Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Physica A 361 (2006) 569–588

PHYSICA A

www.elsevier.com/locate/physa

Long memory analysis in DNA sequences

S.R.C. Lopes*, M.A. Nunes

Instituto de Matemática, UFRGS, Porto Alegre, Brazil

Received 2 September 2004; received in revised form 6 June 2005
Available online 15 September 2005

Abstract

Our goal in this work is to construct empirical confidence intervals for the fractional parameter d in ARFIMA$(0, d, 0)$ processes. Through these confidence intervals one can compare several estimators for d to decide which one is the best estimation method related to long memory time series. We use a FORTRAN routine that simulates random time series to later perform an analysis for detecting long memory. We also apply the methodology to real DNA sequences to evaluate the efficiency of our method in the construction of these confidence intervals.

Keywords: Long memory; DNA sequences; Empirical confidence intervals

1. Introduction

A time series is a register of values for a certain random variable, measured in different discrete times. For instance, the daily temperature of one city measured at the same hour, during some interval of time, where the previous measure is related to the latter one.

We shall use $\{X_t\}_{t \in T}$ to denote a stochastic process in time $t \in T$, where T is an index set. For each $t \in T$, X_t is a random variable.

Recently, many researchers in time series analysis are studying the ones with long memory characteristics, that is, time series with significant dependence between observations apart for a long period of time. The goal here is to use these characteristics to construct an adequate model for the time series.

*Corresponding author.
E-mail address: slopes@mat.ufrgs.br (S.R.C. Lopes).

0378-4371/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2005.06.099

According to the works [1–3], DNA sequences show long memory, and the goal here is to properly estimate the parameter that describes this characteristic. In order to do this, we consider the ARFIMA (*autoregressive fractionally integrated moving average*) models with (p, d, q) parameters where the *fractional parameter* d measures the long memory property when $d \in (0.0, 0.5)$, and p and q are the orders of the autoregressive and moving average processes, respectively. We shall consider five different estimation methods for d.

This paper is organized as follows: Section 2 describes the stochastic processes with long memory characteristic treating, particularly, the case of ARFIMA(p, d, q) models. In Section 3 we present the chemical structure of a DNA sequence. An explanation of the different estimators for d is given in Section 4. In Section 5, we construct the empirical confidence intervals based on each estimator proposed in the previous section. We analyze a real DNA sequence in Section 6, estimating the value of d through the proposed estimator methods obtaining their confidence intervals. Section 7 concludes this paper.

2. Long memory models

In this section, we present the ARFIMA(p, d, q) model (also called Fractional ARIMA model) and some related theoretical results. Models that includes fractional differentiation d in the interval $(0.0; 0.5)$ are able to represent any time series that shows *persistence*, also known by *long memory property* (see Ref. [4] for a complete study of these models). Initial studies of time series with long memory characteristics were given by Hurst [5]. ARFIMA processes first appeared in Refs. [6,7] and are a generalization of the ARMA and ARIMA models. The author of Ref. [8] was the pioneer in the application of long memory in hydrological time series.

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 different forms:

- in 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 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 $w \to 0$.

One of the models that can describe the persistence is the so-called ARFIMA(p, d, q) processes.

2.1. ARFIMA(p, d, q) process

Definition 1. A stochastic process $\{X_t\}_{t \in \mathbb{Z}}$ is *Gaussian* if, for any set of $t_1, t_2, \ldots, t_n \in \mathbb{Z}$, the random variables $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ have a n-dimensional normal distribution.

We observe that weakly stationary process $\{X_t\}_{t\in\mathbb{Z}}$ does not need to be strongly stationary. However, any weakly stationary Gaussian process will be also strongly stationary (see Ref. [9]).

Definition 2. The process $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is said to be a white noise process with zero mean and variance σ_ε^2, denoted by $\varepsilon_t \sim W\mathrm{N}(0,\sigma_\varepsilon^2)$, if

$$\mathbb{E}(\varepsilon_t) = 0, \quad \mathrm{Var}(\varepsilon_t) = \mathbb{E}(\varepsilon_t^2) = \sigma_\varepsilon^2, \quad \text{and} \quad \gamma_\varepsilon(k) = \begin{cases} \sigma_\varepsilon^2, & k = 0\,, \\ 0, & k \neq 0\,. \end{cases} \tag{1}$$

Definition 3. Let $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ be a white noise process with zero mean and variance $\sigma_\varepsilon^2 > 0$, and $\mathscr{B}$ the backward-shift operator, i.e., $\mathscr{B}^k(X_t) = X_{t-k}$. If $\{X_t\}_{t\in\mathbb{Z}}$ is a linear process satisfying

$$\phi(\mathscr{B})(1-\mathscr{B})^d X_t = \theta(\mathscr{B})\varepsilon_t\,, \quad t \in \mathbb{Z}\,, \tag{2}$$

where $d \in (-0.5; 0.5)$, $\phi(\cdot)$, and $\theta(\cdot)$ are polynomials of degree p and q, respectively, given by

$$\phi(\mathscr{B}) = 1 - \phi_1\mathscr{B} - \cdots - \phi_p\mathscr{B}^p\,,$$

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

where ϕ_i, $1 \leqslant i \leqslant p$, and θ_j, $1 \leqslant j \leqslant q$, are real constants, then $\{X_t\}_{t\in\mathbb{Z}}$ is called general fractional differentiation ARFIMA(p,d,q) process, where d is the degree or fractional differentiation parameter.

The term $(1-\mathscr{B})^d$, for $d \in \mathbb{R}$, is defined through the binomial expansion

$$(1-\mathscr{B})^d = \sum_{k=0}^{\infty} \binom{d}{k} (-\mathscr{B})^k = 1 - d\mathscr{B} - \frac{d}{2!}(1-d)\mathscr{B}^2 \cdots\,.$$

If $d \in (-0.5; 0.5)$, then $\{X_t\}_{t\in\mathbb{Z}}$ is a stationary, and an invertible process (see Theorem 4 below, for the case where $p = 0 = q$).

The most important characteristic of an ARFIMA(p,d,q) process is the property of *long dependence*, when $d \in (0.0; 0.5)$, *short dependence*, when $d = 0$, and *intermediate dependence*, when $d \in (-0.5; 0.0)$. In this work we analyze only processes with long memory property.

2.2. ARFIMA$(0,d,0)$ process

In this work we consider ARFIMA processes where p and q are both equal to zero. The ARFIMA$(0,d,0)$ processes are given by

$$(1-\mathscr{B})^d X_t = \varepsilon_t, \quad \text{for all } t \in \mathbb{Z}\,. \tag{3}$$

Important properties for ARFIMA$(0,d,0)$ processes can be found in Ref. [7]. The following theorem supplies the main properties for these processes.

Theorem 4 (*see Hosking [7]*). *Let* $\{X_t\}_{t\in\mathbb{Z}}$ *be an* $ARFIMA(0,d,0)$ *process.*

(a) *When* $d<0.5$, $\{X_t\}_{t\in\mathbb{Z}}$ *is a stationary process with an infinite moving average representation given by*

$$X_t = \psi(\mathscr{B})\varepsilon_t = \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k} ,$$

where

$$\psi_k = \frac{d(1+d)\cdots(k-1+d)}{k!} = \frac{(k+d-1)!}{k!(d-1)!} .$$

When $k \to \infty, \psi_k \simeq k^{d-1}/(d-1)!$.

(b) *When* $d> -0.5$, $\{X_t\}_{t\in\mathbb{Z}}$ *is an invertible process with an infinite autoregressive representation given by*

$$\pi(\mathscr{B})X_t = \sum_{k=0}^{\infty} \pi_k X_{t-k} = \varepsilon_t ,$$

where

$$\pi_k = \frac{-d(1-d)\cdots(k-1-d)}{k!} = \frac{(k-d-1)!}{k!(-d-1)!} .$$

When $k \to \infty, \pi_k \simeq k^{-d-1}/(-d-1)!$.

In items (c)–(e) *below, we assume that* $d \in (-0.5;0.5)$.

(c) *The spectral density function of* $\{X_t\}_{t\in\mathbb{Z}}$ *is given by*

$$f_X(w) = \left[2\sin\left(\frac{w}{2}\right)\right]^{-2d}, \quad \textit{for } 0<w\leqslant\pi .$$

When $w \simeq 0, f_X(w) \simeq w^{-2d}$.

(d) *The autocovariance function of* $\{X_t\}_{t\in\mathbb{Z}}$ *is given by*

$$\gamma_X(k) = \frac{(-1)^k(-2d)!}{(k-d)!(-k-d)!}$$

and the autocorrelation function is given by

$$\rho_X(k) = \frac{(-d)!(k+d-1)!}{(d-1)!(k-d)!}, \quad \textit{for all } k \in \mathbb{Z} .$$

When $k \to \infty$, $\rho_X(k) \simeq ((-d)!/(d-1)!)k^{2d-1}$.

(e) *The partial autocorrelation function of* $\{X_t\}_{t\in\mathbb{Z}}$ *is given by*

$$\phi_X(k,k) = \frac{d}{k-d}, \quad \textit{for all } k \in \mathbb{N} .$$

Remark 5. For $d>0$ the autocorrelation function $\rho_X(k)$ has hyperbolic decay when k increases, and the spectral density function is unbounded for frequencies near to zero frequency demonstrating the capability of the model to show persistence.

Remark 6. If $\{X_t\}_{t\in\mathbb{Z}}$ is defined by expression (2), then its spectral density function is given by

$$f_X(w) = f_U(w)\left[2\sin\left(\frac{w}{2}\right)\right]^{-2d}, \quad \text{for all } 0<w\leqslant\pi\,, \tag{4}$$

where $f_U(\cdot)$ denotes the spectral density function of an ARMA(p,q) process, U_t, given by

$$(1-\mathscr{B})^d X_t = U_t, \quad \text{for all } t\in\mathbb{Z}\,. \tag{5}$$

3. DNA chemical structure

The DNA is the deoxyribonucleic acid. The DNA ribbons are long polymers made of millions of nucleotides connected some to the others. Individually, nucleotides are quite simple, consisting of three distinct parts: one of the four nitrogenized bases, a deoxyribose (a sugar of 5 carbons), and a phosphate group.

The denomination of the nucleotides depends on the nitrogenized basis that composes them. A DNA sequence is composed by four nucleotides called as *adenine*, *guanine*, *cytosine*, and *thymine*, denoted by the capital letters A, G, C, and T, respectively. (Note: In this work, the words nucleotide and basis will be used to represent the same thing, i.e., a nucleotide.)

Adenine and guanine, a two ring composed molecules, are classified as *purines*. Cytosine and thymine are classified as *pyrimidines* and they are molecules formed by only one ring. One purine connects to one pyrimidine in a DNA sequence to form a pair of bases. Adenine, and thymine are connected to each other to form a pair of A–T bases, while guanine, and cytosine form a pair of G–C bases. The bases remain joined for weakly hydrogen bridges, and these hydrogen bridges are responsible in order to maintain the structure of a double helix of the DNA sequence (see Fig. 1).

3.1. Autocorrelation function in DNA sequences

It is not evident what makes the DNA sequence to present long memory characteristics, but they do so, and this may be related to the evolution's mechanism since the growth of the first form of life on Earth.

When life appeared in our planet, billions of years ago, it appeared from the random combinations in the seas in formation. Passing the time by, for natural processes, these particles had been increasing, and combining, to generate more complex, and adaptable organisms to the environment; this increasing or "elongation" occurred through the so-called *oligonucleotide duplication* or *duplication of the genes* process in which a segment was removed, and some times duplicate, after being reinserted in the original sequence. It is clear that such process was not perfect, and from these small mutations the evolution was made.

Introns are nucleotide sequences that do not "generate" proteins; in contrast with *exons*, that do generate them. Sometimes called by "junk genes", *introns* seem not to

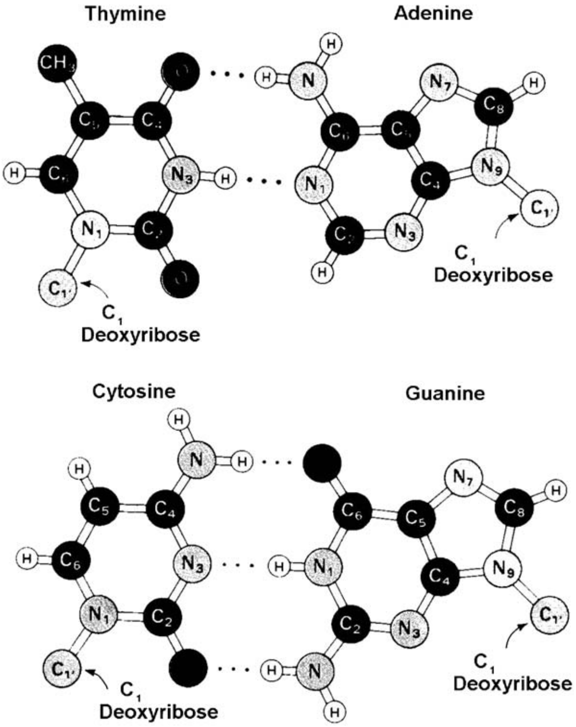

Fig. 1. Illustration showing how the bases pairs are connected by hydrogen bridges.

have any function in the genetic sequence. However, nowadays the biologists have doubts of this (see Ref. [10]), and they do believe that *introns* possess important functions in the mechanism of the evolution. For unknown reasons, *introns* do not suffer many changes as *exons* during the duplication, provoking long memory property more evident through them. Our goal here is to study the parameter of long memory in time series consisting by *introns* and *exons*.

3.2. Random walk

One can consider several different ways to construct a random walk from DNA sequences (see, for instance, Refs. [1,11,12]). Here, in this work, the classification in *purines*, and *pyrimidines* was chosen because its better detection of the long dependence property in DNA sequences (see, for instance, Refs. [1,11]).

In order to study the properties of a DNA sequence we construct a random walk in one dimension, based on this classification. In a DNA sequence, if in the position i one finds a *pyrimidine*, we give one step upward, otherwise, if a *purine* is found we give one step downward (see Ref. [11]). Therefore, we define the function $g(\cdot)$

such that

$$g(i) = \begin{cases} +1 & \text{if } i = \text{ pyrimidine}\,, \\ -1 & \text{if } i = \text{ purine}\,. \end{cases} \tag{6}$$

After t positions, the random walk is the addition of the $g(i)$ steps up to position t, that is,

$$X_t = \sum_{i=1}^{t} g(i)\,.$$

A FORTRAN routine was written to identify the bases, from a standard text archive, determining the steps, and the random walk. We have then a time series $\{X_t\}_{t=1}^{n}$, and we proceed with a long memory analysis based on this data. In general, the time series $\{X_t\}_{t=1}^{n}$ is a sample from a non-stationary stochastic process. In order to obtain a stationary time series we take a first difference of it denoted by

$$Y_t = X_t - X_{t-1} = (1 - \mathscr{B})X_t, \qquad \text{for } t \in \mathbb{N}\,,$$

where $\mathscr{B}$ is the backward-shift operator. Our goal is to study the long memory property of the stochastic process $\{Y_t\}_{t\in\mathbb{N}}$ based on ARFIMA$(0, d, 0)$ processes. We want to estimate properly the parameter d when $d \in (0.0; 0.5)$. We recall that if $\hat{d}_X = 1.2$ is an estimator of d, under the stochastic process $\{X_t\}_{t\in\mathbb{N}}$, then $\hat{d}_Y \approx 0.2$, under the stochastic process $\{Y_t\}_{t\in\mathbb{N}}$ (see Ref. [13]). For the estimation of parameter d when $d \in (0.5, 1.5)$ we refer the reader to Ref. [14].

4. Fractional parameter estimation

We now summarize some methods for the estimation of d: the regression methods using the periodogram function ($\hat{d}_{GPH}$), proposed by Geweke and Porter-Hudak in Ref. [15], and the smoothed version of the periodogram function ($\hat{d}_{SPR}$), proposed by Reisen in Ref. [16]; the estimator proposed by Robinson in Ref. [17] ($\hat{d}_{RP}$) based on the Geweke, and Porter-Hudak's method, where the number of regressors in the regression equation starts from $l > 1$ instead of one and its smoothed version proposed by this work ($\hat{d}_{RSP}$); and the approximated maximum likelihood estimator ($\hat{d}_W$), proposed by Fox, and Taqqu in Ref. [18], based on an idea of [19].

4.1. Estimator $\hat{d}_{GPH}$

Consider the set of Fourier frequencies $w_j = 2\pi j/n, j = 1, \ldots, [n/2]$, where n is the sample size and $[x]$ means the integer part of x. By taking the logarithm of the spectral density function $f_X(\cdot)$, and adding $\ln f_U(0)$, and $\ln I(w_j)$ to both sides of expression (4) we have

$$\ln I(w_j) = \ln f_U(0) - d \ln \left[2 \sin\left(\frac{w_j}{2}\right)\right]^2 + \ln\left[\frac{f_U(w_j)}{f_U(0)}\right] + \ln\left[\frac{I(w_j)}{f_X(w_j)}\right], \tag{7}$$

where $I(\cdot)$ is the periodogram function.

The estimator of d is given by

$$\hat{d}_{GPH} = -\frac{\sum_{j=1}^{g(n)} (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{g(n)} (x_j - \bar{x})^2} , \tag{8}$$

where $g(n) = n^{\alpha}$, $0 < \alpha < 1$ (see Ref. [15]), $y_j = \ln I(w_j)$, $x_j = \ln[2\sin(w_j/2)]^2$, and $\bar{x} = (1/g(n))\sum_{j=1}^{g(n)} x_j$.

4.2. Estimator $\hat{d}_{SPR}$

The regression estimator $\hat{d}_{SPR}$ is obtained by replacing the periodogram function in expression (7) by the smoothed periodogram function, $f_s(\cdot)$, with the Parzen lag window. The parameter m in the lag window generator, usually referred to as the *truncation point*, is a function of the sample size chosen as $m = n^{\beta}$, for $0 < \beta < 1$. The paper [16] shows that $\hat{d}_{SPR}$ is given by the same expression as in (8), where now $y_j = \ln f_s(w_j)$, for $j = 1, \ldots, g(n)$. The value of $g(n)$ is chosen as in the $\hat{d}_{GPH}$ method.

4.3. Estimators $\hat{d}_{RP}$ and $\hat{d}_{RSP}$

We also consider the estimator proposed in Ref. [17], and its smoothed variation proposed by this work. The first one, denoted by $\hat{d}_{RP}$, is a modified version of the estimator $\hat{d}_{GPH}$, where the number of regressors $g(n)$, in expression (7), starts from $l > 1$ instead of one (see Ref. [17]). The second one, denoted by $\hat{d}_{RSP}$, uses the smoothed version of the periodogram function instead of the periodogram itself.

4.4. Estimator $\hat{d}_W$

This estimator involves the function

$$Q(\eta) = \int_{-\pi}^{\pi} \frac{I(w)}{f_X(w;\eta)}\, \mathrm{d}w ,$$

where $f_X(\cdot;\eta)$ is the spectral density function of the $\{X_t\}_{t\in\mathbb{N}}$, and η denotes the vector of unknown parameters. The $\hat{d}_W$ estimator is the value of η which minimizes the function $Q(\cdot)$ (see Ref. [19]). When we are dealing with the situation where $p = 0 = q$, η is given only by the parameter d. For computational purposes, it is easier to minimize the function

$$\mathscr{L}_n(\eta) = \frac{1}{2n}\sum_{j=1}^{n-1}\left\{\ln f_X(w_j;\eta) + \frac{I(w_j)}{f_X(w_j;\eta)}\right\}$$

instead of $Q(\cdot)$, where w_j are the Fourier frequencies, for $j = 1, \ldots, n-1$. For general ARFIMA$(0, d, 0)$ Gaussian processes Fox, and Taqqu have shown in Ref. [18] that the maximum likelihood estimator of d is strongly consistent, and asymptotically normally distributed.

5. Confidence intervals construction

In this section we describe the method used for the construction of the empirical confidence intervals based on the estimators proposed on the previous section. We follow the ideas in the works [20,21].

We do not use the asymptotic theory of all estimators, instead we wrote a FORTRAN routine using the method proposed in Ref. [8] to generate an ARFIMA$(0, d, 0)$ process. The steps of this algorithm are given as follows:

(1) Calculate the partial autocorrelation function $\varphi_X(j,j)$.
(2) Generate a random variable $\mathcal{N}(0, 1)$, through the subroutine RNNOR,[1] of size n, to simulate a Gaussian white noise $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ process.
(3) Calculate the mean and the variance of the random variable X_t, where X_t is obtained from (9) below.
(4) Generate a random variable X_t, with distribution $\mathcal{N}(m_t, v_t)$, for $t \in \{1, 2, \ldots, n\}$, where

$$m_t \equiv \mathbb{E}(X_t|X_\ell, \ell < t) = \sum_{j=1}^{t} \varphi_X(t,j)X_{t-j} ,$$

$$v_t \equiv \mathrm{Var}(X_t|X_\ell, \ell < t) = \sigma_\varepsilon^2 \prod_{j=1}^{t} (1 - \varphi_X^2(j,j)) ,$$

where $\varphi_X(t,j)$ is the partial autocorrelation of an ARFIMA$(0, d, 0)$ process and σ_ε^2 is the variance of the white noise process. For more details, see Ref. [22].

Note 1. For the simulations of any ARFIMA$(0, d, 0)$ process, we always use $\sigma_\varepsilon^2 = 1.0$ in expression (3). We observe that these processes are strongly stationary.

From the generation algorithm, given by (1)–(4), we obtain a sample time series $\{X_t\}_{t=1}^n$ from an ARFIMA$(0, d, 0)$ process given by

$$(1 - \mathcal{B})^d X_t = \varepsilon_t, \text{ for } t \in \{1, 2, \ldots, n\} ,$$

where the sample $\{X_t\}_{t=1}^n$ was obtained from the expression

$$X_t = (1 - \mathcal{B})^{-d}\varepsilon_t, \quad \text{for } t \in \{1, 2, \ldots, n\} . \tag{9}$$

The next step, after obtaining a time series, is to deal with the estimation of the fractional parameter d. In this work we use the estimators proposed in Section 4, namely $\hat{d}_{GPH}$, $\hat{d}_{SPR}$, $\hat{d}_{RP}$, $\hat{d}_{RSP}$, and $\hat{d}_W$. For this we consider 1000 time series. For each series we estimate the value of d through the different methods and later

[1] This subroutine belongs to the IMSL FORTRAN library and generates pseudorandom numbers from a standard normal distribution.

we take the arithmetic average of these values, that is,

$$\overline{d}_i = \frac{1}{1000}\sum_{j=1}^{1000} \widehat{d}_i(j)\,,$$

where $\overline{d}_i$ corresponds to $\hat{d}_{GPH}$, $\hat{d}_{SPR}$, $\hat{d}_{RP}$, $\hat{d}_{RSP}$, and $\hat{d}_W$, respectively, depending on the estimation method used. To compare the different estimators we considered the mean squared error value, denoted hereafter by MSE, i.e.,

$$\text{MSE} = \frac{1}{1000}\sum_{j=1}^{1000} (\widehat{d}_i(j) - d)^2\,,$$

where d is the true parameter value.

In Tables 1–4 we present the simulation results for the fractional parameter $d \in \{0.05; 0.10; 0.15; 0.45\}$ in ARFIMA$(0, d, 0)$ processes, for all estimation methods proposed here.

We now construct empirical confidence intervals for the fractional parameter based on the estimation procedures given in Section 4. The process to construct the empirical confidence intervals consists of the following steps:

(1) For each sample size n (we use $n \in \{256, 512, \ldots, 8192\}$) we generate 1000 replications.
(2) We calculate the lower bound (0.5%, 2.5% and 5.0%) and upper bound (99.5%, 97.5% and 95%) limits of the obtained estimator values. They are denoted by lower bound and upper bound, respectively.

Table 1
The mean value and MSE, using different estimators, for different sample sizes n, based on ARFIMA$(0, d, 0)$ with $(d = 0.05)$

n	Estimation method $(d = 0.05)$				
	$\hat{d}_{GPH}$	$\hat{d}_{SPR}$	$\hat{d}_{RP}$	$\hat{d}_{RSP}$	$\hat{d}_W$
Mean value					
256	0.0550	0.0091	0.0495	0.0460	0.0353
512	0.0496	0.0176	0.0468	0.0456	0.0421
1024	0.0530	0.0215	0.0544	0.0403	0.0446
2048	0.0531	0.0340	0.0532	0.0493	0.0481
4096	0.0515	0.0380	0.0529	0.0488	0.0481
8192	0.0482	0.0369	0.0471	0.0441	0.0493
MSE					
256	0.0436	0.0273	0.0689	0.0384	0.0032
512	0.0285	0.0182	0.0459	0.0233	0.0014
1024	0.0167	0.0114	0.0237	0.0143	0.0007
2048	0.0112	0.0074	0.0161	0.0089	0.0003
4096	0.0078	0.0051	0.0098	0.0058	0.0002
8192	0.0053	0.0034	0.0066	0.0037	0.0001

Table 2
The mean value and MSE, using different estimators, for different sample sizes n, based on ARFIMA$(0, d, 0)$ with $d = 0.10$

n	Estimation method ($d = 0.10$)				
	$\hat{d}_{GPH}$	$\hat{d}_{SPR}$	$\hat{d}_{RP}$	$\hat{d}_{RSP}$	$\hat{d}_{W}$
Mean value					
256	0.1105	0.0656	0.1023	0.0931	0.0843
512	0.1013	0.0684	0.1012	0.0967	0.0911
1024	0.1091	0.0743	0.0941	0.0964	0.0951
2048	0.1069	0.0790	0.0966	0.0972	0.0969
4096	0.1035	0.0812	0.0970	0.0983	0.0986
8192	0.1071	0.0844	0.1027	0.0951	0.0993
MSE					
256	0.0434	0.0293	0.0711	0.0381	0.0033
512	0.0299	0.0209	0.0426	0.0238	0.0015
1024	0.0195	0.0127	0.0244	0.0135	0.0007
2048	0.0111	0.0082	0.0169	0.0091	0.0003
4096	0.0081	0.0052	0.0115	0.0067	0.0002
8192	0.0047	0.0037	0.0061	0.0039	0.0001

Table 3
The mean value and MSE, using different estimators, for different sample sizes n, based on ARFIMA$(0, d, 0)$ with $d = 0.15$

n	Estimation method ($d = 0.15$)				
	$\hat{d}_{GPH}$	$\hat{d}_{SPR}$	$\hat{d}_{RP}$	$\hat{d}_{RSP}$	$\hat{d}_{W}$
Mean value					
256	0.1502	0.0983	0.1502	0.1408	0.1327
512	0.1404	0.1041	0.1428	0.1389	0.1418
1024	0.1486	0.1167	0.1509	0.1407	0.1441
2048	0.1495	0.1315	0.1493	0.1497	0.1475
4096	0.1528	0.1347	0.1526	0.1480	0.1487
8192	0.1527	0.1383	0.1534	0.1484	0.1494
MSE					
256	0.0436	0.0289	0.0751	0.0385	0.0031
512	0.0295	0.0201	0.0475	0.0236	0.0015
1024	0.0173	0.0120	0.0245	0.0133	0.0006
2048	0.0121	0.0081	0.0164	0.0092	0.0003
4096	0.0078	0.0054	0.0103	0.0059	0.0002
8192	0.0054	0.0036	0.0065	0.0038	0.0001

(3) We construct a graph where the sample sizes are in the abscissa and the obtained values from step 2 are in the ordinate axis. The values are fitted by a linear regression method using a MATLAB routine. From the fitted functions we construct the confidence intervals.

Table 4
The mean value and MSE, using different estimators, for different sample sizes n, based on ARFIMA$(0, d, 0)$ with $d = 0.45$

n	Estimation method ($d = 0.45$)				
	$\hat{d}_{GPH}$	$\hat{d}_{SPR}$	$\hat{d}_{RP}$	$\hat{d}_{RSP}$	$\hat{d}_{W}$
Mean value					
256	0.4705	0.4124	0.4687	0.4825	0.4391
512	0.4601	0.4149	0.4545	0.4638	0.4436
1024	0.4696	0.4344	0.4690	0.4730	0.4491
2048	0.4569	0.4349	0.4552	0.4628	0.4497
4096	0.4635	0.4452	0.4609	0.4664	0.4499
8192	0.4583	0.4452	0.4583	0.4617	0.4501
MSE					
256	0.0434	0.0284	0.0737	0.0366	0.0033
512	0.0302	0.0211	0.0473	0.0242	0.0014
1024	0.0190	0.0127	0.0264	0.0147	0.0007
2048	0.0131	0.0090	0.0174	0.0096	0.0003
4096	0.0086	0.0057	0.0106	0.0063	0.0002
8192	0.0057	0.0039	0.0070	0.0041	0.0001

For instance, to get a confidence interval for parameter d at 95% confidence level, we plot the values 2.5% and 97.5% of confidence interval versus the time series sample size n. For these data, the best adjusted function is

$$h(n) = a[\log_2(\log_2(n))^2] + b[\log_2(\log_2(n))] + c$$

with coefficients a, b, and c estimated by linear regression method using a MATLAB routine. Figs. 2–5 present the confidence interval for different values of $d \in \{0.05; 0.10; 0.15; 0.45\}$ only at 95% confidence level, based on all estimation procedures considered here, with $n \in \{256, 512, \ldots, 8192\}$. For other confidence levels, and other different values of d, the results are available upon request.

One observes in Figs. 2–5 that the obtained values for the estimation of d converge to the true parameter value as the sample size increases. This result was expected, since as long as the sample size increases, more precise will be the estimates for d, independently of the estimation procedure.

With the data used to construct the graph we can approximate functions that return the estimated value of d for different sample sizes that we choose for the simulation. Table 5 supplies the confidence intervals for d based on time series with sample size n belonging to $\{256, 512, \ldots, 8192\}$. With this procedure, we obtain the confidence interval for d based on each estimation method. The upper and lower bounds of d are calculated when we change N to $\log_2(\log_2(n))$ in the fitted equations of Tables 5–8.

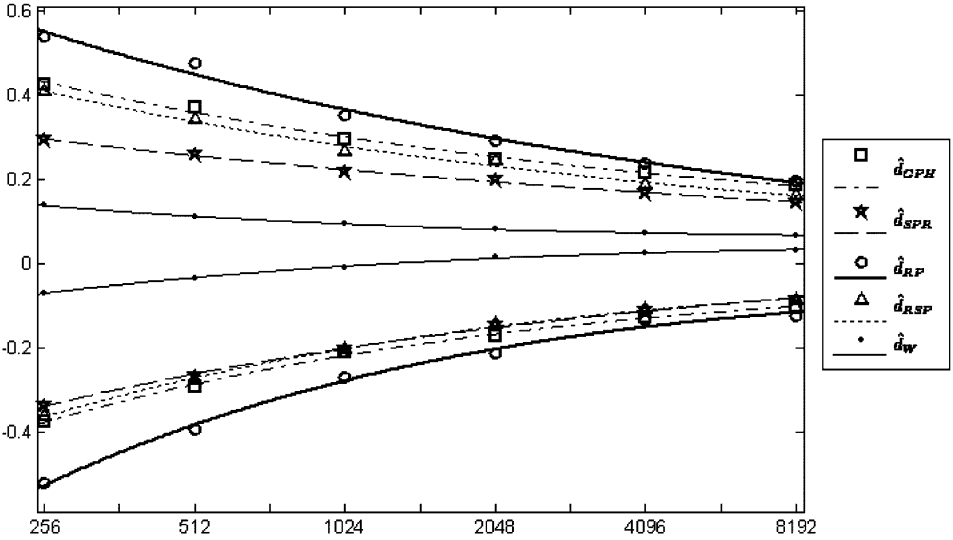

Fig. 2. Confidence interval for $d = 0.05$ at 95% based on the five considered estimation methods and on six different sample sizes.

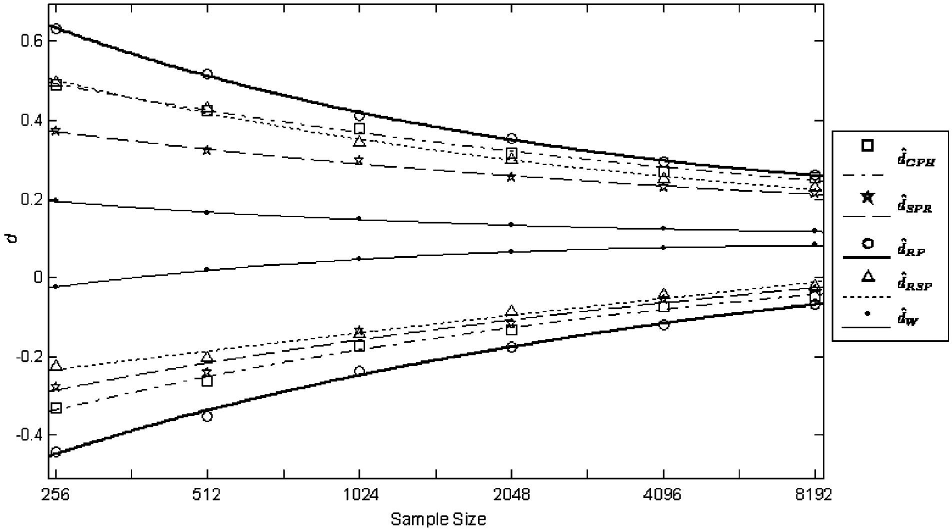

Fig. 3. Confidence interval for $d = 0.10$ at 95% based on the five considered estimation methods and on six different sample sizes.

Tables 1–4 show the simulation results for $d \in \{0.05; 0.10; 0.15; 0.45\}$. One observes that, for the case where $d = 0.10$, the best result is attained by $\hat{d}_W$.

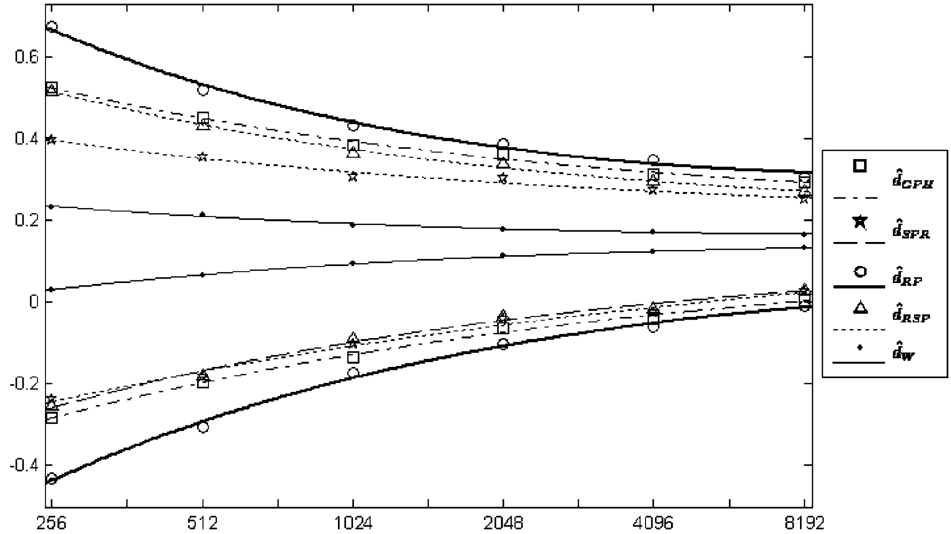

Fig. 4. Confidence interval for $d = 0.15$ at 95% based on the five considered estimation methods and on six different sample sizes.

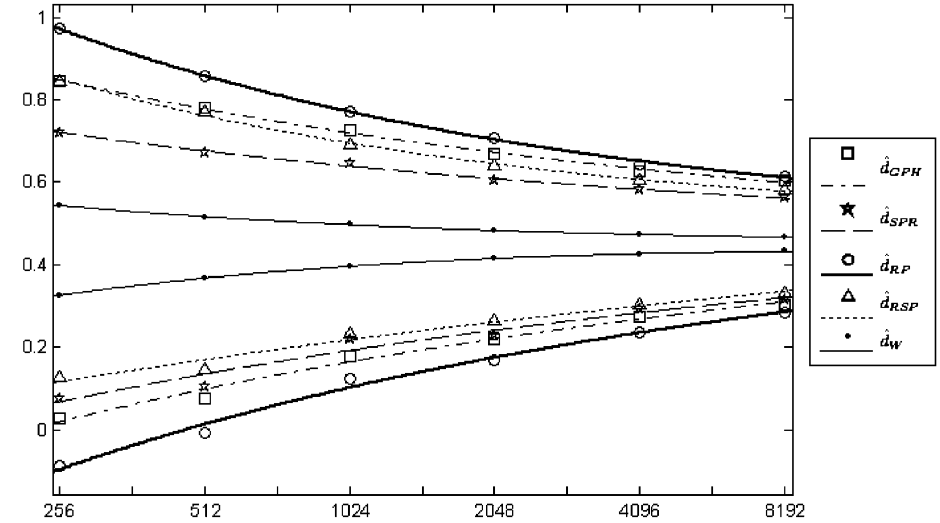

Fig. 5. Confidence interval for $d = 0.45$ at 95% based on the five considered estimation methods and on six different sample sizes.

6. Application

To test the effectiveness of the described procedure in Section 5, we analyze a real DNA sequence, calculating the confidence interval for all estimators proposed in

Table 5
Confidence intervals for $d = 0.05$ at 95%, where $N = \log_2(\log_2(n))$, with different estimators

Estimator	Interval	Fitted equation
$\hat{d}_{GPH}$	Upper bound	$h = 0.1429N^2 - 1.311N + 3.078$
	Lower bound	$h = -0.2635N^2 + 2.160N - 4.485$
$\hat{d}_{SPR}$	Upper bound	$h = 0.0335N^2 - 0.438N + 1.308$
	Lower bound	$h = -0.1589N^2 + 1.433N - 3.210$
$\hat{d}_{RP}$	Upper bound	$h = 0.1704N^2 - 1.654N + 3.978$
	Lower bound	$h = -0.4893N^2 + 3.864N - 7.715$
$\hat{d}_{RSP}$	Upper bound	$h = 0.1372N^2 - 1.273N + 2.993$
	Lower bound	$h = -0.2678N^2 + 2.198N - 4.547$
$\hat{d}_{W}$	Upper bound	$h = 0.0931N^2 - 0.724N + 1.470$
	Lower bound	$h = -0.1334N^2 + 1.042N - 1.997$

Table 6
Confidence intervals for $d = 0.10$ at 95%, where $N = \log_2(\log_2(n))$, with different estimators

Estimator	Interval	Fitted equation
$\hat{d}_{GPH}$	Upper bound	$h = 0.0881N^2 - 0.941N + 2.521$
	Lower bound	$h = -0.1491N^2 + 1.419N - 3.250$
$\hat{d}_{SPR}$	Upper bound	$h = 0.0711N^2 - 0.704N + 1.843$
	Lower bound	$h = -0.0656N^2 + 0.811N - 2.129$
$\hat{d}_{RP}$	Upper bound	$h = 0.3447N^2 - 2.845N + 6.068$
	Lower bound	$h = -0.2062N^2 + 1.923N - 4.362$
$\hat{d}_{RSP}$	Upper bound	$h = 0.1819N^2 - 1.614N + 3.707$
	Lower bound	$h = 0.0704N^2 - 0.154N - 0.4058$
$\hat{d}_{W}$	Upper bound	$h = 0.0921N^2 - 0.731N + 1.553$
	Lower bound	$h = -0.1334N^2 + 1.027N - 1.063$

Section 4. We use available sequences in the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov). Chosen a sequence, we use routines developed in this work to analyze it.

The first routine constructs a random walk (described in Section 3) for this sequence. From this random walk we use another routine developed in FORTRAN computational language (see Refs. [23,24]) to estimate d based on the estimation procedures proposed in Section 4.

In this section, the methodology is applied to the homo sapiens dystrophin sequence (muscular dystrophy, Duchenne and Becker types) (DMD, transcript

Table 7
Confidence intervals for $d = 0.15$ at 95%, where $N = \log_2(\log_2(n))$, with different estimators

Estimator	Interval	Fitted equation
$\hat{d}_{GPH}$	Upper bound	$h = 0.2032N^2 - 1.691N + 3.768$
	Lower bound	$h = -0.1934N^2 + 1.706N - 3.665$
$\hat{d}_{SPR}$	Upper bound	$h = 0.1088N^2 - 0.929N + 2.204$
	Lower bound	$h = -0.1170N^2 + 1.168N - 2.699$
$\hat{d}_{RP}$	Upper bound	$h = 0.5444N^2 - 4.148N + 8.210$
	Lower bound	$h = -0.4586N^2 + 3.679N - 7.348$
$\hat{d}_{RSP}$	Upper bound	$h = 0.2434N^2 - 1.977N + 4.256$
	Lower bound	$h = -0.2390N^2 + 2.012N - 4.147$
$\hat{d}_{W}$	Upper bound	$h = 0.0985N^2 - 0.758N + 1.623$
	Lower bound	$h = -0.1305N^2 + 1.021N - 1.860$

Table 8
Confidence intervals for $d = 0.45$ at 95%, where $N = \log_2(\log_2(n))$, with different estimators

Estimator	Interval	Fitted equation
$\hat{d}_{GPH}$	Upper bound	$h = 0.1145N^2 - 1.123N + 3.187$
	Lower bound	$h = -0.0933N^2 + 1.040N - 2.262$
$\hat{d}_{SPR}$	Upper bound	$h = 0.0711N^2 - 0.704N + 2.193$
	Lower bound	$h = -0.0716N^2 + 0.842N - 1.813$
$\hat{d}_{RP}$	Upper bound	$h = 0.2917N^2 - 2.470N + 5.756$
	Lower bound	$h = -0.1900N^2 + 1.818N - 3.840$
$\hat{d}_{RSP}$	Upper bound	$h = 0.2435N^2 - 2.018N + 4.713$
	Lower bound	$h = -0.0109N^2 + 0.388N - 0.949$
$\hat{d}_{W}$	Upper bound	$h = 0.0909N^2 - 0.719N + 1.882$
	Lower bound	$h = -0.1779N^2 + 1.343N - 2.102$

variant Dp140bc, mRNA from NCBI #NM 004023). For other applications of DNA sequences we refer the reader to Ref. [25].

The sequence of this example presents 7048 nucleotides. Fig. 6 shows the plot of the random walk for this DNA sequence. Table 9 shows the analysis result for all estimators proposed in Section 4 for this DNA sequence.

Remark 7. The right choice of the number of regressors $g(n) = n^{\alpha}$ in expression (8) gave raise too many works among researchers and practitioners for the

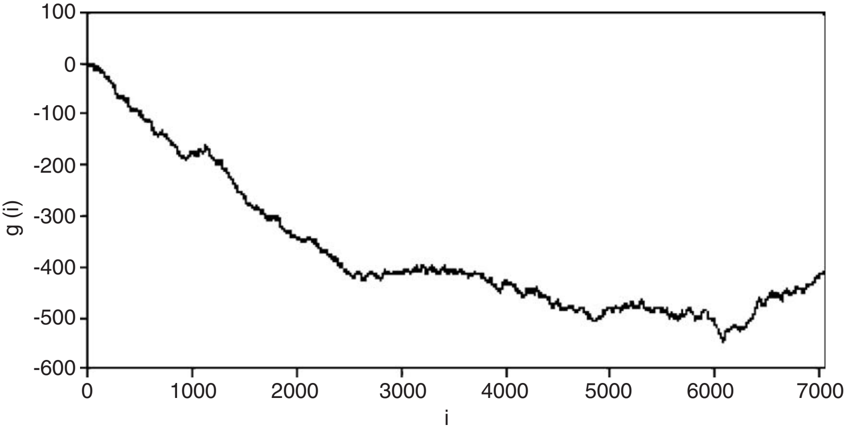

Fig. 6. Random walk for the homo sapiens dystrophin sequence.

Table 9
Estimation results for d using different estimation methods

Estimation method	$\hat{d}_{GPH}$	$\hat{d}_{SPR}$	$\hat{d}_{RP}$	$\hat{d}_{RSP}$	$\hat{d}_{W}$
$\hat{d}_i$ value	0.0922	0.0915	0.0902	0.0928	0.0930

Table 10
Results for the upper and lower bounds for the estimated value of d using different estimation methods

Estimation method	$\hat{d}_{GPH}$	$\hat{d}_{SPR}$	$\hat{d}_{RP}$	$\hat{d}_{RSP}$	$\hat{d}_{W}$
Upper bound	0.2523	0.2158	0.2677	0.2319	0.1104
Lower bound	−0.0485	−0.0342	−0.0794	−0.0205	0.9097

semiparametric estimation of d. For theoretical purpose, $g(n)$ is a function of n such that $g(n)/n \to 0$, as $n \to \infty$. We refer the reader to [4,9,17] for more details.

To calculate the confidence intervals for the estimators we use Table 6. In this table one finds the fitted equations that allow to evaluate the upper and lower bounds of the estimators for the considered cases. The results are shown in Table 10 below.

According to Table 10 one observes that the estimated values for d are in between the limits of the calculated fitted equations, at 95% confidence level (see the upper and lower bounds in Table 10).

Table 11
Results for the estimation of d using different values of α

α	Estimation method					Absolute error			
	$\hat{d}_{GPH}$	$\hat{d}_{SPR}$	$\hat{d}_{RP}$	$\hat{d}_{RSP}$	$\hat{d}_{W}$	$\hat{d}_{GPH}$	$\hat{d}_{SPR}$	$\hat{d}_{RP}$	$\hat{d}_{RSP}$
0.50	0.1513	0.1286	0.1137	0.0995	0.0930	0.0584	0.0356	0.0207	0.0065
0.51	0.1485	0.1151	0.1136	0.0865	0.0930	0.0555	0.0221	0.0207	0.0065
0.52	0.1231	0.0887	0.0878	0.0590	0.0930	0.0302	0.0042	0.0052	0.0340
0.53	0.1296	0.0910	0.0979	0.0639	0.0930	0.0366	0.0020	0.0049	0.0291
0.54	0.1355	0.0915	0.1071	0.0666	0.0930	0.0425	**0.0015**	0.0141	0.0264
0.55	0.1672	0.1212	0.1448	0.1019	0.0930	0.0743	0.0283	0.0519	0.0089
0.56	0.1632	0.1187	0.1421	0.1006	0.0930	0.0702	0.0258	0.0491	0.0076
0.57	0.1551	0.1132	0.1350	0.0960	0.0930	0.0622	0.0202	0.0420	0.0030
0.58	0.1686	0.1217	0.1513	0.1067	0.0930	0.0756	0.0288	0.0583	0.0137
0.59	0.1477	0.1134	0.1299	0.0988	0.0930	0.0548	0.0205	0.0369	0.0059
0.60	0.1475	0.1068	0.1311	0.0928	0.0930	0.0546	0.0138	0.0381	**0.0002**
0.61	0.1673	0.1102	0.1538	0.0975	0.0930	0.0744	0.0173	0.0608	0.0046
0.62	0.1498	0.0968	0.1359	0.0841	0.0930	0.0568	0.0039	0.0430	0.0089
0.63	0.1358	0.0854	0.1219	0.0728	0.0930	0.0428	0.0076	0.0289	0.0202
0.64	0.1282	0.0881	0.1149	0.0766	0.0930	0.0352	0.0049	0.0219	0.0164
0.65	0.1157	0.0759	0.1026	0.0646	0.0930	0.0227	0.0170	0.0096	0.0284
0.66	0.1093	0.0784	0.0967	0.0680	0.0930	0.0163	0.0146	0.0038	0.0250
0.67	0.0897	0.0672	0.0769	0.0570	0.0930	0.0033	0.0258	0.0160	0.0360
0.68	0.0946	0.0719	0.0831	0.0627	0.0930	0.0017	0.0211	0.0099	0.0303
0.69	0.0825	0.0618	0.0712	0.0528	0.0930	0.0105	0.0312	0.0217	0.0402
0.70	0.0821	0.0631	0.0716	0.0548	0.0930	0.0109	0.0299	0.0213	0.0382
0.71	0.0747	0.0562	0.0647	0.0482	0.0930	0.0182	0.0368	0.0282	0.0448
0.72	0.0762	0.0563	0.0670	0.0489	0.0930	0.0168	0.0367	0.0260	0.0441
0.73	0.0761	0.0565	0.0676	0.0497	0.0930	0.0169	0.0365	0.0254	0.0433
0.74	0.0473	0.0328	0.0383	0.0255	0.0930	0.0456	0.0602	0.0546	0.0675
0.75	0.0444	0.0289	0.0359	0.0220	0.0930	0.0486	0.0641	0.0571	0.0709
0.76	0.0413	0.0309	0.0334	0.0246	0.0930	0.0516	0.0620	0.0596	0.0683
0.77	0.0483	0.0393	0.0411	0.0337	0.0930	0.0447	0.0537	0.0518	0.0592
0.78	0.0442	0.0364	0.0374	0.0312	0.0930	0.0488	0.0565	0.0555	0.0617
0.79	0.0446	0.0389	0.0384	0.0342	0.0930	0.0483	0.0540	0.0546	0.0588
0.80	0.0533	0.0513	0.0478	0.0473	0.0930	0.0396	0.0416	0.0452	0.0457
0.81	0.0515	0.0505	0.0463	0.0467	0.0930	0.0414	0.0425	0.0466	0.0463
0.82	0.0556	0.0555	0.0509	0.0521	0.0930	0.0373	0.0375	0.0421	0.0409
0.83	0.0673	0.0638	0.0632	0.0608	0.0930	0.0256	0.0292	0.0298	0.0322
0.84	0.0638	0.0640	0.0599	0.0612	0.0930	0.0291	0.0289	0.0330	0.0317
0.85	0.0661	0.0652	0.0625	0.0627	0.0930	0.0269	0.0277	0.0305	0.0303
0.86	0.0681	0.0651	0.0648	0.0627	0.0930	0.0248	0.0279	0.0282	0.0303
0.87	0.0745	0.0691	0.0716	0.0669	0.0930	0.0184	0.0238	0.0214	0.0260
0.88	0.0809	0.0759	0.0782	0.0740	0.0930	0.0121	0.0170	0.0148	0.0189
0.89	0.0816	0.0758	0.0791	0.0740	0.0930	0.0113	0.0172	0.0139	0.0190
0.90	0.0834	0.0764	0.0811	0.0747	0.0930	0.0095	0.0166	0.0119	0.0183
0.91	0.0884	0.0835	0.0863	0.0820	0.0930	0.0045	0.0095	0.0067	0.0110
0.92	0.0922	0.0882	0.0902	0.0869	0.0930	**0.0008**	0.0048	**0.0027**	0.0061
0.93	0.0918	0.0880	0.0898	0.0866	0.0930	0.0012	0.0050	0.0031	0.0063
0.94	0.0955	0.0841	0.0883	0.0856	0.0930	0.0025	0.0089	0.0047	0.0074
0.95	0.0978	0.0879	0.0869	0.0876	0.0930	0.0048	0.0051	0.0061	0.0054

7. Conclusion

In this work we analyzed five different estimators for the long memory parameter d. From all these analyzed estimators, we can observe that $\hat{d}_W$ is the estimator that better behaves. Tables 1, 3, and 4 and Figs. 2, 4, and 5 show that $\hat{d}_W$ has lesser variation among the maximum, and minimum values. Also, $\hat{d}_W$ has the smallest mean squared error, as we can see in Tables 1, 3, and 4. The best estimation procedure, in the statistical sense, is the maximum likelihood method, hereafter denoted by $\hat{d}_W$. It is always asymptotically unbiased and normally distributed (see Refs. [4,9,18,26]). The semiparametric methods ($\hat{d}_{GPH}, \hat{d}_{SPR}, \hat{d}_{RP}$, and $\hat{d}_{RSP}$) are easier to be implemented and even more flexible, but if one has the information that the data comes from an ARFIMA model, then the right method to be used is the maximum likelihood procedure. This is in accordance with the simulation results obtained in the tables (Table 11).

A question can be asked from the results when we use real data. Which estimator is the best choice when analyzing DNA sequences? As we noted in the previous paragraph, $\hat{d}_W$ has lesser variation among the maximum, and minimum values and the smallest mean squared error value. So, if we have to choose a method to estimate the parameter d in a time series, obtained from an ARFIMA$(0, d, 0)$ process, we must choose the $\hat{d}_W$ method.

One can also note that $\hat{d} \in (0.0, 0.5)$ for all methods. With this result, we conjecture that DNA sequences have long range dependence. However, the $\hat{d}_W$ estimator has slow convergence. When one is dealing with large size of observations in a time series (for instance, DNA sequences with more than 4000 base pairs), the $\hat{d}_W$ estimator takes some minutes to converge, while the other estimators ($\hat{d}_{GPH}$, $\hat{d}_{SPR}$, $\hat{d}_W$, $\hat{d}_{RP}$, and $\hat{d}_{RSP}$) converge more quickly. We believe that this fact is due to the FORTRAN routines used in this work, since they were not optimized to a statistical use.

Acknowledgements

SRC Lopes was partially supported by CNPq-Brazil, by Pronex *Probabilidade e Processos Estocásticos* (Convênio MCT/CNPq/FAPERJ - Edital 2003) and also by *Fundação de Amparo à Pesquisa no Estado do Rio Grande do Sul* (FAPERGS). The authors would like to thank to the editor and one anonymous referee for the valuable comments and suggestions that improved this version of the manuscript.

References

[1] W. Li, K. Kaneko, Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence, Europhys. Lett. 17 (7) (1992) 655–660.

[2] M. Osaka, K. Gohara, S. Ishii, H. Kishida, H. Hayakawa, N. Ito, Symbolic strings and spatial $1/f$ spectra, Physica D 125 (1–2) (1999) 142–154.

[3] B. Borstnik, D. Pumpernik, D. Lukman, Analysis of apparent $1/f^{\alpha}$ spectrum in DNA sequences, Europhys. Lett. 23 (6) (1993) 389–394.

[4] J. Beran, Statistics for Long Memory Processes, Chapman & Hall, New York, 1994.

[5] H.E. Hurst, Long-term storage capacity of reservoirs, Trans. Amer. Soc. Civil Eng. 116 (1951) 770–799.

[6] C.W.J. Granger, R. Joyeux, An introduction to long-memory time series models and fractional differencing, J. Time Ser. Anal. 1 (1980) 15–30.

[7] J.R.M. Hosking, Fractional differencing, Biometrika 68 (1) (1981) 165–176.

[8] J.R.M. Hosking, Modelling persistence in hydrological time series using fractional differencing, Water Resour. Res. 20 (12) (1984) 1898–1908.

[9] P.J. Brockwell, R.A. Davis, Time Series: Theory and Methods, Springer, New York, 1991.

[10] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.K. Peng, M. Simons, Scaling features of noncoding DNA, Physica A 273 (1) (1999) 1–18.

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

[12] Z.G. Yu, V.V. Anh, B. Wang, Correlation property of length sequences based on global structure of the complete genome, Phys. Rev. E 63 (2000) (011903).

[13] B.P. Olbermann, S.R.C. Lopes, V.A. Reisen, Invariance of the first difference in ARFIMA models, Comput. Stat. (2005), accepted for publication.

[14] S.R.C. Lopes, B.P. Olbermann, V.A. Reisen, A comparison of estimation methods in non-stationary arfima processes, J. Stat. Comput. Simulation 74 (5) (2004) 339–347.

[15] J. Geweke, S. Porter-Hudak, The estimation and application of long-memory time series models, J. Time Ser. Anal. 4 (4) (1983) 221–238.

[16] V.A. Reisen, Estimation of the fractional difference parameter in the $ARIMA(p,d,q)$ model using the smoothed periodogram, J. Time Ser. Anal. 15 (3) (1994) 335–350.

[17] P.M. Robinson, Log-periodogram regression of time series with long range dependence, Ann. Stat. 23 (3) (1995) 1040–1072.

[18] R. Fox, M. Taqqu, Large sample properties of parameter estimates for strongly dependent stationary time series, Ann. Stat. 14 (1986) 517–532.

[19] P. Whittle, Estimation and information in stationary time series, Arkiv Math. 2 (1953) 423–434.

[20] M.S. Taqqu, V. Teverovsky, W. Willinger, Estimators for long-range dependence: an empirical study, Fractals 3 (4) (1995) 785–798.

[21] R. Weron, Estimating long-range dependence: finite sample properties and confidence intervals, Physica A 312 (1–2) (2002) 285–299.

[22] C. Bisognin, S.R.C. Lopes, Estimating the long memory parameter in the presence of seasonality, submitted for publication.

[23] V.A. Reisen, S.R.C. Lopes, Some simulations and applications of forecasting long memory time series models, J. Stat. Planning Inference 80 (2) (1999) 269–287.

[24] V.A. Reisen, B. Abraham, S.R.C. Lopes, Estimation of parameters in ARFIMA processes: a simulation study, Commun. Stat. Simulation Comput. 30 (4) (2001) 787–803.

[25] M.A. Nunes, S.R.C. Lopes, Análise de Longa Dependência em Sequências de DNA, Technical Report, Serie A, No. 60, Instituto de Matemática, UFRGS, Porto Alegre, 2004.

[26] F. Sowell, Maximum likelihood estimation of stationary univariate fractionally integrated time series models, J. Econometrics 53 (1992) 165–188.