α-STABLE LAWS FOR NONCODING REGIONS IN DNA SEQUENCES

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

September 10, 2009

Abstract

In this work, we analyze the *long-range dependence* parameter for a nucleotide sequence in several different transformations. The long-range dependence parameter is estimated by the approximated maximum likelihood method, by a novel estimator based on the spectral envelope theory, by a regression method based on the periodogram function, and also by the *detrended fluctuation analysis* method. We study the length distribution of coding and noncoding regions for all *Homo sapiens* chromosomes available from the European Bioinformatics Institute (EBI). The parameter of the tail rate decay is estimated by the Hill estimator $\hat{\alpha}$. We show that the tail rate decay is greater than 2 for coding regions, while for almost all noncoding regions it is less than 2.

Mathematics Subject Classification (2000). Primary 60G10, 62G05, 62G35, 62M10, 62M15; Secondary 62M20.

Keywords. Long Memory Models, α-Stable Law, Generalized Pareto Distribution, Hill Estimator.

1 Introduction

The statistical properties of DNA genomes are of great interest since they reflect biological features that are important for life (see Percus, 2002). One of the main recent findings is the existence of correlation on the sequence, i.e., occurrence of a nucleotide in a specific position depends statistically on the previous nucleotides (memory). Moreover, it has been found that the memory of genome structure decomposition is of a particular type: the so-called *long memory*.

The search for intrinsic patterns, correlations, and parameters measuring self-similarity by scaling exponents has been carried out in past years by several statistical methods. Peng et al. (1992), Buldyrev et al. (1995), Li an Kaneko (1992) and Stanley et al. (1999) use the

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

detrended fluctuation analysis to characterize long-range correlations in both coding and noncoding regions of DNA sequences. Karlin and Brendel (1993) show the relationship between the effect of patchiness and correlations in genome sequences. Bernardi et al. (1985) show that the DNA nucleotides form a mosaic of long homogeneous segments or "isochores".

In order to study the long-range correlations in genomes it is necessary to transform them into numerical sequences. There are many studies on correlations in DNA sequences using different representations for the four nucleotides (see Stanley et al., 1999; Guharay et al., 2000; Cristea, 2003; Garcia and José, 2005; Lopes and Nunes, 2006 and Podobnik et al., 2007).

One of the most appropriated methods proposed in recent years for the study of long-range correlations in genomes is the detrended fluctuation analysis (DFA) (see Peng et al., 1994; Buldyrev et al., 1995; Podobnik et al., 2007 and Garcia et al., 2008). More recently, Crato et al. (2009) estimate the fractional parameter d by considering semiparametric regression method based on the periodogram function, in both classical and robust versions. They also use the semiparametric R/S(n) method, proposed by Hurst (1951), and the maximum likelihood method (see Fox and Taqqu, 1986), with the approximation suggested by Whittle (1953).

Here we study the behavior of the *long-range dependence* parameter of one DNA sequence for the most relevant transformations of the type $f : (\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}) \rightarrow \{0, 1, 2, 3\}$, where A, C, G and T are the four nucleotides, by considering the DFA and other estimation methods for the *fractional parameter d*.

A common problem in analyzing long nucleotide sequence data is the proper identification of coding (*cds*) regions dispersed throughout the sequence and separated by noncoding (*ncds*) regions (see Bergen and Antoniou, 2005). Here we are interested in analyzing the distribution of coding and noncoding regions, based on the length distribution of these two regions (see Stuart et al., 2006). From the tail rate decay estimator, we identify differences between length distribution of coding and noncoding regions of DNA sequences.

This paper is organized as follows. Section 2 describes the DFA and other estimation methods for the *fractional parameter d*. Section 3 presents the most relevant transformations for the four nucleotides. We study the long-range dependence for a DNA sequence numerically transformed into several different maps. Section 4 presents the relationship between long-range dependence and the tail index parameter and also the Hill estimator $\hat{\alpha}$. Section 5 presents the discussion of the length distribution for coding and noncoding regions for all *Homo sapiens* chromosomes available from the European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/). Section 6 concludes the paper.

2 Estimation Methods

In this section we want to emphasize the existence of long memory in a DNA sequence, without separately considering its coding and noncoding regions (see Crato et al., 2008). With this purpose we consider the estimation of the fractional parameter d in four different methods. First, the approximated maximum likelihood estimator ($\hat{d}_W$), proposed by Fox and Taqqu (1986); the $\hat{d}_{WS}$ estimator, obtained by using the spectral envelope theory proposed by Stoffer et al. (1993); the regression method using the periodogram function ($\hat{d}_{GPH}$), a very well-known estimator proposed by Geweke and Porter-Hudak (1983)

and, finally, the *detrended fluctuation analysis* estimator ($\hat{d}_{DFA}$), proposed by Peng et al. (1992).

The parameter of interest is the *fractional integration parameter* d, which can be introduced in the general framework of the autoregressive fractionally integrated moving average processes (ARFIMA).

An ARFIMA(p, d, q) process is defined as follows. Let $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ be a white noise process with zero mean and variance $\sigma_\epsilon^2 > 0$, let $\mathcal{B}$ be the backward-shift operator, that is, $\mathcal{B}^k(X_t) = X_{t-k}$, and let $\Phi(\cdot)$ and $\Theta(\cdot)$ be polynomials of orders p and q, respectively. If

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

where μ is the mean of the process and $(1-\mathcal{B})^d = \sum_{k=0}^{\infty}\binom{d}{k}(-\mathcal{B})^k$, then $\{X_t\}_{t\in\mathbb{Z}}$ is an ARFIMA(p, d, q).

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, $U_t \equiv (1-\mathcal{B})^d(X_t - \mu)$.

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 $w \to 0$.

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 Beran (1994).

2.1 Estimator $\hat{d}_W$

Let $\{X_t\}_{t\in\mathbb{Z}}$ be an ARFIMA(p, d, q), defined in expression (2.1). The estimator for d, by using the maximum likelihood method, denoted by $\hat{d}_W$, is the value

$$\boldsymbol{\eta} = (\sigma_X^2, d, \phi_1, \phi_2, \cdots, \phi_p, \theta_1, \theta_2, \cdots, \theta_q) \tag{2.3}$$

that minimizes the function

$$\mathcal{Q}(\boldsymbol{\eta}) = \sum_{j=1}^{[\frac{n-1}{2}]}\left(\frac{I(w_j)}{f_X(w_j, \boldsymbol{\eta})}\right), \tag{2.4}$$

where $\boldsymbol{\eta}$ is the vector of unknown parameters given in (2.3), $f_X(\cdot, \boldsymbol{\eta})$ is the spectral density function of the $\{X_t\}_{t\in\mathbb{Z}}$, $[\cdot]$ is the greatest integer function, $w_j = \frac{2\pi j}{n}$ is the j-th Fourier frequency, $j \in \{1, \cdots, [\frac{n-1}{2}]\}$, and $I(\cdot)$ is the periodogram function.

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

2.2 Estimator $\hat{d}_{WS}$

Let $\{X_t\}_{t\in\mathbb{Z}}$ be a categorical-valued process with finite state-space $\mathcal{C} = \{c_1, c_2, \cdots, c_k\}$ and $p_j = \mathbb{P}(X_t = c_j) > 0$, for $j = 1, 2, \cdots, k$. For $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_k) \in \mathbb{R}^k$ denote by $\{X_t(\boldsymbol{\beta})\}_{t\in\mathbb{Z}}$ the real-valued process corresponding to the scaling that assigns the category c_j to the numerical value β_j, $j = 1, 2, \cdots, k$. Assume $\{X_t(\boldsymbol{\beta})\}_{t\in\mathbb{Z}}$ is a stationary process. Its spectral density will be denoted by $f_X(\omega; \boldsymbol{\beta})$. Our goal is to find scalings $\boldsymbol{\beta}$ in such way that one can maximize the power at each frequency ω, across all frequencies $\omega \in (0, \pi]$, relative to the total power $\sigma^2(\boldsymbol{\beta}) = \mathrm{Var}(X_t(\boldsymbol{\beta}))$. That is, we choose $\boldsymbol{\beta}(\omega)$, at each ω of interest, so that

$$\lambda(\omega) = \max_{\boldsymbol{\beta}} \left\{ \frac{f_X(\omega; \boldsymbol{\beta})}{\sigma^2(\boldsymbol{\beta})} \right\}, \tag{2.5}$$

over all $\boldsymbol{\beta}$ not proportional to $\mathbf{1}_k$, the $k \times 1$ vector of ones. The function $\lambda(\cdot)$ is defined to be the *spectral envelope* of the stationary categorical process $\{X_t\}_{t\in\mathbb{Z}}$.

The name *spectral envelope* is very much appropriate since $\lambda(\cdot)$ envelopes the standardized spectrum of any scaled process. That is, given any $\boldsymbol{\beta}$ normalized so that $\{X_t(\boldsymbol{\beta})\}_{t\in\mathbb{Z}}$ has total power equal to one, $f_X(\omega; \boldsymbol{\beta}) \leq \lambda(\omega)$, where equality holds, if and only if, $\boldsymbol{\beta}$ is proportional to $\boldsymbol{\beta}(\omega)$. For more details, we refer the reader to Stoffer et al. (1993).

Here we suggest an estimator for the fractional parameter d, denoted by $\hat{d}_{WS}$, that consists in replacing the periodogram function in expression (2.4) by the spectral envelope, given in expression (2.5).

2.3 Estimator $\hat{d}_{GPH}$

Let $\{X_t\}_{t\in\mathbb{Z}}$ be an ARFIMA(p, d, q), defined in expression (2.1). The first estimation method based on the periodogram function was proposed by Geweke and Porter-Hudak (1983). These authors obtain an estimate for d by considering $\ln(I(\omega_j))$ regressed on $\ln\left(2\sin\left(\frac{w_j}{2}\right)\right)^2$. The estimator $\hat{d}_{GPH}$ is given by

$$\hat{d}_{GPH} = -\frac{\sum_{j=1}^{g(n)} (x_j - \overline{x})(y_j - \overline{y})}{\sum_{j=1}^{g(n)} (x_j - \overline{x})^2}, \tag{2.6}$$

where $y_j = \ln(I(w_j))$, $x_j = \ln\left(2\sin\left(\frac{w_j}{2}\right)\right)^2$ and $\overline{x} = \frac{1}{g(n)} \sum_{j=1}^{g(n)} x_j$, with $g(n) = n^\alpha$, for $\alpha \in (0, 1)$.

2.4 Estimator $\hat{d}_{DFA}$

Given a time series $\{X_t\}_{t=1}^n$, the *Detrended Fluctuation Analysis* (DFA) method, proposed by Peng et al. (1994), 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 $\ln(l)$.

The estimator $\hat{d}_{DFA}$ is given by

$$\hat{d}_{DFA} = \frac{\sum_{j=1}^{m}(x_j - \overline{x})y_j}{\sum_{j=1}^{m}(x_j - \overline{x})^2} - \frac{1}{2}, \tag{2.7}$$

where $y_j = \ln(\mathrm{F}(j+3))$, $x_j = \ln(j+3)$, $\overline{x} = \frac{1}{m}\sum_{j=1}^{m} x_j$ and $m = [g(n) - 3]$. The function $F(l)$, for each block of size l, is the *root mean square fluctuation* given by

$$\mathrm{F}^2(l) = \frac{1}{\widetilde{n}}\sum_{t=1}^{\widetilde{n}} Z_t^2, \tag{2.8}$$

with $Z_t = Y_t - Y_t^l$, where $Y_t = \sum_{j=1}^{t} X_j$, for each $t \in \{1, 2, \cdots, n\}$, Y_t^l is the adjusted fit on each block and $\widetilde{n} = [\mathrm{M} \cdot l] \leq n$, with $M = [n/l]$.

For technical details on the estimator $\hat{d}_{DFA}$ we refer the reader to Crato et al. (2009).

3 Transformations

A nucleotide sequence is composed by the basis A (adenine), C (cytosine), T (thymine) and G (guanine). In order to apply numerical methods to a nucleotide sequence it is necessary to transform it into a numerical sequence.

From the biological point of view, it is common to apply the SW (strong-weak pairing) rule, which maps C and G to 1 and A and T to 0. We could also try scaling according to the purine-pyrimidine alphabet (RY rule), that is A = G = 1 and C = T = −1. This rule describes how purines (A and G) and pyrimidines (C and T) are distributed along the sequence. Another interesting scaling (Real rule) is the one that assigns A = −1.5, C = 0.5, G = −0.5 and T = 1.5 (see Chakravarthy et al., 2004).

Guharay et al. (2000) symmetrically assign a four dimensional vector to each of the four bases A, C, G and T, which are equidistant from each other, and their sum is equal to zero, that is, maps A to $(0.75, -0.25, -0.25, -0.25)$, C to $(-0.25, 0.75, -0.25, -0.25)$, G to $(-0.25, -0.25, 0.75, -0.25)$ and T to $(-0.25, -0.25, -0.25, 0.75)$.

Cristea (2003) proposes a tetrahedral representation of the nucleotides: four vectors in three dimensions, symmetrically placed with respect to each other, i.e., oriented towards the corners of a tetrahedron, are placed in correspondence with the nucleotides, in which it is emphasized that the vertices of a regular tetrahedron are a subset of the vertices of a cube. Thus, this rule maps A to $(1, 1, 1)$, C to $(-1, 1, -1)$, G to $(-1, -1, -1)$ and T to $(1, -1, -1)$.

Table 3.1: Estimators for the Parameter d, with their Respective 95% Confidence Levels for the AL163202 Nucleotide Sequence, using 30 Different Transformations.

(A,C,G,T)	$\hat{d}_{\mathrm{W}}$	$\hat{d}_{\mathrm{GPH}}$	$\hat{d}_{\mathrm{DFA}}$	(A,C,G,T)	$\hat{d}_{\mathrm{W}}$	$\hat{d}_{\mathrm{GPH}}$	$\hat{d}_{\mathrm{DFA}}$
(0,0,0,1)	0.0766^{*}	0.4016^{*}	0.1073^{*}	(0,1,1,1)	0.0749^{*}	0.3015^{*}	0.1133^{*}
(0,0,1,0)	0.0593^{*}	0.2660^{*}	0.1033^{*}	(0,1,1,2)	0.0923^{*}	0.3665^{*}	0.0872^{*}
(0,1,0,0)	0.0597^{*}	0.2581^{*}	0.0955^{*}	(0,1,1,3)	0.0901^{*}	0.3664^{*}	0.0892^{*}
(0,0,1,1)	0.0804^{*}	0.3836^{*}	0.0649^{*}	(0,0,2,3)	0.0823^{*}	0.3980^{*}	0.0744^{*}
(0,1,0,1)	0.0965^{*}	0.3060^{*}	0.0817^{*}	(0,2,0,3)	0.0947^{*}	0.3453^{*}	0.0884^{*}
(0,1,1,0)	0.0382^{*}	0.3137^{*}	0.1748^{*}	(0,0,3,1)	0.0652^{*}	0.2613^{*}	0.0776^{*}
(0,0,1,2)	0.0820^{*}	0.4016^{*}	0.0820^{*}	(0,0,3,2)	0.0750^{*}	0.3377^{*}	0.0633^{*}
(0,1,0,2)	0.0915^{*}	0.3606^{*}	0.0931^{*}	(0,1,2,1)	0.0559^{*}	0.2988^{*}	0.1247^{*}
(0,1,2,0)	0.0422^{*}	0.3081^{*}	0.1602^{*}	(0,1,2,2)	0.0809^{*}	0.3561^{*}	0.0870^{*}
(0,0,1,3)	0.0808^{*}	0.3976^{*}	0.0907^{*}	(0,1,2,3)	0.0880^{*}	0.3875^{*}	0.0812^{*}
(0,1,0,3)	0.0871^{*}	0.3840^{*}	0.0982^{*}	(0,1,3,1)	0.0526^{*}	0.2843^{*}	0.1211^{*}
(0,1,3,0)	0.0464^{*}	0.2931^{*}	0.1446^{*}	(0,1,3,2)	0.0696^{*}	0.3180^{*}	0.0908^{*}
(0,0,2,1)	0.0704^{*}	0.2863^{*}	0.0680^{*}	(0,1,3,3)	0.0816^{*}	0.3772^{*}	0.0779^{*}
(0,2,0,1)	0.0855^{*}	0.2489^{*}	0.0790^{*}	(0,2,1,2)	0.0907^{*}	0.3066^{*}	0.0951^{*}
(0,2,1,0)	0.0425^{*}	0.3024^{*}	0.1534^{*}	(0,2,1,3)	0.0954^{*}	0.3407^{*}	0.0888^{*}

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

Rather than choose values arbitrarily, the spectral envelope approach (see Stoffer et al., 1993) selects scales that help to emphasize any periodic feature that may exist in a categorical time series. This approach can be used for any categorical time series in a quick and automated fashion.

If we consider the naive approach of arbitrarily assigning numerical values (scales) to the categories and then proceeding with a spectral analysis, the result will depend on the particular assignment of the numerical values. Here we consider the estimation of the fractional parameter d, based on the four methods given in Section 2, relatively to all possible transformations $f : (\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}) \to \{0, 1, 2, 3\}$. To analyze the long-range dependence we consider the AL163202 nucleotide sequence, corresponding to a part of the *Homo sapiens* chromosome 21 (with 340,000 bp).

We have a total of 4^4 different maps. Some of these 4^4 transformations are equivalent for the estimations results. For instance, the transformations $(1, 0, 0, 0)$, $(2, 0, 0, 0)$, $(3, 0, 0, 0)$, and $(4, 0, 0, 0)$ lead necessarily to the same correlation and spectral estimates. In this regard, they constitute a class of equivalent transformations. In Table 3.1 we chose one mapping representative for each of such classes. This is what we call the set of the relevant transformations. Although each map brings out different property of the sequence, note that all of them display long-range dependence. For each sequence, we test the hypothesis $H_0 : d = 0$ versus $H_1 : d \neq 0$, based on the long-range dependence parameter, at 95% confidence level, for all estimators proposed.

The value for the spectral envelope estimator is $\hat{d}_{WS} = 0.0722$, calculated for all transformations together, applied to the AL163202 nucleotide sequence. Since it is a single value, we do not report it in Table 3.1. From Table 3.1 one observes that the $\hat{d}_{GPH}$ estimator always give higher estimates for the parameter d among all four methods. The estimators $\hat{d}_W$, $\hat{d}_{WS}$ and also $\hat{d}_{DFA}$ are in the same order of magnitude. What is important is that all estimators are significantly pointing into the direction of long memory.

4 Long-Range Dependence and the Tail Index Estimator

The most important parameter to estimate on heavy-tailed data is the tail rate of decay α which determines the probability of occurrence of extreme values of the underlying distribution. That is, α is the parameter such that

$$\mathbb{P}(|X| > x) \sim C\, x^{-\alpha}, \quad \text{for} \quad x \in \mathbb{R},$$

where C is a positive constant. Our interest here lies on the probabilistic modeling and on the inference statistics for the extreme part of the tail of the distribution of X_t for coding (*cds*), and Y_t for noncoding (*ndcs*) regions, of some DNA sequences (see Definition 5.1) by estimating the parameter α.

Let $X_1, X_2, \cdots$ be i.i.d. random variables representing risks or losses with an unknown cumulative distribution function (CDF) $F(x) = \mathbb{P}(X_i \leq x)$. A loss is treated as a positive number and extreme events occur when losses take values in the right tail of the loss distribution $F(\cdot)$. Define $M_n = \max(X_1, \cdots, X_n)$ as the worst-case loss in a sample of n losses. An important part of the extreme value theory focuses on the distribution of M_n. From the i.i.d. assumption, the CDF of M_n is given by

$$\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x, \cdots, X_n \leq x) = \prod_{i=1}^{n} F(x) = F^n(x).$$

Since $F^n(\cdot)$ is assumed to be unknown and the empirical distribution function is often a very poor estimator for $F^n(\cdot)$, an asymptotic approximation to $F^n(\cdot)$ based on the *Fisher-Tippett Theorem* (Fisher and Tippett, 1928) is used to make inferences on M_n. Furthermore, since $F^n(x) \to 0$ or 1, as $n \to \infty$ and x is fixed, the asymptotic approximation is based on the standardized maximum value

$$Z_n = \frac{M_n - \mu_n}{\sigma_n}, \tag{4.1}$$

where σ_n and μ_n are sequences of real numbers such that $\sigma_n > 0$ is interpreted as a scale measure and μ_n as a location measure.

A natural measure of extreme events are values of the X_i that exceed a high threshold u. Define the *excess distribution* above the threshold u as the conditional probability

$$F_u(y) = \mathbb{P}(X - u \leq y | X > u) = \frac{F(y+u) - F(u)}{1 - F(u)}, \; y > 0. \tag{4.2}$$

It can be shown (see Embrechts et al., 1997) that for large enough u there exists a positive function $\beta(u)$ such that the excess distribution (4.2) is well approximated by the *generalized Pareto distribution* (GPD)

$$G_{\xi,\beta(u)}(y) = \begin{cases} 1 - (1 + \xi y/\beta(u)), & \text{if } \xi \neq 0 \\ 1 - \exp(-y/\beta(u)), & \text{if } \xi = 0, \end{cases} \tag{4.3}$$

defined for $y > 0$, when $\xi \geq 0$, and for $0 \leq y \leq -\beta(u)/\xi$, when $\xi < 0$, where $\beta(u) > 0$.

For more details see Embrechts et al. (1997) and Samorodnitsky and Taqqu (1994).

4.1 Long-Range Dependence

We can define *long-range dependence* in the *symmetric α-stable ($s\alpha s$)* case (see Embrechts et al., 1997), but certainly not via the covariance function since the second moments for the marginal distributions do not exist. An example of $s\alpha s$ H-selfsimilar process is given by the integral

$$X_t^{(H)} = \int_{\mathbb{R}} \left[\left((t-x)^+\right)^{H-1/\alpha} - \left((-x)^+\right)^{H-1/\alpha} \right] dM(x),\ t \in \mathbb{R}, \tag{4.4}$$

where H $\in (0,1)$ is the *Hurst's effect* parameter for estimating the long-range dependence. In the above expression, M is an $s\alpha s$ random measure with Lebesgue control measure. For $H = 1/\alpha$ the process is formally interpreted as an $s\alpha s$ motion. The $s\alpha s$ process so defined in expression (4.4) is called *linear fractional stable motion*. It is a process with stationary increments, but the subclass of selfsimilar processes does not consist only of these fractional motions. The corresponding *fractional $s\alpha s$ noise* can be defined as

$$Y_t^{(H)} = X_{t+1}^{(H)} - X_t^{(H)},\ t \in \mathbb{Z}. \tag{4.5}$$

When $H = 1/\alpha$, in expression (4.5), the process is formally interpreted as a *symmetric α-stable ($s\alpha s$) noise*, whereas for $H \in (\frac{1}{\alpha}, 1)$ and $1 < \alpha < 2$, it is considered as a process with *long-range dependence*, in analogy to the fractional Brownian noise.

Remark 4.1. The Hurst parameter H, in expression (4.4), is related to the fractional parameter d (see Beran, 1994) by the equation

$$d = H - \frac{1}{2}. \tag{4.6}$$

Since $H = \frac{1}{\alpha}$, from expression (4.6) we obtain

$$d = \frac{1}{\alpha} - \frac{1}{2} \Longleftrightarrow \alpha = \frac{2}{2\,d+1}, \tag{4.7}$$

that is, one has the relationship between the fractional parameter d and the tail index α.

4.2 Hill Estimator for the Tail Index

For the *generalized Pareto distribution* (GPD), the shape α may be estimated non-parametrically in quite different ways. A popular method due to Hill (1975) applies to the case where $\alpha > 0$ so that the data is generated by some fat-tailed distribution in the domain of attraction of Fréchet family or GPD distribution. To describe the Hill estimator, let us consider a sample of losses $X_1, \cdots, X_n$ and define the order statistics as

$$X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}.$$

For a positive integer k, the *Hill estimator* for α is defined as

$$\hat{\alpha}(k) = \left(k^{-1} \sum_{j=1}^{k} \ln(X_{(j)}) - \ln(X_{(k)}) \right)^{-1}. \tag{4.8}$$

The Hill estimators for α depend on the integer k. It can be shown that if $F(\cdot)$ is in the domain of attraction of a GPD distribution, then $\hat{\alpha}(k)$ is asymptotically normally distributed with asymptotic variance given by $avar(\hat{\alpha}(k)) = k^{-1}\,\alpha^2$.

5 Length Distribution Analysis in DNA Sequences

In this section we analyze the coding and noncoding length distribution by considering the Hill estimator. This analysis will help us to identify differences between these distributions and the existence of long memory for noncoding regions of DNA sequences.

A common problem in analyzing long nucleotide sequence data is how to identify coding regions (*cds*) dispersed throughout the sequence and separated by noncoding regions (*ncds*). Distinguishing the distribution for each one of these regions can help to locate both coding and noncoding regions in DNA sequences.

DEFINITION 5.1. Let $\{n_i\}_{i=1}^n$ be any nucleotide sequence. By ordering all lengths of coding (or noncoding) segments, according to their order in the complete genome, the obtained integer sequence is called a *coding (or noncoding) length sequence*.

REMARK 5.1. From Definition 5.1, one obtain the time series $\{X_t\}_{t=1}^n$ (analogously, the time series $\{Y_t\}_{t=1}^n$) derived from the *coding (or noncoding) length sequence* given by

$$\begin{aligned} X_t &= \text{the length of the } \textit{coding segment} \text{ at position } t \\ Y_t &= \text{the length of the } \textit{noncoding segment} \text{ at position } t, \end{aligned}$$

where the *length of a coding (or noncoding) segment* means the number of its bases in the segment.

Figure 5.1 shows the histograms of the *coding* and *noncoding length sequences*, respectively, for the *Homo sapiens* chromosome 21. The observed long tail in Figure 5.1 illustrates how difficult it can be to attribute probabilities to extreme events using all data set. One can note there is a difference between *coding* and *noncoding length sequence* histograms.

Our interest is to analyze the distribution of coding and noncoding regions, by considering, respectively, the *coding* and *noncoding length sequences* (see Definition 5.1). Table 5.1 presents, respectively, the Hill estimator $\hat{\alpha}_{cds}$ for the *coding* and $\hat{\alpha}_{ncds}$ for the *noncoding length sequences* for all *Homo sapiens* chromosomes. This set of data is available from the European Bioinformatics Institute (EBI, http://www.ebi.ac.uk/). Since each set of sequences begins and ends with a coding segment, we always have one more segment for coding (column 3) than for noncoding (column 5).

The α parameter evaluates the tail rate of decay, which determines the probability of occurrence of extreme events and the existence of moments for the underlying distribution. Here k is given by $n^{0.5}$, where n is the sample size.

From Table 5.1, one observes that coding regions present α estimates usually larger than those from noncoding regions. This happens both on average and within the same chromosome. Only in two cases (chromosomes 15 and 18) it does not happen that the first estimate is larger than the corresponding second one. Although this needs further investigation, one may suspect that the α parameters are generally above the $\alpha = 2$ mark for coding segments and below this mark for noncoding segments.

6 Conclusions

In this paper, we analyze the performance of the estimator $\hat{d}_{WS}$, obtained by replacing the periodogram in expression (2.4) by the spectral envelope. Considering several different

Table 5.1: Hill Estimator $\hat{\alpha}_{cds}$ ($\hat{\alpha}_{ncds}$) for Coding (Noncoding) Length Sequences for *Homo sapiens* Chromosomes, with $k = n^{0.5}$.

Homo sapiens Chr.	Sequence	Coding's Size	$\hat{\alpha}_{cds}$	Noncoding's Size	$\hat{\alpha}_{ncds}$
1	CM000252	45119	2.6216	45118	1.8538
2	CM000253	35685	2.0099	35684	1.8062
3	CM000254	27943	2.1347	27942	1.9630
4	CM000255	16325	1.9904	16324	1.8526
5	CM000256	19582	2.2984	19581	2.0941
6	CM000257	23586	2.3579	23585	1.6658
7	CM000258	22096	2.4524	22095	1.9517
8	CM000259	14854	2.3956	14853	2.0456
9	CM000260	18837	1.9965	18836	1.7615
10	CM000261	20235	2.2170	20234	1.7593
11	CM000262	25522	2.2307	25521	1.8608
12	CM000263	25970	3.0919	25969	2.2118
13	CM000264	9149	1.7409	9148	1.7010
14	CM000265	15114	2.2924	15113	1.4330
15	CM000266	15032	1.9473	15031	2.5036
16	CM000267	19098	1.9741	19097	1.7487
17	CM000268	26740	2.2214	26739	1.8042
18	CM000269	7317	2.2275	7316	2.6834
19	CM000270	23421	2.6367	23420	2.3779
20	CM000271	13203	2.4240	13202	1.6034
21	CM000272	5243	2.1805	5242	1.6884
22	CM000273	5243	2.1805	5242	1.6884
X	CM000274	15604	2.3963	15603	1.5784
Y	CM000275	361	2.1131	360	2.0556

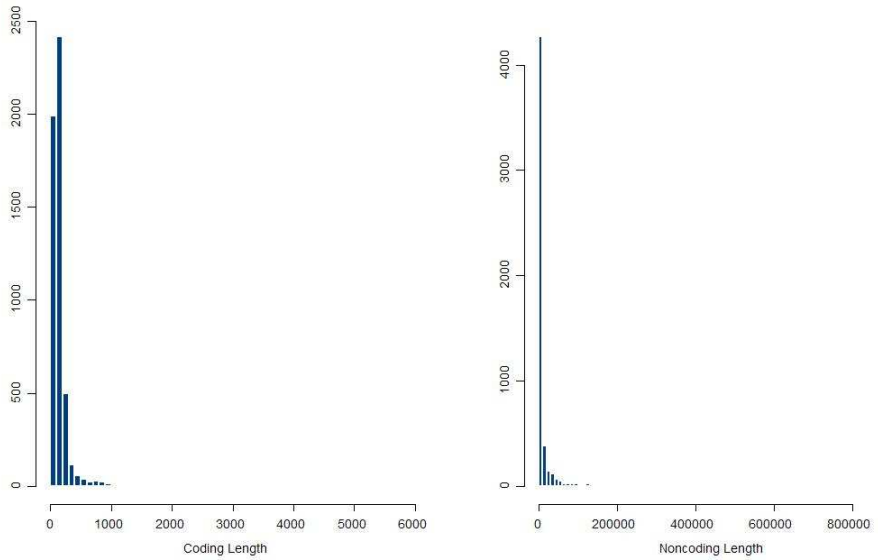

Figure 5.1: Histogram for *Coding* and *Noncoding Length Sequences* for the *Homo sapiens* Chromosome 21.

transformations, we compare the $\hat{d}_{GPH}$ estimator with $\hat{d}_W$, $\hat{d}_{WS}$ and $\hat{d}_{DFA}$, where the first one is the well-known estimator based on the linear regression equation, the second one is

the approximated maximum likelihood method, and the last one is the estimator obtained from the DFA method. For the AL163202 nucleotide sequence applied to these different transformations, the estimators for the fractional parameter d considered in Section 2 are statistically different from zero. This indicates the existence of long-range dependence for the considered sequence when the four bases are numerically assigned by these maps. This type of dependence cannot be induced by the scaling; it is a property of the sequence.

We also performed an analysis of the tail rate decay of coding and noncoding regions' distribution, for all considered *Homo sapiens* chromosomes. We applied the $\hat{\alpha}$ estimator to both *coding* and *noncoding length sequences*. Although this needs further investigation, results seem to indicate that for most *noncoding regions* the length distributions tend to have finite first and second moments ($\alpha \geq 2$), while for *coding regions* the length distributions tend to have no finite second moments, being in the domain of attraction of an α-stable law, with $\alpha < 2$.

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 thank the editor and two anonymous referees for valuable comments that greatly improved the paper.

References

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.

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.

Chakravarthy, N., A. Spanias, L.D. Iasemidis and K. Tsakalis (2004). "Autoregressive Modeling and Feature Analysis of Sequences". *EURASIP Journal on Applied Signal Processing*, Vol. **2004**(1), 13-28.

Crato, N., R.R. Linhares and S.R.C. Lopes (2009). "Statistical Properties of Detrended Fluctuation Analysis". Accepted for publication in *Journal of Statistical Computation and Simulation.*

Cristea, P.D. (2003). "Large scale features in DNA genomic signals". *Journal on Applied Signal Processing*, Vol. **83**, 871-888.

Embrechts, P., C. Klüppelberg and T. Mikosch (1997). *Modelling extremal events for insurance and finance.* Berlin: Springer-Verlag.

Fisher, R.A. and L.H. Tippett (1928). "Limiting forms of the frequency distribution of the largest or smallest member of a sample". *Proc. Cambridge Philos. Soc.* **24**, 180-190.

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.

Garcia, J.A.L. , F. Bartumeusa, D. Rocheb, J. Giraldob, H.E. Stanley and E.O. Casamayora (2008). "Ecophysiological significance of scale-dependent patterns in prokaryotic genomes unveiled by a combination of statistic and genometric analyses". *Genomics*, Vol **91**, 538-543.

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.

Guharay, S., B.R. Hunt, J.A. Yorke and O.R. White (2000). "Correlations in DNA Sequences Across the Three Domains of Life". *Physica D: Nonlinear Phenomena*, Vol. **146**(1-4), 388-396.

Hill, B.M. (1975). "A Simple General Approach to Inference about the Tail of a Distribution". *The Annals of Statistics*, Vol. **3**, 1163-1174.

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

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.

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.

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.

Podobnik, B., J. Shaoc, N.V. Dokholyand, V. Zlatice, H.E. Stanley and I. Grossef (2007). "Similarity and dissimilarity in correlations of genomic DNA". *Physica A*, Vol. **373**, 497-502.

Samorodnitsky, G. and M.S. Taqqu (1994). *Stable Non Gaussian Random Processes.* London: Chapman & Hall.

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.

Stoffer, D.S., D.E. Tyler and A.J. McDougall (1993). "Spectral analysis for categorical time series: Scaling and the spectral envelope". *Biometrika*, Vol. **80**, 611-622.

Stuart, W., A. Bergen and A. Antoniou (2006). "A Stochastic Model for DNA Sequences Using Prescribed Nucleotide and Length Distributions". *International Symposium on Signal Processing and Information Technology*, 95-100.

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