1}^{n}$ into blocks of size l. In each block, one computes the partial sums $\{Y_t\}_{t=1}^{l}$, and then fits a least squared line $Y_t^l = a + bt$. We showed that, if the random variables $Y_1 - Y_1^l, Y_2 - Y_2^l, \ldots, Y_{\tilde{n}} - Y_{\tilde{n}}^l$, are independent and identically distributed with common distribution function $\mathcal{N}(0, \sigma_l^2)$, then $F^2(l)$ has the exact distribution function $\Gamma(\tilde{n}/2, \tilde{n}/2\sigma_l^2)$, where $\tilde{n}$ is the maximum multiple of l, smaller or equal to the sample size. We observed that ${\sigma_l}^2$ is the theoretical variance of the random variables $Y_j - Y_j^l$, $j = 1, \ldots, \tilde{n}$. We proved that $F^2(l)$ is unbiased for the variance σ_l^2 and, if $0 < \sigma_l^4 < \infty$, the statistic $F^2(l)$ is a consistent estimator for σ_l^2 and it has minimum variance as $\tilde{n}$ tends to infinity.

According to the results of the estimation methods discussed in Section 3, all four nucleotide sequences studied here display long-range dependence. For each sequence, this conclusion is statistically significant at the 5% level for all estimators proposed.

Acknowledgements

N. Crato acknowledges the partial support by FCT-Fundação para a Ciência e Tecnologia (Programme FEDER/POCI 2010), Portugal. The work of R.R. Linhares was supported by CNPq-Brazil. S.R.C. Lopes acknowledges the partial supported by CNPq-Brazil, by CAPES-Brazil, by *Millennium Institute in Probability* and also by Pronex *Probabilidade e Processos Estocásticos* - E-26/170.008/2008 -APQ1.

The authors thank two anonymous referees and both the editor and the associate editor for their valuable comments and suggestions that improved the final version of the manuscript.

References

[1] J. Beran, *Statistics for Long Memory Processes*, Chapman & Hall, New York, 1994.
[2] C. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, and A.L. Goldberger, *Mosaic organization of DNA nucleotides*, Phys. Rev. E 49(5) (1994), pp. 1685–1689.
[3] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.K. Peng, M. Simons, and H.E. Stanley, *Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis*, Phys. Rev. E 51(5) (1995), pp. 5084–5091.
[4] J.K. Percus, *Mathematics of Genome Analysis*, Cambridge University Press, Cambridge, 2002.
[5] S.W.A. Bergen and A. Antoniou, *Application of parametric window functions to the STDFT method for gene prediction*, IEEE Pacific Rim Conf. Commun. Comput. Signal Process. 1 (2005), pp. 324–327.
[6] C.A. Chatzidimitriou-Dreismann and D. Larhammar, *Long-range correlations in DNA*, Nature 361 (1993), pp. 212.
[7] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.K. Peng, and M. Simons, *Scaling features of noncoding DNA*, Phys. A 273(1) (1999), pp. 1–18.
[8] Z-G. Yu, V. V. Anh, and B. Wang, *Correlation property of length sequences based on global structure of complete genome*, Phys. Rev. E 63 (2000), 011903.
[9] B. Audit, C. Vaillant, A. Arneodo, Y. d'Aubenton-Carafa, and C. Thermes, *Long-range correlations between DNA bending sites: Relation to the structure and dynamics of nucleosomes*, J. Mol. Biol. 316 (2002), pp. 903–918.
[10] S.R.C. Lopes and M.A. Nunes, *Long memory analysis in DNA sequences*, Phys. A 361(2) (2006), pp. 569–588.
[11] W. Li and K. Kaneko, *Long-range correlation and partial* $1/f^{\alpha}$ *spectrum in a noncoding DNA sequence*, Europhys. Lett. 17(7) (1992), pp. 655–660.
[12] R.F. Voss, *Evolution of long-range fractal correlations and 1/f noise in DNA base sequences*, Phys. Rev. Lett. 68 (1992), pp. 3805–3808.
[13] G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, and J. Salinas, *The mosaic genome of warm-blooded vertebrates*, Science 228 (1985), pp. 953–958.
[14] G. Bernardi, *Structural and Evolutionary Genomics*, Elsevier, Amsterdam, 2004.
[15] J.L. Oliver, P. Carpena, M. Hackenberg, and P. Bernaola-Galván, *IsoFinder: computational prediction of isochores in genome sequences*, Nucleic Acids Res. 32 (2004), pp. W287–W292.
[16] S. Karlin and V. Brendel, *Patchiness and correlations in DNA sequences*, Science 259 (5095) (1993), pp. 677–680.
[17] P. Carpena, P.B. Galván, A.V. Coronado, M. Hackenberg, and J.L. Oliver, *Identifying characteristic scales in the human genome*, Phys. Rev. E 75 (2007), 032903.
[18] C. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley, *Long-range correlations in nucleotide sequences*, Nature 356 (1992), pp. 168–170.
[19] Y.H. Liu, P. Cizeau, M. Meyer, C.-K. Peng, and H.E. Stanley, *Correlations in economic time series*, Phys. A, 245 (1997), pp. 437–440.
[20] R.-G. Yeh, J-S. Shieh, Y-Y. Han, Y-J. Wang, and S-C. Tseng, *Detrended fluctuation analyses of short-term heart rate variability in surgical intensive care units*, Biomed. Eng.—Applications, Basis & Communications, 18 (2006), pp. 67–72.
[21] E. Koscielny-Bunde, H.E. Roman, A. Bunde, S. Havlin, and H.-J. Schellnhuber, *Long-range power-law correlations in local daily temperature fluctuations*, Phil. Mag. B 77 (1998), pp. 1331–1340.
[22] H.R. Hurst, *Long-term storage in reservoirs*, Trans. Am. Soc. Civil Eng. 116 (1951), pp. 770–799.
[23] J. Hosking, *Fractional differencing*, Biometrika 68 (1981), pp. 165–167.
[24] P. Doukhan, G. Oppenheim, and M.S. Taqqu, *Theory and Applications of Long-Range Dependence*, Birkhäuser, Boston, 2003.
[25] R. Fox and M.S. Taqqu, *Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time Series*, Ann. Stat. 14 (1986), pp. 517–532.
[26] P. Whittle, Hypothesis Testing in Time Series Analysis, Hafner, New York, 1953.
[27] J. Geweke and S. Porter-Hudak, *The estimation and application of long memory time series model*, J. Time Ser. Anal. 4(4) (1983), pp. 221–238.
[28] P.M. Robinson, *Log-periodogram regression of time series with long range dependence*, Ann. Stat. 23(3) (1995), pp. 1048–1072.
[29] S.R.C Lopes and B.V.M. Mendes, *Bandwidth selection in classical and robust estimation of long memory*, Int. J. Stat. Systems 1(2) (2006), pp. 167–190.
[30] N. Crato and B.K. Ray, *Semiparametric smoothing estimators for long memory processes with added noise*, J. Stat. Plan. Inference 105 (2002), pp. 283–297.

[31] P.J. Rousseeuw, *Least median of squares regression*, J. Am. Stat. Assoc. 79 (1984), pp. 871–880.
[32] V.J. Yohai, *High breakdown point and high efficiency robust estimates for regression*, Ann. Stat. 15 (1987), pp. 642–656.
[33] A.W. Lo, *Long term memory in stock market prices*, Econometrica 59 (1991), pp. 1279–1313.
[34] V. Teverovsky, M.S. Taqqu, and W. Willinger, *A critical look at Lo's modified R/S statistic*, J. Stat. Plan. Inference 80 (1999), pp. 211–227.
[35] M.S. Taqqu, V. Teverovsky, and W. Willinger, *Estimators for long range dependence: An empirical study*, Fractals 3(4) (1995), pp. 785–798.
[36] L. Giraitis, P. Kokoszka, R. Leipus, and G. Teyssière, *On the power of the R/S-type tests against contiguous and semi long memory alternatives*, Actae Appl. Math. 78 (2003), pp. 285–299.
[37] K. Hu, P.C. Ivanov, Z. Chen, P. Carpena, and H.E. Stanley, *Effect of trends on detrended fluctuation analysis*, Phys. Rev. E 64 (2001), 011114.
[38] R.R. Linhares, *Propriedades Estatísticas do Método da Análise de Flutuações Destendenciadas em Seqüências de DNA*. Tese de Mestrado, Programa de Pós-Graduação em Matemática, Instituto de Matemática, UFRGS, Porto Alegre, 2007.