Statistical Properties of Detrended Fluctuation Analysis

N. Crato[a], R.R. Linhares[b] and S.R.C. Lopes[b1]

[a]Cemapre, ISEG, Technical University of Lisbon
Lisbon - Portugal

[b]Mathematical Institute - UFRGS
Porto Alegre - RS - Brazil

January 18, 2009

Abstract

The main goal of this work is to consider the *detrended fluctuation analysis* (DFA), proposed by Peng et al. (1994). This is a well known method for analyzing 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 analyze 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 (1951) and the maximum likelihood method (see Fox and Taqqu (1986), by considering the approximation suggested by Whittle (1953).

AMS Classification: Primary 91B70; Secondary 00A05.

Keywords: Long Memory, Detrended Fluctuation Analysis, Semiparametric Estimation, Robustness.

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 (1994). One of the models that exhibits 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 Peng et al., 1994). To obtain a time series from a nucleotide sequence it is necessary to consider some type of transformation (see Buldyrev et al., 1995).

[1]Corresponding author. E-mail: silvia.lopes@ufrgs.br

The statistical properties of DNA genomes are of interest because they reflect biological features (see Percus, 2002). 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 Bergen and Antoniou, 2005). Consequently, this property can be used to help identifying coding regions.

Several papers (see Peng et al., 1994; Chatzidimitriou-Dreismann and Larhammar, 1993; Buldyrev et al. 1995; Stanley et al., 1999; Yu et al., 2000; Audit et al., 2002; and Lopes and Nunes, 2006, among others) study the existence of long-range or power-law correlations in DNA sequences. Peng et al. (1994), Li and Kaneko (1992) and Voss (1992) 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 Bernardi et al., 1985; Bernardi, 2004 and Oliver et al., 2004). For some authors the existence of long range power-law correlations seems to be related to such "isochore" segments (see Karlin and Brendel, 1993). Carpena et al. (2007) argue that the DNA correlations are much more complex than power-laws with a single scaling exponent. In fact, these authors propose to analyze 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 nonfractal behavior 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. (2004) explore the phylogenetic distribution of large-scale genome patchiness by considering the deviations of the power-law behavior in long-range correlations.

In the literature the "*Detrended Fluctuation Analysis*" (DFA), proposed by Peng et al. (1994), has successfully been applied to different fields of interest, such as DNA sequences (see Buldyrev et al., 1995 and Peng et al., 1992), economical time series (see Liu et al., 1997) heart rate variability analysis (see Yeh et al., 2006) and long-time weather records (see Koscielny-Bunde et al., 1998). The DFA method 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 Hurst, 1951) and also similar to another method based on wavelet transform (see Koscielny-Bunde et al., 1998). 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 analyze the statistical properties of the DFA method. We are interested in analyzing 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

$$\Phi(\mathcal{B})(1-\mathcal{B})^d(X_t - \mu) = \Theta(\mathcal{B})\varepsilon_t,\ t \in \mathbb{Z}, \tag{2.1}$$

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

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

given by

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

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

$$f_X(w) = f_U(w)\left[2\,\sin(\frac{w}{2})\right]^{-2d},\ \text{for } 0 < w \leq \pi, \tag{2.2}$$

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 (2.1), is the binomial expansion

$$(1-\mathcal{B})^d = \sum_{k=0}^{\infty}\binom{d}{k}(-\mathcal{B})^k = 1 - d\mathcal{B} - \frac{d}{2!}(1-d)\mathcal{B}^2\cdots,\ \text{for } d \in \mathbb{R}. \tag{2.3}$$

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 (1981), Beran (1994) and Doukhan et al. (2003).

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 (1951) and the maximum likelihood method (see Fox and Taqqu, 1986), by considering the approximation suggested by Whittle (1953).

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 (1983), both classical and robust versions.
- The semiparametric regression method based on GPH with trimming l and bandwidth $g(n)$, proposed by Robinson (1995), both classical and robust versions.
- The semiparametric method based on Hurst (1951) 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 (2.1). Taking the logarithm of the spectral density function $f_X(\cdot)$ given by (2.2), we have

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

or writing

$$\ln(f_X(w)) = \ln(f_U(0)) - d\ln\left(4\sin^2\left(\frac{w}{2}\right)\right) + \ln\left(\frac{f_U(w)}{f_U(0)}\right). \tag{3.1}$$

Substituting w by $w_j = \frac{2\pi j}{n}$ and adding $\ln(I_n(w_j))$ to both sides of (3.1), we obtain

$$\ln(I_n(w_j)) = \ln(f_U(0)) - d\ln\left(4\sin^2\left(\frac{w_j}{2}\right)\right) + \ln\left(\frac{f_U(w_j)}{f_U(0)}\right) + \ln\left(\frac{I_n(w_j)}{f_X(w_j)}\right), \tag{3.2}$$

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

$$I_n(w) = \frac{1}{2\pi}\left(\widehat{\gamma}_X(0) + 2\sum_{k=1}^{n-1}\widehat{\gamma}_X(k)\cos(wk)\right), \quad w \in (0, \pi], \tag{3.3}$$

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 $\ln\left(\frac{f_U(w_j)}{f_U(0)}\right)$ may be discarded (see Geweke and Porter-Hudak, 1983). Then, we may rewrite (3.2) in the context of a simple linear regression model

$$y_j = a + bx_j + \epsilon_j, \qquad j = 1, \cdots, g(n), \tag{3.4}$$

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

A semiparametric regression estimator (see Lopes and Mendes, 2006 and Crato and Ray, 2002) may be obtain by minimizing some loss function of the residuals

$$r_j = y_j - a - bx_j, \quad \text{for} \quad j = 1, \cdots, g(n). \tag{3.5}$$

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 (1984) and the MM method, proposed by Yohai (1987).

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

$$L_1(g(n)) = \sum_{j=1}^{g(n)} (r_j)^2, \tag{3.6}$$

where r_j is given by expression (3.5), for $j \in \{1, \cdots, g(n)\}$.

Definition 3.2. The *Robust Estimators LTS* (see Rousseeuw, 1984) are the values $(\hat{a}, \hat{b})$ that minimize the loss function

$$L_2(g(n)) = \sum_{j=1}^{g^*(n)} (r^2)_{j:g(n)}, \tag{3.7}$$

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 Yohai, 1987) are the values $(\hat{a}, \hat{b})$ that minimize the loss function

$$L_3(g(n)) = \sum_{j=1}^{g(n)} \rho_2\left(\frac{r_j}{s}\right)^2, \tag{3.8}$$

subject to the constraint

$$\frac{1}{g(n)} \sum_{j=1}^{g(n)} \rho_1\left(\frac{r_j}{s}\right) \leq C, \tag{3.9}$$

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 (1983). To obtain an estimate for d, these authors propose to apply the Ordinary Least Squares method (see Lopes and Mendes, 2006) in (3.5) based on (3.3), which we denote it by GPH. The estimator of d is given by

$$\text{GPH} = -\frac{\sum_{j=1}^{g(n)}(x_j - \overline{x})(y_j - \overline{y})}{\sum_{j=1}^{g(n)}(x_j - \overline{x})^2}, \tag{3.10}$$

where

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

The variance of the GPH estimator (see Geweke and Porter-Hudak, 1983) is given by

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

To obtain the robust version of the GPH estimator we just apply the Least Trimmed Squared (LTS) and MM methodologies (see Lopes and Mendes, 2006), respectively, to the regression model (3.5). This gives rise to the GPH-LTS and the GPH-MM estimators.

3.1.2 R, R-LTS e R-MM Estimators

The regression estimator, proposed by Robinson (1995), is obtained by applying the Ordinary Least Squared method in (3.5) based on (3.3), but considering only the frequencies ω_j, for $j \in \{l, l+1, \cdots, g(n)\}$, where $l > 1$ is a trimming value that tends to infinity more slowly than $g(n)$.

The asymptotic variance of the estimator R (see Robinson, 1995) is given by

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

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 Lopes and Mendes, 2006) to the regression model (3.5).

3.1.3 R/S(n) and R/S(q) Estimators

Here we introduced the R/S(n) *statistic* proposed by Hurst (1951) and a modified version of it, denoted by R/S(q) and proposed by Lo (1991).

Definition 3.4. Let $\{X_t\}_{t=1}^n$ be a time series. The *rescaled range statistic R/S*(n), introduced by Hurst (1951), is defined by

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

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

For the fractional Gaussian noise process or the ARFIMA process (see Teverovsky et al., 1999),

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

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 $j \in \{1, \cdots, s\}$, one divides the time series $\{X_t\}_{t=1}^n$ into $\left[\frac{n}{k_j}\right]$ blocks, each one of size k_j, where $k_j = \ell k_{j-1}$.
- For each block, one computes the R/S(k_j) statistic.
- One adjusts a regression line, by regressing $\ln(\mathrm{R/S}(k_j))$ on $\ln(k_j)$, $j = 1, \cdots, s$, 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 Taqqu et al., 1995)

$$d = H - \frac{1}{2}. \tag{3.11}$$

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

$$\widehat{\sigma}_n^2(q) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2 + \frac{2}{n}\sum_{j=1}^{q}\omega_j(q)\Big(\sum_{l=j+1}^{n}(X_l - \overline{X})(X_{l-j} - \overline{X})\Big), \tag{3.12}$$

where $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ 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 Giraitis et al. 2003).

Definition 3.6. The R/S(n) *modified statistic*, proposed by Lo (1991) and denoted by R/S(q), is defined as

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

where $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ and $\widehat{\sigma}_n(q)$ is defined in (3.12).

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 Fox and Taqqu, 1986), by considering the approximation suggested by Whittle (1953).

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

$$\boldsymbol{\eta} = (\sigma_X^2, d, \phi_1, \phi_2, \cdots, \phi_p, \theta_1, \theta_2, \cdots, \theta_q) \tag{3.13}$$

that minimizes the function

$$\mathcal{Q}(\boldsymbol{\eta}) = \sum_{j=1}^{[\frac{n-1}{2}]}\left(\frac{I(w_j)}{f_X(w_j,\boldsymbol{\eta})}\right), \tag{3.14}$$

where $\boldsymbol{\eta}$ is the vector of unknown parameters given in (3.13), $f_X(\cdot,\boldsymbol{\eta})$ is the spectral density function of the $\{X_t\}_{t\in\mathbb{Z}}$, $[x]$ is the integer part of x, $w_j = \frac{2\pi j}{n}$ is the Fourier frequencies, for $j \in \{1,\cdots,[\frac{n-1}{2}]\}$, and $I(\cdot)$ is the periodogram function given by (3.3).

The asymptotic variance of the estimator W (see Fox and Taqqu, 1986) is given by

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

More details on this estimator can be found in Fox and Taqqu (1986) and Beran (1994).

4 DFA Method and Some Properties

Given a time series $\{X_t\}_{t=1}^n$, the *Detrended Fluctuation Analysis* (DFA), proposed by Peng et al. (1994), consists on five steps. In the first one, for each $t \in \{1,2,\cdots,n\}$, we calculate

$$Y_t = \sum_{j=1}^{t} X_j. \tag{4.1}$$

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 $\left[\frac{n}{l}\right]$ nonoverlapping 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

$$Z_t = Y_t - Y_t^l, \tag{4.2}$$

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

To illustrate the DFA method we show, in Figure 4.1, a 1,000-nucleotide subsequence of the *Enterobacteria phage lambda* (genbank name: LAMCG, with 48,502 base pair). The *"Detrended Fluctuation Analysis"* (DFA) is applied to blocks of size $l = 100$.

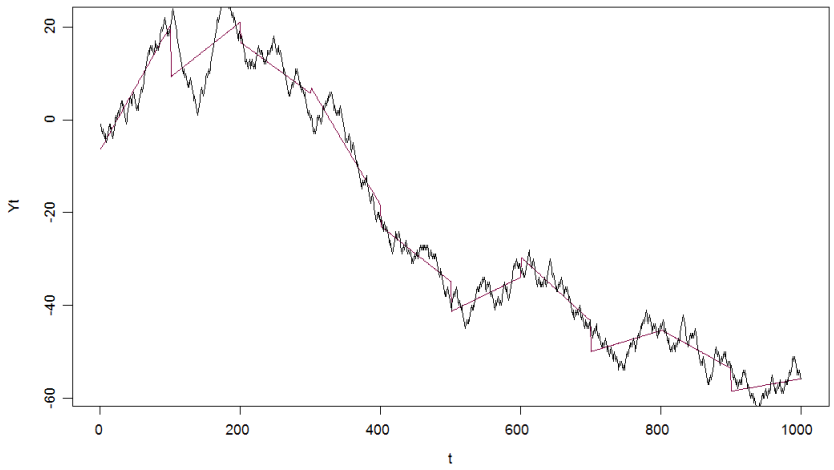

Figure 4.1: The application of DFA method for the first 1,000 nucleotides of the LAMCG sequence, with blocks of $l = 100$ observations.

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

Definition 4.1. The *root mean square fluctuation* is defined by

$$\mathrm{F}(l) = \sqrt{\frac{1}{\widetilde{n}} \sum_{t=1}^{\widetilde{n}} Z_t^2}, \tag{4.3}$$

where Z_t is given by (4.2) and $\widetilde{n}$ is the maximum multiple of l, smaller or equal to n, that is, $\widetilde{n} = [\mathrm{M} \cdot l] \leq n$, with $\mathrm{M} = [n/l]$.

Remark 4.1. In the literature an optimal choice of $g(n)$ is $\left[\frac{n}{10}\right]$ (see Hu et al., 2001). In Section 5, we consider $g(n) = \left[\frac{n}{10}\right]$.

Observe that $\mathrm{F}(l)$, given by (4.3), will increase with block size l. A linear relationship on a $\log - \log$ plot indicates the presence of power law scaling

$$\mathrm{F}(l) \sim \varphi l^{\alpha}. \tag{4.4}$$

Under such condition, 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 (4.4), we obtain

$$\ln(\mathrm{F}(l)) \sim \ln(\varphi) + \alpha \ln(l). \tag{4.5}$$

Then, we may rewrite (4.5) in the context of a simple linear regression model given by (3.4), where now

$$y_j = \ln(\mathrm{F}(l)), \quad a = \ln(\varphi), \quad b = \alpha, \quad x_j = \ln(l) \quad \text{and} \qquad l = j + 3, \tag{4.6}$$

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

$$\widehat{\alpha} = \frac{\displaystyle\sum_{j=1}^{m}(x_j - \overline{x})y_j}{\displaystyle\sum_{j=1}^{m}(x_j - \overline{x})^2} = \frac{\overline{x}(1-\overline{y})}{\frac{1}{m}\displaystyle\sum_{j=1}^{m}(x_j - \overline{x})^2}, \tag{4.7}$$

where $y_j = \ln(\mathrm{F}(j+3))$, $x_j = \ln(j+3)$, $\overline{x} = \dfrac{1}{m}\displaystyle\sum_{j=1}^{m} x_j$ 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 (3.4), with $m = g(n) - 3$, are independent and identically distributed random variables, with distribution function $\mathcal{N}(0, \sigma^2)$, then $\widehat{\alpha}$, given by the expression (4.7), is an* U.M.V.U. *estimator.*

Proof. For a proof see Linhares (2007), page 34. □

Remark 4.2. If the set $\{\epsilon_j\}_{j=1}^m$ in the regression model given by the expression (3.4), are independent and identically distributed random variables, with distribution function $\mathcal{N}(0, \sigma^2)$, then

a) the exponent $\widehat{\alpha}$, given by expression (4.7), is an U.M.V.U. estimator (from Theorem 4.1). Therefore $\widehat{\alpha}$ is a consistent estimator;

b) the expected value of $\widehat{\alpha}$ is given by

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

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

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

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

Theorem 4.2. *[Taqqu et al. (1995)] 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,*

$$\mathbb{E}\Big(\sum_{t=1}^{l}(Y_t - Y_t^l)^2\Big) \sim C_{\mathrm{H}}\, l^{2\mathrm{H}+1}, \quad \text{as } l \to \infty, \tag{4.8}$$

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

$$C_{\mathrm{H}} = \Big(\frac{2}{2\mathrm{H}+1} + \frac{1}{\mathrm{H}+2} - \frac{2}{\mathrm{H}+1}\Big). \tag{4.9}$$

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,*

$$\mathbb{E}(\mathrm{F}^2(l)) \sim C_{\mathrm{H}}\, l^{2\mathrm{H}}, \quad \text{as } l \to \infty, \tag{4.10}$$

where $\mathrm{F}^2(l)$ is the root mean square fluctuation value given by (4.3) and C_{H} is given by (4.9).

Proof. One observes that

$$\begin{aligned}
\mathbb{E}(\mathrm{F}^2(l)) &= \frac{1}{\widetilde{n}}\mathbb{E}\Big(\sum_{t=1}^{\widetilde{n}} Z_t^2\Big) = \frac{1}{\widetilde{n}}\mathbb{E}\Big(\sum_{t=1}^{\widetilde{n}}(Y_t - Y_t^l)^2\Big) \\
&= \frac{1}{\widetilde{n}}\mathbb{E}\Big(\sum_{t=1}^{l}(Y_t - Y_t^l)^2 + \sum_{t=l+1}^{2l}(Y_t - Y_t^l)^2 + \cdots + \\
&+ \sum_{t=[(n/l)-1]l+1}^{\widetilde{n}}(Y_t - Y_t^l)^2\Big) \\
&= \frac{1}{\widetilde{n}}\Big[\mathbb{E}\Big(\sum_{t=1}^{l}(Y_t - Y_t^l)^2\Big) + \mathbb{E}\Big(\sum_{t=l+1}^{2l}(Y_t - Y_t^l)^2\Big) + \cdots + \\
&+ \mathbb{E}\Big(\sum_{t=[(n/l)-1]l+1}^{\widetilde{n}}(Y_t - Y_t^l)^2\Big)\Big].
\end{aligned} \tag{4.11}$$

Therefore, from Theorem 4.2 and the expression (4.11) we obtain

$$\mathbb{E}(\mathrm{F}^2(l)) \sim \frac{1}{\widetilde{n}}\Big(C_{\mathrm{H}}\, l^{2\mathrm{H}+1} + \cdots + C_{\mathrm{H}}\, l^{2\mathrm{H}+1}\Big) = \frac{1}{\widetilde{n}}\frac{\widetilde{n}}{l}C_{\mathrm{H}}\, l^{2\mathrm{H}+1} = C_{\mathrm{H}}\, l^{2\mathrm{H}},$$

where $\mathrm{F}^2(l)$ is the root mean square fluctuation value given by the expression (4.3) and C_{H} is given by the expression (4.9).

□

Remark 4.3. By the expression (4.4) we obtain

$$\mathbb{E}(F^2(l)) \sim \varphi^2 l^{2\alpha}. \tag{4.12}$$

Comparing the expressions (4.12) and (4.10), we find $\alpha = \mathrm{H}$. Thus, by using the equation (3.11) we obtain the following relationship

$$\alpha = \mathrm{H} = d + \frac{1}{2}. \tag{4.13}$$

Theorem 4.4. *Suppose that the random variables $Z_1, Z_2, \cdots, Z_{\widetilde{n}}$, given by expression (4.2), are independent and identically distributed random variables with common distribution function $\mathcal{N}(0, \sigma_l^2)$. Then, $\mathrm{F}^2(l)$, defined by the expression (4.3), has the exact distribution function $\Gamma\big(\frac{\widetilde{n}}{2}, \frac{\widetilde{n}}{2\sigma_l^2}\big)$.*

Proof. Since the random variables $Z_1, Z_2, \cdots, Z_{\widetilde{n}}$, given by the expression (4.2), are independent and identically distributed random variables with distribution function $\mathcal{N}(0, \sigma_l^2)$, then for each $j \in \{1, 2, \cdots, \widetilde{n}\}$, the random variable $\frac{Z_j}{\sigma_l}$ has a standard Normal distribution. Therefore, the random variable $\displaystyle\sum_{j=1}^{\widetilde{n}} \frac{Z_j^2}{\sigma_l^2}$ has distribution function $\chi^2(\widetilde{n}) = \Gamma(\frac{\widetilde{n}}{2}, \frac{1}{2})$, where $\widetilde{n} = [\mathrm{M} \cdot l] \leq n$.

Denote $X \equiv \displaystyle\sum_{j=1}^{\widetilde{n}} \frac{Z_j^2}{\sigma_l^2}$ and $Y \equiv \left(\frac{\sigma_l^2}{\widetilde{n}}\right) X$. Then, by using the expression (4.3), we obtain

$$\mathrm{F}^2(l) = \frac{1}{\widetilde{n}} \sum_{j=1}^{\widetilde{n}} Z_j^2 = \frac{\sigma_l^2}{\widetilde{n}} \sum_{j=1}^{\widetilde{n}} \frac{Z_j^2}{\sigma_l^2} = \left(\frac{\sigma_l^2}{\widetilde{n}}\right) X = Y. \tag{4.14}$$

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

$$\varphi_Y(t) = \mathbb{E}\big(e^{itY}\big) = \mathbb{E}\left(e^{it\frac{\sigma_l^2}{\widetilde{n}}X}\right) = \varphi_X\left(\frac{t\sigma_l^2}{\widetilde{n}}\right) = \left[\frac{1}{1 - 2i\big(\frac{t\sigma_l^2}{\widetilde{n}}\big)}\right]^{\frac{\widetilde{n}}{2}}$$

$$= \left[\frac{1}{\frac{\widetilde{n} - 2it\sigma_l^2}{\widetilde{n}}}\right]^{\frac{\widetilde{n}}{2}} = \left[\frac{\frac{\widetilde{n}}{2\sigma_l^2}}{\frac{\widetilde{n}}{2\sigma_l^2} - it}\right]^{\frac{\widetilde{n}}{2}}, \quad \text{for all } t < \frac{\widetilde{n}}{2\sigma_l^2}, \tag{4.15}$$

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

One observes that the characteristic function in the expression (4.15), is one of a random variable with distribution function $\Gamma\big(\frac{\widetilde{n}}{2}, \frac{\widetilde{n}}{2\sigma_l^2}\big)$. From the uniqueness of the characteristic function, it follows that Y has distribution function $\Gamma\big(\frac{\widetilde{n}}{2}, \frac{\widetilde{n}}{2\sigma_l^2}\big)$, that is, $F^2(l)$, given by expression (4.3), has distribution function $\Gamma\big(\frac{\widetilde{n}}{2}, \frac{\widetilde{n}}{2\sigma_l^2}\big)$. □

Corollary 4.1. *Suppose that the random variables $Z_1, Z_2, \cdots, Z_{\widetilde{n}}$, given by expression (4.2), are independent and identically distributed random variables with distribution function $\mathcal{N}(0, \sigma_l^2)$. Then $F^2(l)$, given by expression (4.3), has expected value and variance, respectively, given by*

$$\mathbb{E}(\mathrm{F}^2(l)) = \sigma_l^2 \quad \textit{and} \quad \mathrm{Var}(\mathrm{F}^2(l)) = \frac{2\sigma_l^4}{\widetilde{n}}, \tag{4.16}$$

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

Proof. For a proof see Linhares (2007), page 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 nitrogenized basis that composes it. A DNA sequence has four nitrogenous basis: adenine (A), thymine (T), cytosine (C) and guanine (G). Adenine and guanine basis, are classified as purines and cytosine and thymine basis are classified as pyrimidines.

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

Given a nucleotide sequence $\{n_i\}_{i=1}^n \equiv \{n_1, n_2, \cdots, n_n\}$ 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 Buldyrev et al, 1995).

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

$$f(n_i) = \begin{cases} 1, & \text{se } n_i \in \{\mathrm{C,G}\} \\ 0, & \text{se } n_i \in \{\mathrm{A,T}\}. \end{cases} \tag{5.1}$$

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

$$X_t = f(n_t), \tag{5.2}$$

where $f(\cdot)$ is given by the expression (5.1).

In this section we analyze 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).

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 following relationship among H, d and α given by the expression (4.13).

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.

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

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

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

b) For Table 5.1, we consider the following notation:

* : Rejects H_0 at 5% significance level;

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

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

where $z_{\frac{\alpha}{2}} = 1.96$ and $\sigma_{\widehat{d}} = \sqrt{Var(\widehat{d})}$.

Table 5.1: 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.

From Table 5.1 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.

The Figures 5.1 to 5.4 represent the 95% confidence intervals for the fractional parameter d, respectively, for each sequence in Table 5.1.

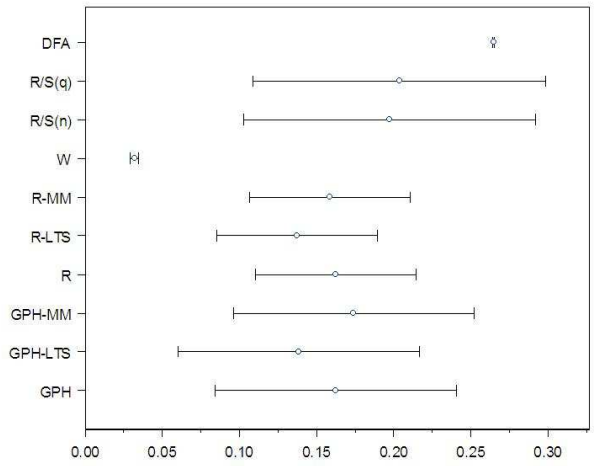

Figure 5.1: The 95% Confidence Intervals for the Fractional Parameter d of the Sequence AL163202, Based on the Considered Estimators.

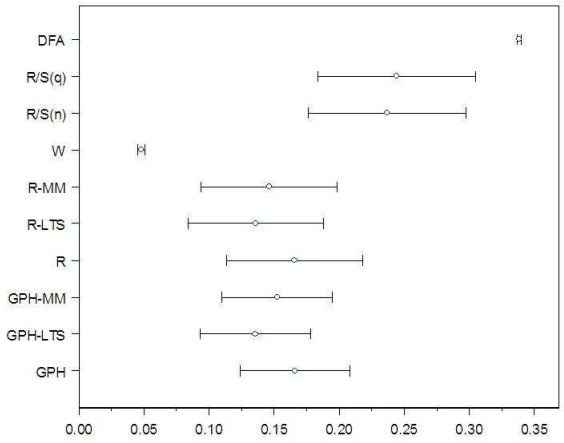

Figure 5.2: The 95% Confidence Intervals for the Fractional Parameter d of the Sequence AL163203, Based on the Considered Estimators.

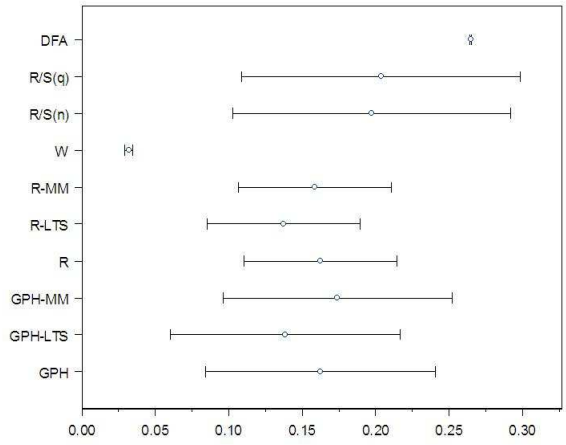

Figure 5.3: The 95% Confidence Intervals for the Fractional Parameter d of the Sequence AL163204, Based on the Considered Estimators.

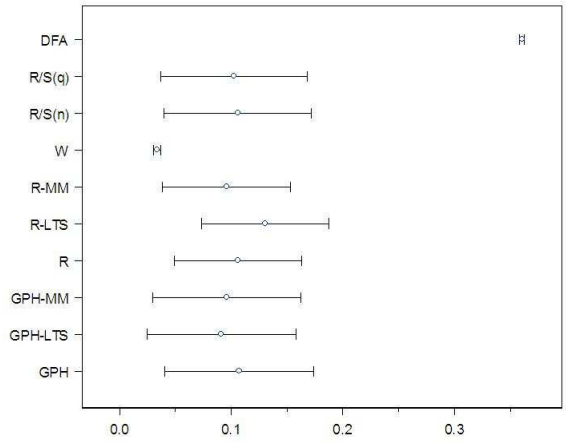

Figure 5.4: The 95% Confidence Intervals for the Fractional Parameter d of the Sequence AM494938, Based on the Considered Estimators.

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 (1951) and the "*Detrended Fluctuation Analysis"* (DFA), proposed by Peng et al. (1994) to estimate the fractional parameter d, by using the following relationship

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

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 analyzed 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 $Y_1 - Y_1^l, Y_2 - Y_2^l, \cdots, Y_{\widetilde{n}} - Y_{\widetilde{n}}^l$, are independent and identically distributed with common distribution function $\mathcal{N}(0, \sigma_l^2)$, then $\mathrm{F}^2(l)$ has the exact distribution function $\Gamma\big(\frac{\widetilde{n}}{2},\ \frac{\widetilde{n}}{2\sigma_l^2}\big)$, where $\widetilde{n}$ is is the maximum multiple of l, smaller or equal to the length of the sample size. We observed that ${\sigma_l}^2$ is the theoretical variance of the random variables $Y_j - Y_j^l$, $j = 1, \cdots, \widetilde{n}$. 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 $\widetilde{n}$ 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.

Acknowledgments

N. Crato research was partially supported by FCT-Fundação para a Ciência e Tecnologia (Programme FEDER/POCI 2010), Portugal. R.R. Linhares was supported by CNPq-Brazil. S.R.C. Lopes research was partially 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 would like to 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

Audit, B., C. Vaillant, A. Arneodo, Y. d'Aubenton-Carafa and C. Thermes (2002). "Long-range Correlations between DNA Bending Sites: Relation to the Structure and Dynamics of Nucleosomes". *Journal of Molecular Biology*, Vol **316**, 903-918.

Beran, J. (1994). *Statistics for Long Memory Processes*. New York: Chapman & Hall.

Bergen, S.W.A. and A. Antoniou (2005). "Application of parametric window functions to the STDFT method for gene prediction". *IEEE Pacific Rim Conf. on Communications, Computers and Signal Processing*, Vol. **1**, 324-327.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial and J. Salinas (1985). "The mosaic genome of warm-blooded vertebrates". *Science*, Vol. **228**, 953-958.

Bernardi, G. (2004). *Structural and Evolutionary Genomics*. Elsevier: Amsterdam.

Buldyrev, S.V., A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.K. Peng, M. Simons and H.E. Stanley (1995). "Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis". *Physical Review E*, Vol. **51**(5), 5084-5091.

Carpena, P., P.B. Galván, A.V. Coronado, M. Hackenberg and J.L. Oliver (2007). "Identifying characteristic scales in the human genome". *Physical Review E*, Vol**75**, 032903.

Chatzidimitriou-Dreismann, C.A. and D. Larhammar (1993). "Long-Range Correlations in DNA". *Nature*, Vol. **361**, 212.

Crato, N. and B.K. Ray (2002). "Semiparametric smoothing estimators for long memory processes with added noise". *Jornal of Statistical Planning and Inference*, Vol. **105**, 283-297.

Doukhan, P., G. Oppenheim and M.S. Taqqu (2003). *Theory and Applications of Long-Range Dependence.* Boston: Birkhäuser.

Fox, R. and M.S. Taqqu (1986). "Large-sample Properties of Parameter Estimates for Strongly Dependent Stationary Gaussian Time Series". *The Annals of Statistics*, Vol. **14**, 517-532.

Geweke, J. and S. Porter-Hudak (1983). "The Estimation and Application of Long Memory Time Series Model". *Journal of Time Series Analysis*, Vol. **4**(4), 221-238.

Giraitis, L., P. Kokoszka, R. Leipus and G. Teyssière (2003). "On the power of the R/S-type tests against contiguous and semi long memory alternatives". *Actae Applicandae Mathematicae*, Vol. 78, 285-299.

Hosking, J. (1981). "Fractional Differencing". *Biometrika*, Vol. **68**, 165-167.

Hu, K., P.C. Ivanov, Z. Chen, P. Carpena and H.E. Stanley (2001). "Effect of Trends on Detrended Fluctuation Analysis". *Phyical Review E*, Vol. **64**, 011114.

Hurst, H.R. (1951). "Long-term storage in reservoirs". *Trans. Am. Soc. Civil Eng.*, Vol. **116**, 770-799.

Karlin, S. and V. Brendel (1993). "Patchiness and correlations in DNA sequences". *Science*, Vol. **259**(5095), 677-680.

Koscielny-Bunde, E., H.E. Roman, A. Bunde, S. Havlin and H.-J. Schellnhuber (1998). "Long-range power-law correlations in local daily temperature fluctuations". *Phil. Mag. B*, Vol. **77**, 1331-1340.

Li, W. and K. Kaneko (1992). "Long-Range Correlation and Partial $1/f^{\alpha}$ Spectrum in a Noncoding DNA Sequence". *Europhysics Letters*, Vol. **17**(7), 655-660.

Liu, Y.H., P. Cizeau, M. Meyer, C.-K. Peng and H.E. Stanley (1997). "Correlations in Economic Time Series". *Physica A*, Vol. **245**, 437-440.

Linhares, R.R. (2007). *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.

Lo, A.W. (1991). "Long term memory in stock market prices". *Econometrica*, Vol. **59**, 1279-1313.

Lopes, S.R.C. and B.V.M. Mendes (2006). "Bandwidth Selection in Classical and Robust Estimation of Long Memory". *International Journal of Statistics and Systems*, Vol. **1**(2), 167-190.

Lopes, S.R.C. and M.A. Nunes (2006). "Long Memory Analysis in DNA Sequences". *Physica A: Statistical Mechanics and its Applications*, Vol. **361**(2), 569-588.

Oliver, J.L., P. Carpena, M. Hackenberg and P. Bernaola-Galván (2004). "IsoFinder: computational prediction of isochores in genome sequences". *Nucleic Acids Research*, Vol. **32**, W287-W292.

Peng, C., S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H.E. Stanley (1992). "Long-range Correlations in Nucleotide Sequences". *Nature*, Vol. **356**, 168-170.

Peng, C., S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley and A.L. Goldberger (1994). "Mosaic organization of DNA nucleotides". *Physical Review E*, Vol. **49**(5), 1685-1689.

Percus, J.K. (2002). *Mathematics of Genome Analysis.* Cambridge: Cambridge University Press.

Robinson, P.M. (1995). "Log-Periodogram Regression of Time Series with Long Range Dependence". *Annals of Statistics*, Vol. **23**(3), 1048-1072.

Rousseeuw, P.J. (1984). "Least Median of Squares Regression". *Journal of the American Statistical Association*, Vol. **79**, 871-880.

Stanley, H.E., S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.K. Peng and M. Simons (1999). "Scaling features of noncoding DNA". *Physica A: Statistical Mechanics and its Applications*, Vol. **273**(1), 1-18.

Taqqu, M.S., V. Teverovsky and W. Willinger (1995). "Estimators for Long Range Dependence: An Empirical Study". *Fractals*, Vol. **3**(4), 785-798.

Teverovsky, V., M.S. Taqqu and W. Willinger (1999). "A critical look at Lo's modified R/S statistic". *Journal of Statistical Planning and Inference*, Vol. **80**, 211-227.

Voss. R.F. (1992). "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences". *Physical Review Letters*, Vol. **68**, 3805 - 3808.

Whittle, P. (1953). *Hypothesis Testing in Time Series Analysis.* New York: Hafner.

Yeh, R-G., J-S. Shieh, Y-Y. Han, Y-J. Wang and S-C. Tseng (2006)."Detrended Fluctuation Analyses of Short-Term Heart Rate Variability in Surgical Intensive Care Units". *Biomedical Engineering-Applications, Basis & Communications*, Vol. **18**, 67-72.

Yohai, V.J. (1987). "High breakdown point and high efficiency robust estimates for regression". *Annals of Statistics*, Vol. **15**, 642-656.

Yu, Z-G., V. V. Anh and B. Wang (2000). "Correlation property of length sequences based on global structure of complete genome". *Physical Review E*, Vol. **63**, 011903.