UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE MATEMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA

Propriedades Estatísticas do Método da Análise de Flutuações Destendenciadas em Seqüências de DNA

Dissertação de Mestrado

Raquel Romes Linhares

Porto Alegre, 13 de Setembro de 2007.

Dissertação submetida por Raquel Romes Linhares[1] como requisito parcial para a obtenção do grau de Mestre em Matemática pelo Programa de Pós-Graduação em Matemática do Instituto de Matemática da Universidade Federal do Rio Grande do Sul.

Professora Orientadora:
Dr[a]. Sílvia Regina Costa Lopes

Banca Examinadora:
Dr. Alexandre Tavares Baraviera
Dr. Hildete Prisco Pinheiro
Dr. Rafael Rigão Souza

Data da Defesa: 13 de Setembro de 2007.

[1]Conselho Nacional de Desenvolvimento Científico e Tecnológico

Agradecimentos

Gostaria de expressar meu agradecimento primeiramente à DEUS por me guiar sempre.

À minha mãe Aparecida Marcia Silva Linhares, ao meu pai Eidmar Oliveira Linhares e ao meu irmão Erick Oliveira Linhares pelo incentivo e compreensão desde o início até a conclusão desta jornada.

À minha orientadora Dr^{a}. Sílvia Regina Costa Lopes pelo apoio na definição do trabalho, pela confiança em mim depositada, pelo incentivo e, principalmente, pela orientação.

Ao meu pai, ao Ricardo Barbosa e ao Cristiano Arend Lima pelo apoio na Souza Cruz.

Aos colegas do Programa de Pós-Graduação em Matemática da UFRGS, especialmente ao meu colega Cleber Bisognin pelas contribuições e comentários.

Aos professores do Programa de Pós-Graduação em Matemática da UFRGS pelo comprometimento na busca da qualidade de trabalho, conceito e conhecimento.

Aos bolsista do LCPM, pelo auxílio e amizade.

À banca examinadora pela atenção.

À Rosane, secretária do Programa de Pós-Graduação, pela atenção e paciência.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico pelo auxílio financeiro.

À Universidade Federal do Rio Grande do Sul pela oportunidade.

E a todos que torceram pela realização deste trabalho.

Resumo

Conforme diversos artigos, as seqüências de DNA apresentam *longa dependência*, isto é, mesmo para tempos bastante distantes entre si, a correlação entre as variáveis aleatórias é não desprezível. Neste trabalho, verificamos se esta longa dependência pode ser explicada pelos processos auto-regressivos médias móveis fracionariamente integráveis (ARFIMA(p, d, q)), através da análise de diversas seqüências de DNA em todos os domínios da vida. Para estimar o parâmetro de diferenciação d utilizamos os seguintes métodos de estimação: semiparamétrico baseado na equação de regressão linear utilizando a função periodograma, em versão clássica e robusta; o da máxima verossimilhança (ver Fox e Taqqu, 1986), utilizando a aproximação sugerida por Whittle (1953) e o método semiparamétrico R/S(n), proposto por Hurst (1951). O objetivo principal deste trabalho é analisar o método da análise de flutuações destendenciadas ("*Detrended Fluctuation Analysis*" - DFA), proposto por Peng et al. (1994). Este método é estabelecido como uma importante ferramenta para detectar longa dependência em séries temporais não estacionárias. Descrevemos o método DFA e analisamos sua consistência e distribuição assintótica como um estimador para o parâmetro fracionário d.

Abstract

In the literature it is stated that the DNA sequences present the long-range dependence property. In this work, we analyze this long dependence property in view of the autoregressive moving average fractionally integrated ARFIMA(p, d, q) processes through the analysis of several DNA sequences in all life domain. For estimating the fractional parameter d we consider the following estimation methods: the semiparametric regression method based on the periodogram function, in both classical and robust version; the maximum likelihood method (see Fox and Taqqu, 1986), by considering the approximation suggested by Whittle (1953) and the semiparametric R/S(n) method, proposed by Hurst (1951). 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. In this work we describe the DFA method and we prove its consistency and its asymptotic distribution as an estimator for the fractional parameter d.

ÍNDICE

Capítulo 1

Introdução

Uma das áreas de séries temporais mais frutíferas, no momento, é aquela de séries temporais que apresentam característica de *longa dependência*, isto é, mesmo para tempos bastante distantes entre si, a correlação entre as variáveis aleatórias é não desprezível.

O estudo de séries temporais com a característica de longa dependência foi apresentado, primeiramente, pelo hidrologista Harold E. Hurst em 1951 enquanto investigava a série temporal das vazões do rio Nilo.

Um dos modelos que descreve a *persistência* ou *longa dependência* são os chamados processos auto-regressivos médias móveis fracionariamente integráveis, denotados por ARFIMA(p, d, q), onde d é o *parâmetro fracionário* e p e q são, respectivamente, os graus dos polinômios auto-regressivo e média móvel.

Existem diferentes propostas para a estimação dos parâmetros do modelo ARFIMA, tanto na classe paramétrica como na semiparamétrica. Métodos paramétricos consistem da estimação simultânea dos parâmetros do modelo, em geral por máxima verossimilhança (ver Fox e Taqqu, 1986 e Sowell, 1992). No procedimento semiparamétrico, a estimação dos parâmetros do modelo é feita em dois passos: primeiro estima-se o parâmetro de diferenciação d através, por exemplo, de um modelo de regressão linear baseado na função periodograma e, posteriormente, estimam-se os parâmetros auto-regressivos e de médias móveis. O estimador mais conhecido dentro dessa classe foi proposto por Geweke e Porter-Hudak (1983).

Qualquer registro no tempo pode ser considerado uma série temporal: a temperatura média de uma cidade durante um certo período de tempo, o valor das ações de uma determinada empresa na Bovespa ou o nível das águas de um rio através dos tempos. As seqüências de DNA podem também serem consideradas como séries temporais (ver Peng et al., 1992; Guharay et al., 2000; Cristea, 2002 e Stoffer e Rosen, 2007). Para obter uma série temporal à partir de uma seqüência de DNA é necessário fazer uso de alguma transformação.

Conforme diversos artigos (ver Li e Kaneko, 1992; Peng et al., 1992;

Borstnik et al., 1993 e Lopes e Nunes, 2006 entre outros) as seqüências de DNA apresentam *longa dependência*.

Introns são seqüências nucleotídicas que não geram proteínas, ao contrário dos *exons*, que as geram. No Capítulo 4 apresentamos conceitos básicos de biologia molecular. O estudo da detecção de longa dependência em *introns* e *exons* em seqüências de DNA foi apresentado, primeiramente, por Peng et al. (1992). Estes autores mostram que existe longa dependência em seqüências de *introns*, enquanto que as seqüências formadas por *exons* não apresentam tal característica. Através de uma análise mais detalhada, Chatzidimitriou-Dreismann e Larhammar (1993) concluem que *exons* e *introns* apresentam a característica de longa dependência. Buldyrev et al. (1995) utilizam a seqüência de DNA completa para mostrar que existe longa dependência em *introns*. Portanto, não há ainda uma resposta conclusiva sobre esta análise e este é ainda assunto de pesquisa na literatura. Neste trabalho analisamos seqüências de DNA completas, ou seja, seqüências com *exons* e *introns*.

O método da análise de flutuações destendenciadas ("*Detrended Fluctuation Analysis*" - DFA), criado por Peng et al. (1994), é um exemplo de metodologia recente, sendo utilizada em um crescente número de aplicações. Técnicas como esta tem como objetivo o cálculo de uma flutuação estatística $F(l)$, onde l representa o tamanho de uma janela, para mapear um conjunto de medidas. Variando o tamanho de l, as flutuações podem ser caracterizadas através de um expoente de escala obtido a partir da curva ajustada ao gráfico $\ln(F(l))$ versus $\ln(l)$ (ver Peng et al., 2004 e Kantelhardt et al., 2001). O método DFA é conhecido como uma importante ferramenta para detectar longa dependência em séries temporais não estacionárias.

Conforme diversos artigos, o método DFA vem sendo aplicado em diferentes campos de interesse, como por exemplo, para identificar longa dependência em seqüências de DNA (ver Buldyrev et al., 1995; Peng et al., 1994 e Buldyrev et al., 1998), para analisar séries temporais econômicas (ver Liu et al., 1997) e séries temporais climáticas (ver Koscielny-Bunde et al., 1998). O DFA é baseado na teoria do passeio aleatório ("*random walk theory*") (ver Shlesinger e Klafter, 1987 e Ben-Avraham e Havlin, 2000) e é similar ao método R/S(n) ("*Rescaled Range Analysis*") (ver Hurst et al., 1965) e ao método baseado na transformada de *wavelet* (ver Koscielny-Bunde et al., 1998).

Para confiar na correta identificação de longa dependência em uma série temporal, é essencial distinguir tendências de altas flutuações, que estão intrínsecas aos dados. As tendências são causadas por efeitos externos e elas são usualmente supostas terem comportamento monótono e suave. Dependendo do método utilizado, altas tendências nos dados podem levar à falsa identificação de longa dependência. A vantagem do método da análise de flutuações destendenciadas é que a tendência pode ser sistematicamente eliminada (ver Kantelhardt et al., 2001).

O objetivo deste trabalho é analisar as propriedades estatísticas do método da análise de flutuações destendenciadas (DFA). Estamos interessados em analisar o parâmetro de *longa dependência* em seqüências de DNA, por meio do método DFA e de diversos outros métodos de estimação para o *parâmetro de diferenciação* d, tanto dentro das classes paramétrica como semiparamétrica.

O Capítulo 2 apresenta conceitos básicos para a análise de processos estocásticos e de séries temporais. Neste capítulo são apresentados importantes processos estocásticos, tais como os processos Gaussianos, os processos ruído branco e os processos Browniano fracionários. No final deste capítulo descrevemos os processos auto-regressivos médias móveis fracionariamente integráveis (ARFIMA) e suas principais propriedades.

No Capítulo 3 apresentamos os estimadores para o parâmetro de diferenciação d, dentro das classes paramétrica e semiparamétrica e em seguida descrevemos o método R/S(n), proposto por Hurst (1951). No final do capítulo descrevemos o método DFA ("*Detrended Fluctuation Analysis*"), proposto por Peng et al. (1994) e analisamos as suas propriedades estatísticas.

O Capítulo 4 apresenta uma breve introdução à *molécula de DNA*. Definimos uma *seqüência* de DNA e descrevemos funções que transformam *seqüências de DNA* em seqüências numéricas. Por fim, apresentamos a definição da série temporal utilizada para representar uma seqüência de DNA neste trabalho.

A análise de diversas seqüências de DNA, com o uso dos softwares S-Plus e R-project, é apresentada no Capítulo 5. Utilizamos seqüências de nucleotídeos disponíveis no Instituto Europeu de Bioinformáticas ("*European Bioinformatics Institute*" - EBI) e no Centro Nacional de Informação Biotecnológica ("*National Center for Biotechnology Information*" - NCBI).

As conclusões finais e propostas para futuros trabalhos são apresentadas no Capítulo 6.

Capítulo 2

Processos Estocásticos

Apresentamos, neste capítulo, alguns conceitos básicos para a análise de processos estocásticos e séries temporais, tais como função de autocovariância, função de autocorrelação, estacionariedade e longa dependência. Apresentamos também importantes processos estocásticos, como por exemplo: os processos Gaussianos, os processos ruído branco e os processos auto-regressivos médias móveis (ARMA). Por fim, descrevemos os modelos auto-regressivos médias móveis fracionariamente integráveis (ARFIMA) e suas principais propriedades.

Um dos modelos que descrevem a *persistência* ou *longa dependência* são os chamados processos ARFIMA(p, d, q), onde d é o parâmetro fracionário e p e q são os graus dos polinômios auto-regressivo e média móvel, respectivamente. A característica de *longa dependência* em uma série temporal, significa que mesmo para tempos bastante distantes entre si, a correlação entre as variáveis aleatórias é não desprezível. Esta característica tem sido observada em séries temporais de diferentes áreas de estudos tais como meteorologia, astronomia, hidrologia e economia. Podemos caracterizar a *persistência* de duas formas:

• no domínio do tempo, a função de autocorrelação decai lentamente a zero.

• no domínio da freqüência, a função densidade espectral tende ao infinito quando a freqüência se aproxima de zero.

Estudos iniciais de séries temporais com características de *longa dependência* foram realizados por Hurst (1951), Mandelbrot e Van Ness (1968). Hosking (1981 e 1984) foi o pioneiro na aplicação de *longa dependência* em séries hidrológicas. Neste capítulo apresentamos o modelo ARFIMA(p, d, q) e alguns resultados teóricos a ele relacionados (ver Lopes et al., 2004; Lopes, 2007 e Olbermann, 2002). Modelos que incluem diferenciação fracionária $d \in (0,0; 0,5)$ são capazes de representar séries temporais que apresentam característica de *persistência*, também chamada *longa dependência* ("*long memory*") (ver Beran, 1994 e Doukhan et al., 2003 para um estudo completo

destes processos).

Definição 2.1 (Processo Estocástico). Um *processo estocástico* é uma família de variáveis aleatórias $\{X_t\}_{t\in T}$ definidas em um mesmo espaço de probabilidades $(\Omega, \mathfrak{F}, \mathbb{P})$ sendo T um conjunto de índices, Ω o espaço amostral, $\mathfrak{F}$ a classe de eventos aleatórios e $\mathbb{P} : \mathfrak{F} \longrightarrow [0;1]$ a função que associa probabilidade a um evento aleatório qualquer. O conjunto T pode ser discreto ou contínuo.

Definição 2.2 (Série Temporal). Uma *série temporal* é uma amostra aleatória $\{X_t\}_{t=1}^{n}$ de um processo estocástico $\{X_t\}_{t\in T}$, onde n é o tamanho amostral.

Observação 2.1. Neste trabalho consideramos somente o caso em que $T = \mathbb{Z}$.

Para se ter uma idéia do grau de dependência entre um número finito de variáveis aleatórias é sempre útil calcular a matriz de covariância. Para um processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ é preciso estender a noção de matriz de covariância, e esta extensão é proporcionada pela *função de autocovariância.*

Definição 2.3 (Função de Autocovariância). Se $\{X_t\}_{t\in\mathbb{Z}}$ é um processo estocástico tal que $Var(X_t) < \infty$, para todo $t \in \mathbb{Z}$, então a *função de autocovariância* $\gamma_X(\cdot,\cdot)$ de $\{X_t\}_{t\in\mathbb{Z}}$ é definida por

$$\gamma_X(r,s) = Cov(X_r, X_s) = \mathbb{E}[(X_r - \mathbb{E}(X_r))(X_s - \mathbb{E}(X_s))], \quad \text{para todo } r,s \in \mathbb{Z}. \tag{2.1}$$

Definição 2.4 (Função de Autocorrelação). Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico tal que $Var(X_t) < \infty$, para todo $t \in \mathbb{Z}$. A *função de autocorrelação* do processo, denotada por $\rho_X(\cdot,\cdot)$ é dada por

$$\rho_X(r,s) = \frac{\gamma_X(r,s)}{\sqrt{Var(X_r)}\sqrt{Var(X_s)}}, \quad \text{para todo } r,s \in \mathbb{Z}.$$

Definição 2.5 (Processo Estacionário). Um processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ é dito *estacionário* se e somente se:

(a) $\mathbb{E}(|X_t|^2) < \infty$, $\forall t \in \mathbb{Z}$

(b) $\mathbb{E}(X_t) = m$, $\forall t \in \mathbb{Z}$

(c) $\gamma_X(r,s) = \gamma_X(r+t, s+t)$, $\forall r,s,t \in \mathbb{Z}$.

Observação 2.2.

1) A estacionariedade definida acima é também chamada *estacionariedade fraca, estacionariedade no sentido largo, estacionariedade de segunda ordem* ou *estacionariedade de covariância.*

2) Se $\{X_t\}_{t\in\mathbb{Z}}$ é estacionário então $\gamma_X(r,s) = \gamma_X(r-s,0)$, para todo $r, s \in \mathbb{Z}$ (basta tomar $t = -s$ no item (c) da Definição 2.5). Portanto, é conveniente redefinir a *função de autocovariância* por

$$\gamma_X(k) \equiv \gamma_X(k,0) = Cov(X_{t+k}, X_t), \quad \text{para todo } t, k \in \mathbb{Z}. \tag{2.2}$$

Note que $Var(X_k) = \gamma_X(0,0) \equiv \gamma_X(0)$.

3) Se $\{X_t\}_{t\in\mathbb{Z}}$ é estacionário então a *função de autocorrelação* é definida por

$$\rho_X(k) \equiv \frac{\gamma_X(k)}{\gamma_X(0)} = Corr(X_{t+k}, X_t), \quad \text{para todo } t, k \in \mathbb{Z},$$

onde $\gamma_X(0) = Var(X_k)$.

Definição 2.6 (Processo Estritamente Estacionário). Um processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ é dito *estritamente estacionário* ou *fortemente estacionário* se e somente se as distribuições conjuntas de $(X_{t_1}, \cdots, X_{t_l})'$ e de $(X_{t_1+k}, \cdots, X_{t_l+k})'$ são as mesmas, para todo $l \in \mathbb{N}$ e para todo $t_1, \cdots, t_l$, $k \in \mathbb{Z}$.

Definição 2.7 (Processo Gaussiano). Um processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ é dito *Gaussiano* se, para qualquer conjunto $t_1, t_2, \cdots, t_n \in \mathbb{Z}$, as variáveis aleatórias $X_{t_1}, X_{t_2}, \cdots, X_{t_n}$ têm uma distribuição normal n-dimensional.

Observamos que um processo $\{X_t\}_{t\in\mathbb{Z}}$ fracamente estacionário não precisa ser fortemente estacionário. No entanto, como um processo Gaussiano, com variância finita, é determinado pelas médias e covariâncias, se ele for fracamente estacionário, será também fortemente estacionário (ver Brockwell e Davis, 1991 e Priestley, 1981).

Proposição 2.1. *Seja $\gamma_X(\cdot)$ a função de autocovariância de um processo estacionário $\{X_t\}_{t\in\mathbb{Z}}$. Então,*

(a) $\gamma_X(0) \geq 0$

(b) $| \gamma_X(k) | \leq \gamma_X(0), \quad \forall k \in \mathbb{Z}$

(c) $\gamma_X(k) = \gamma_X(-k), \quad \forall k \in \mathbb{Z}.$

Observação 2.3. O item (b) da Proposição 2.1 é uma conseqüência imediata da desigualdade de Cauchy-Schwarz,

$$|Cov(X_{t+k}, X_t)| \leq \sqrt{Var(X_{t+k})}\sqrt{Var(X_t)}.$$

Definição 2.8 (Estimador Não Viciado, U.M.V.U. e Consistente). Sejam $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico estacionário e $\{X_t\}_{t=1}^n$ uma amostra deste processo. Seja $\theta \in \mathbb{R}$ um parâmetro qualquer do processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ e seja $T(X_1, \cdots, X_n)$ um estimador qualquer de θ.

i) Se $\mathbb{E}(T(X_1, \cdots, X_n)) = \theta$, então $T(X_1, \cdots, X_n)$ é um *estimador não viciado* de θ.

ii) Se $Var(T^*(X_1, \cdots, X_n)) \leq Var(T(X_1, \cdots, X_n))$, para todo estimador $T(X_1, \cdots, X_n)$ não viciado de θ, então $T^*(X_1, \cdots, X_n)$ é dito ser um *estimador U.M.V.U.* ("*Uniformly Minimum Variance Unbiased*") para θ.

iii) Seja $\{T_n(X_1, \cdots, X_n)\}_{n\geq 1}$ uma seqüência de estimadores de θ, que são gerados similarmente para cada $n \geq 1$. Então, T_n é um estimador *consistente* para θ, se e somente se,

$$\lim_{n\to\infty} P\big[|T_n(X_1, \cdots, X_n) - \theta| \geq \epsilon\big] = 0,$$

para todo $\epsilon > 0$.

Freqüentemente, desejamos estimar a função de autocovariância $\gamma_X(\cdot)$ a partir das observações X_t, $1 \leq t \leq n$. Esta estimativa nos dará informações sobre a estrutura de dependência do processo. Usaremos como estimador de $\gamma_X(\cdot)$ a função de autocovariância amostral definida a seguir.

Definição 2.9 (Função de Autocovariância Amostral). A *função de autocovariância amostral* é definida por

$$\widehat{\gamma}_X(k) = \frac{1}{n}\sum_{j=1}^{n-k}(X_{j+k} - \overline{X})(X_j - \overline{X}), \quad 0 \leq k < n, \tag{2.3}$$

e $\widehat{\gamma}_X(k) = \widehat{\gamma}_X(-k)$, $-n < k \leq 0$, onde $\overline{X} = \frac{1}{n}\sum_{j=1}^n X_j$ *é a média amostral.*

Definição 2.10 (Função de Autocorrelação Amostral). A *função de autocorrelação amostral* é definida por

$$\widehat{\rho}_X(k) \equiv \frac{\widehat{\gamma}_X(k)}{\widehat{\gamma}_X(0)}, \quad |\, k \,| < n,$$

onde $\widehat{\gamma}_X(\cdot)$ é dada pela expressão (2.3).

O teorema a seguir caracteriza a função de autocovariância que pode ser escrita na forma

$$\gamma_X(k) = \int_{(-\pi,\pi]} e^{iwk} dF_X(w),$$

para alguma função $F_X(\cdot)$ com massa concentrada em $(-\pi, \pi]$.

Teorema 2.1 (Herglotz). *Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico estacionário e seja $\gamma_X(\cdot)$ sua função de autocovariância, conforme expressão (2.2). Então, a função $\gamma_X(\cdot)$ é função definida não negativa se e somente se*

$$\gamma_X(k) = \int_{(-\pi,\pi]} e^{iwk} dF_X(w), \quad \textit{para todo } k \in \mathbb{Z},$$

onde $F_X(\cdot)$ é uma função contínua à direita, não decrescente e limitada em $[-\pi, \pi]$ e $F_X(-\pi) = 0$. (A função $F_X(\cdot)$ é denominada ***função de distribuição espectral*** *de $\gamma_X(\cdot)$ (ou de $\{X_t\}_{t\in\mathbb{Z}}$). Se $F_X(w) = \int_{-\pi}^{w} f_X(\nu)d\nu$, $-\pi \leq w \leq \pi$, então $f_X(\cdot)$ é denominada* ***função densidade espectral*** *de $\gamma_X(\cdot)$ (ou de $\{X_t\}_{t\in\mathbb{Z}}$).)*

A demonstração do Teorema 2.1 pode ser encontrada em Brockwell e Davis (1991), página 116.

Observação 2.4. A análise de Fourier nos diz que a função de autocovariância do processo estocástico estacionário $\{X_t\}_{t\in\mathbb{Z}}$ pode ser obtida através da função densidade espectral $f_X(\cdot)$, usando a transformada inversa de Fourier

$$\gamma_X(k) = \int_{-\pi}^{\pi} f_X(w) e^{iwk} dw. \tag{2.4}$$

A seguir definimos dois estimadores da função densidade espectral (ver Definições 2.11 e 2.12).

Definição 2.11 (Função Periodograma). Sejam $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico estacionário e $\{X_t\}_{t=1}^{n}$ uma série temporal obtida a partir deste processo. A *função periodograma* calculada à partir da série temporal $\{X_t\}_{t=1}^{n}$, é definida por

$$I_n(w) = \frac{1}{2\pi}\left(\widehat{\gamma}_X(0) + 2\sum_{k=1}^{n-1} \widehat{\gamma}_X(k)\cos(wk)\right), \quad w \in [-\pi, \pi], \tag{2.5}$$

onde $\widehat{\gamma}_X(\cdot)$ é a função de autocovariância amostral do processo $\{X_t\}_{t\in\mathbb{Z}}$ dada na Definição 2.9.

A *função periodograma* é um estimador não-viciado mas, inconsistente, para a função densidade espectral $f_X(\cdot)$. Na próxima definição apresentamos um estimador consistente para a função densidade espectral.

Definição 2.12 (Função Periodograma Suavizado). Sejam $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico estacionário e $\{X_t\}_{t=1}^{n}$ uma série temporal obtida a partir deste processo. A *função periodograma suavizado* calculada à partir da série temporal $\{X_t\}_{t=1}^{n}$, denotada por $f_s(\cdot)$, é definida por

$$f_s(w) = \frac{1}{2\pi}\sum_{k=-(n-1)}^{n-1} \lambda\left(\frac{k}{n-1}\right)\widehat{\gamma}_X(k)\cos(wk), \quad w \in [-\pi, \pi],$$

onde $\widehat{\gamma}_X(\cdot)$ é a função de autocovariância amostral do processo definida em (2.3) e $\lambda(\cdot)$ é uma função de ponderação, sendo função par e contínua, satisfazendo $\lambda(0) = 1$, $|\lambda(x)| \leq 1$, para todo x e $\lambda(x) = 0$, para todo $|x| > 1$.

Notações:

1) Se, para a seqüência $\{a_n\}_{n\in\mathbb{N}}$, existe um número real $u \in \mathbb{R}$ e constantes $c_1, c_2 > 0$ tais que, para todo $n \in \mathbb{N}$, vale

$$c_1 \leq \left| \frac{a_n}{n^{-u}} \right| \leq c_2,$$

então, denotamos $a_n \approx n^{-u}$.

2) Se, para a função $g(\cdot)$, existe um número real $b \in \mathbb{R}$ e constantes $d_1, d_2 > 0$ tais que, para todo x, vale

$$d_1 \leq \left| \frac{g(x)}{x^b} \right| \leq d_2,$$

então, denotamos $g(x) \approx x^b$.

A seguir, introduzimos uma definição formal para a propriedade de *longa dependência*.

Definição 2.13 (Longa Dependência). Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico estacionário. Se existe um número real $u \in (0, 1)$ tal que

$$\rho_X(k) \approx k^{-u},$$

onde $\rho_X(\cdot)$ é a função de autocorrelação do processo, ou equivalentemente, se existe um número real $b \in (0, 1)$ tal que

$$f_X(w) \approx w^b,$$

onde $f_X(\cdot)$ é a função densidade espectral do processo, então dizemos que $\{X_t\}_{t\in\mathbb{Z}}$ é um *processo estocástico estacionário com longa dependência* (ou com *memória longa*).

Definição 2.14 (Filtro Linear). Sejam $\{X_t\}_{t\in\mathbb{Z}}$ e $\{Y_t\}_{t\in\mathbb{Z}}$ dois processos estocásticos quaisquer. Dizemos que o processo $\{Y_t\}_{t\in\mathbb{Z}}$ é obtido do processo $\{X_t\}_{t\in\mathbb{Z}}$ pela aplicação de um *filtro linear* $C = \{c_{t,k}; t, k \in \mathbb{Z}\}$ se

$$Y_t = \sum_{k\in\mathbb{Z}} c_{t,k} X_k, \quad \text{para todo } t \in \mathbb{Z}, \tag{2.6}$$

onde os coeficientes $c_{t,k}$ são chamados de *pesos do filtro*. O filtro C é dito ser *invariante no tempo* se $c_{t,k}$ depende somente de $t - k$, i.e., se $c_{t,k} = h_{t-k}$.

Exemplo 2.1. Seja o processo estocástico $\{Y_t\}_{t\in\mathbb{Z}}$ obtido do processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ através da aplicação do *filtro linear* dado por

$$Y_t = X_t + 1,1\, X_{t-1} + 0,2\, X_{t-2}.$$

Com base na Definição 2.14, observamos que o filtro linear $C = \{c_{t,k}; t, k \in \mathbb{Z}\}$ tem coeficientes $c_{t,0} = 1,0$, $c_{t,1} = 1,1$, $c_{t,2} = 0,2$ e $c_{t,k} = 0$ para todo $k \notin \{0,1,2\}$.

Definição 2.15 (Processo de Incremento Ortogonal). Um processo estocástico de valor complexo $\{Z(w)\}_{w=-\pi}^{\pi}$ é dito ser um *processo de incremento ortogonal* sobre $[-\pi,\pi]$ se

(a) $\big\langle Z(w), Z(w)\big\rangle < \infty, \quad -\pi \le w \le \pi,$

(b) $\big\langle Z(w), 1\big\rangle = 0, \quad -\pi \le w \le \pi,$

(c) $\big\langle Z(w_4) - Z(w_3), Z(w_2) - Z(w_1)\big\rangle = 0, \quad$ se $(w_1, w_2] \cap (w_3, w_4] = \emptyset,$

onde o produto interno é definido por $\big\langle X, Y\big\rangle = \mathbb{E}(X\overline{Y})$.

Observação 2.5. Na Definição 2.15, o item (a) diz que $\mathbb{E}(Z^2(w)) < \infty$; os itens (a) e (b) garantem que $Var(Z(w)) < \infty$ e o item (c) diz que $Cov(Z(w_4) - Z(w_3), Z(w_2) - Z(w_1)) = 0$.

Definição 2.16 (Contínuo à Direita). O processo $\{Z(w)\}_{w=-\pi}^{\pi}$, dado na Definição 2.15, é dito ser *contínuo à direita* se, para todo $w \in [-\pi,\pi)$,

$$\|Z(w+\delta) - Z(w)\|^2 = \mathbb{E}|Z(w+\delta) - Z(w)|^2 \to 0, \quad \text{quando } \delta \downarrow 0.$$

O teorema a seguir fornece a representação espectral para um processo estacionário $\{X_t\}_{t\in\mathbb{Z}}$ e sua demonstração pode ser encontrada em Brockwell e Davis (1991).

Teorema 2.2 (Teorema da Representação Espectral). *Se $\{X_t\}_{t\in\mathbb{Z}}$ é um processo estocástico estacionário com média $\mu = 0$ e função de distribuição espectral $F_X(\cdot)$, então existe um processo de incremento ortogonal $\{Z_X(w)\}_{w=-\pi}^{\pi}$ contínuo à direita tal que*

(a) $\mathbb{E}|Z_X(w) - Z_X(-\pi)|^2 = F_X(w), \quad -\pi \le w \le \pi,$

(b) $X_t = \int_{(-\pi,\pi)} e^{it\nu} dZ_X(\nu).$

Observação 2.6. A integral mencionada no Teorema da Representação Espectral é uma integral estocástica com respeito a um processo de incremento ortogonal. Para mais detalhes ver Brockwell e Davis (1991), página 135.

Dado um processo estocástico estacionário $\{X_t\}_{t\in\mathbb{Z}}$, o Teorema 2.3, a seguir, mostra a relação entre a função de distribuição espectral do processo estocástico $\{Y_t\}_{t\in\mathbb{Z}}$, definido por

$$Y_t = \sum_{j\in\mathbb{Z}} h_j X_{t-j},$$

e a função de distribuição espectral do processo $\{X_t\}_{t\in\mathbb{Z}}$. Observe que $C = \{h_j; j \in \mathbb{Z}\}$ é um filtro linear, dado na Definição 2.14, e é invariante no tempo.

Teorema 2.3. *Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico estacionário com média zero, representação espectral*

$$X_t = \int_{(-\pi,\pi)} e^{it\nu} dZ_X(\nu) \tag{2.7}$$

e função de distribuição espectral $F_X(\cdot)$. Suponha que $C = \{h_j\}_{j\in\mathbb{Z}}$ é um filtro linear invariante no tempo tal que

$$\sum_{j=-n}^{n} h_j e^{-ij\cdot} \tag{2.8}$$

converge na norma $L^2(F_X)$ para

$$\sum_{j\in\mathbb{Z}} h_j e^{-ij\cdot} \equiv h(e^{-i\cdot}), \quad \textit{quando } n \to \infty. \tag{2.9}$$

Então, o processo estocástico

$$Y_t = \sum_{j\in\mathbb{Z}} h_j X_{t-j} \tag{2.10}$$

é estacionário com média zero, função distribuição espectral dada por

$$F_Y(w) = \int_{-\pi}^{w} \left|h(e^{-i\nu})\right|^2 dF_X(\nu) \tag{2.11}$$

e representação espectral dada por

$$Y_t = \int_{-\pi}^{\pi} e^{it\nu} h(e^{-i\nu}) dZ_X(\nu). \tag{2.12}$$

A demonstração pode ser encontrada em Brockwell e Davis (1991), página 149.

Observação 2.7. Como conseqüência do Teorema 2.3 observe que se $\{X_t\}_{t\in\mathbb{Z}}$ tem função densidade espectral $f_X(\cdot)$ e $\{Y_t\}_{t\in\mathbb{Z}}$ é um processo dado por (2.10), com pesos $\{h_j\}_{j\in\mathbb{Z}}$ tais que

$$\sum_{j\in\mathbb{Z}} \left|h_j\right| < \infty,$$

então o processo $\{Y_t\}_{t\in\mathbb{Z}}$ tem função densidade espectral $f_Y(\cdot)$ dada por

$$f_Y(w) = \left|h(e^{-iw})\right|^2 f_X(w), \tag{2.13}$$

onde

$$h(e^{-iw}) = \sum_{j\in\mathbb{Z}} h_j e^{-iwj}.$$

Definição 2.17 (Ruído Branco). O processo $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ é denominado um *ruído branco* com média zero e variância σ_ε^2, denotado por $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, se

$$\mathbb{E}(\varepsilon_t) = 0, \quad Var(\varepsilon_t) = \mathbb{E}(\varepsilon_t^2) = \sigma_\varepsilon^2 \quad \text{e} \quad \gamma_\varepsilon(k) = \begin{cases} \sigma_\varepsilon^2, & k = 0, \\ 0, & k \neq 0. \end{cases} \tag{2.14}$$

Observação 2.8. Observe que um processo ruído branco é formado por variáveis aleatórias com média zero, variância finita e constante e não correlacionadas.

Para se ter uma idéia mais geral sobre processos estocásticos com *longa dependência*, tratamos a seguir, daqueles processos onde o conjunto de índices T, da Definição 2.1, é $\mathbb{R}^+$.

Definição 2.18 (Processo de Wiener Padrão). Um processo estocástico $\{W_t\}_{t\in\mathbb{R}^+}$ é denominado um *processo de Wiener padrão* ("*standard Wiener process*") ou *Movimento Browniano* se, para $t \geq 0$, os incrementos são independentes e estacionários e, para cada $t \in \mathbb{R}^+$, W_t é variável aleatória normalmente distribuída com $\mathbb{E}(W_t) = 0$ e $Var(W_t) = t$.

Definição 2.19 (Processo Ponte Browniana). Um processo estocástico $\{W_t^0\}_{t\in\mathbb{R}^+}$ é denominado um processo *Ponte Browniana* ("*Brownian bridge*"), se

$$W_t^0 = W_t - tW_1,$$

onde $\{W_t\}_{t\in\mathbb{R}^+}$ é dado pela Definição 2.18.

Definição 2.20 (Movimento Browniano Fracionário). O processo estocástico $\{B_{\mathrm{H}}(t)\}_{\mathbb{R}^+}$ é denominado um *movimento Browniano fracionário*, se é um processo Gaussiano (ver Definição 2.7) com média $\mu = 0$, incrementos estacionários, variância $\mathbb{E}(B_{\mathrm{H}}^2(t)) = t^{2\mathrm{H}}$ e autocovariância dada por

$$\mathbb{E}(B_{\mathrm{H}}(s)B_{\mathrm{H}}(t)) = \frac{1}{2}\left\{ s^{2\mathrm{H}} + t^{2\mathrm{H}} - \left|s-t\right|^{2\mathrm{H}} \right\}, \quad \text{para todo } t, s \in \mathbb{R}^+. \tag{2.15}$$

Observação 2.9. O índice H na Definição 2.20 é o parâmetro sugerido por Harold Edwin Hurst (1880-1978), para medir longa dependência, nomeado H *de Hurst*.

Definição 2.21 (Ruído Gaussiano Fracionário). O processo estocástico $\{X_t\}_{t\in\mathbb{R}^+}$ é denominado um *ruído Gaussiano fracionário*, se é o incremento do *movimento Browniano fracionário*, ou seja,

$$X_t = B_{\mathrm{H}}(t+1) - B_{\mathrm{H}}(t), \quad \text{para todo } t \in \mathbb{R}^+. \tag{2.16}$$

Observação 2.10.

1) A função de autocovariância de um *ruído Gaussiano fracionário* é dada por

$$\gamma_X(k) = \frac{1}{2}\left\{(k+1)^{2\mathrm{H}} - 2k^{2\mathrm{H}} + |k-1|^{2\mathrm{H}}\right\}, \quad k \geq 0. \tag{2.17}$$

2) Se $\mathrm{H} \neq \frac{1}{2}$ então

$$\gamma_X(k) \sim \mathrm{H}(2\mathrm{H}-1)k^{2\mathrm{H}-2}, \quad \text{quando } k \to \infty. \tag{2.18}$$

3) Se $\mathrm{H} = \frac{1}{2}$ e $k \geq 1$ então $\gamma_X(k) = 0$. Neste caso, $\{X_t\}_{t\in\mathbb{R}^+}$ é um *ruído branco.*

4) Se $\frac{1}{2} < \mathrm{H} < 1$ então o processo $\{X_t\}_{t\in\mathbb{R}^+}$ é formado por variáveis aleatórias possivelmente correlacionadas e podemos dizer que existe *longa dependência* ou *longa correlação.*

5) A função densidade espectral de um *ruído Gaussiano fracionário* é dada por

$$f_X(w) = C_{\mathrm{H}}\left(2\,\mathrm{sen}\left(\frac{w}{2}\right)\right)^2 \sum_{j=-\infty}^{\infty} \frac{1}{|w+2\pi j|^{2\mathrm{H}+1}} \sim C_{\mathrm{H}}|w|^{1-2\mathrm{H}}, \quad w \to 0, \tag{2.19}$$

onde C_{H} é uma constante.

A partir de agora, o conjunto de índices T, dado na Definição 2.1, é considerado ser $T = \mathbb{Z}$.

Definição 2.22 (Operador Diferença). Seja $\mathcal{B}$ o *operador defasagem*, i.e., $\mathcal{B}^j(X_t) = X_{t-j}$, para todo $j \in \mathbb{N}\cup\{0\}$. Para todo $d \in \mathbb{R}$, definimos o *operador diferença* $\nabla^d \equiv (1-\mathcal{B})^d$ através da expansão binomial

$$\nabla^d \equiv (1-\mathcal{B})^d = \sum_{j=0}^{\infty} \binom{d}{j} (-\mathcal{B})^j = 1 - d\mathcal{B} - \frac{d}{2!}(1-d)\mathcal{B}^2 \cdots \tag{2.20}$$

onde

$$\binom{d}{j} = \frac{\Gamma(d+1)}{\Gamma(j+1)\Gamma(d-j+1)}, \tag{2.21}$$

com $\Gamma(\cdot)$ a função Gama.

Apresentamos, a seguir, processos estocásticos definidos em termos de equações de diferenças lineares com coeficientes constantes, chamados *processos auto-regressivos médias móveis de ordens p e q*, denotados por ARMA(p, q). Para maiores detalhes, ver Brockwell e Davis (1991) e Box et al. (1994).

Definição 2.23 (Processo ARMA(p, q)). O processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ é denominado um *processo auto-regressivo média móvel de ordens*, respectivamente, p e q, com média μ, denotado por ARMA(p, q), se é um processo estacionário tal que

$$\Phi(\mathcal{B})(X_t - \mu) = \Theta(\mathcal{B})\varepsilon_t, \quad \text{para todo } t \in \mathbb{Z}, \tag{2.22}$$

onde $\Phi(\cdot)$ e $\Theta(\cdot)$ são os polinômios de ordens p e q dados, respectivamente, por

$$\begin{aligned} \Phi(\mathcal{B}) &= 1 - \phi_1\mathcal{B} - \ldots - \phi_{\mathrm{p}}\mathcal{B}^{\mathrm{p}} \\ \Theta(\mathcal{B}) &= 1 - \theta_1\mathcal{B} - \ldots - \theta_{\mathrm{q}}\mathcal{B}^{\mathrm{q}} \end{aligned} \tag{2.23}$$

onde ϕ_l, $1 \leq l \leq p$, e θ_j, $1 \leq j \leq q$, são constantes reais.

Se $\Theta(\mathcal{B}) \equiv 1$ o processo $\Phi(\mathcal{B})(X_t - \mu) = \varepsilon_t$ é um *processo auto-regressivo de ordem p*, denotado por AR(p). Da mesma forma, se $\Phi(\mathcal{B}) \equiv 1$, o processo $X_t - \mu = \Theta(\mathcal{B})\varepsilon_t$ é um *processo média móvel de ordem q*, denotado por MA(q).

Definição 2.24 (ARMA(p, q) Causal). Um processo ARMA(p, q), dado pela expressão (2.22) é denominado *causal* se existe uma seqüência de constantes $\{\psi_j\}_{j\in\mathbb{N}\cup\{0\}}$ tal que

$$\sum_{j\geq 0} |\psi_j| < \infty \qquad \text{e} \qquad X_t = \sum_{j\geq 0} \psi_j \varepsilon_{t-j}, \quad \text{para todo } t \in \mathbb{Z}. \tag{2.24}$$

Definição 2.25 (ARMA(p, q) Inversível). Um processo ARMA(p, q), definido pela expressão (2.22), é denominado *inversível* se existe uma seqüência de constantes $\{\pi_j\}_{j\in\mathbb{N}\cup\{0\}}$ tal que

$$\sum_{j\geq 0} |\pi_j| < \infty \quad \text{e} \quad \varepsilon_t = \sum_{j\geq 0} \pi_j X_{t-j}, \quad \text{para todo } t \in \mathbb{Z}. \tag{2.25}$$

No teorema, a seguir, é apresentada a função densidade espectral de um processo ARMA(p, q) e a demonstração pode ser encontrada em Brockwell e Davis (1991).

Teorema 2.4. *Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico estacionário ARMA(p, q), definido em (2.22), onde $\Phi(\cdot)$ e $\Theta(\cdot)$, dados pela expressão (2.23), não*

possuem raízes em comum e $\Phi(\cdot)$ *não tem raízes no círculo unitário. Então, o processo* $\{X_t\}_{t\in\mathbb{Z}}$ *tem função densidade espectral dada por*

$$f_X(w) = \frac{\sigma_\varepsilon^2}{2\pi}\left|\frac{\Theta(e^{-iw})}{\Phi(e^{-iw})}\right|^2, \quad \forall w \in [-\pi, \pi]. \tag{2.26}$$

Apresentamos, a seguir, os *processos auto-regressivos integrados de médias móveis*.

Definição 2.26 (Processo ARIMA(p, d, q)). Se d é um inteiro não negativo, então o processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ é denominado um *processo auto-regressivo integrado de média móvel de ordens p e q*, respectivamente, denotado por ARIMA(p, d, q), com média μ, se $Y_t \equiv \nabla^d(X_t - \mu)$ é um processo ARMA(p, q) causal.

A generalização natural do processo ARIMA(p, d, q) é permitir que o parâmetro d assuma valores fracionários (ver Hosking, 1981). A seguir estendemos a classe de processos ARIMA(p, d, q), onde $d \in \mathbb{N}$, a uma classe de processos chamados *processos com diferenciação fracionária*, onde $d \in (-0,5; 0,5)$.

2.1 Processo ARFIMA(p, d, q)

Definição 2.27 (Processo ARFIMA(p, d, q)). Um processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$ é denominado um *processo geral com diferenciação fracionária*, denotado por ARFIMA(p, d, q), com média μ, se satisfaz a seguinte equação

$$\Phi(\mathcal{B})\nabla^d(X_t - \mu) = \Theta(\mathcal{B})\varepsilon_t, \quad \text{para todo } \in \mathbb{Z}, \tag{2.27}$$

onde $d \in (-0,5; 0,5)$ é o *parâmetro ou grau de diferenciação*, ∇^d é o *operador diferença* definido na expressão (2.20), $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ o *processo ruído branco* com média zero e variância $\sigma_\varepsilon^2 > 0$, $\mathcal{B}$ o *operador defasagem* e $\Phi(\cdot)$ e $\Theta(\cdot)$ são os polinômios de ordem p e q, respectivamente, dados na expressão (2.23).

Os processos ARFIMA(p, d, q) exibem a característica de *longa dependência* quando $d \in (0,0; 0,5)$, a de *curta dependência* quando $d = 0,0$ e a de *dependência intermediária* quando $d \in (-0,5; 0,0)$.

Se $d \in (-0,5; 0,5)$, então $\{X_t\}_{t\in\mathbb{Z}}$ é um processo estacionário e inversível, como veremos no Teorema 2.5, para o caso em que $p = 0 = q$ e nos Teoremas 2.6 e 2.7, para o caso geral.

Definição 2.28 (ARFIMA(p, d, q) Puro). O processo ARFIMA(p, d, q) é denominado um *processo fracionariamente integrado puro* se $p = 0 = q$, isto é, processos ARFIMA em que p e q, os graus dos polinômios $\Phi(\cdot)$ e $\Theta(\cdot)$, respectivamente, são ambos zero.

Da expressão (2.27), quando $p = 0 = q$, o processo ARFIMA$(0, d, 0)$ é representado por

$$\nabla^d (X_t - \mu) = \varepsilon_t, \quad \text{para todo } t \in \mathbb{Z}. \tag{2.28}$$

Importantes propriedades dos processos ARFIMA(p, d, q) podem ser encontradas em Hosking (1981 e 1984). A seguir apresentamos um teorema que fornece as principais propriedades dos processos ARFIMA$(0, d, 0)$. Assumiremos, por conveniência, mas sem perda de generalidade, que $\mu = 0$.

Teorema 2.5. (ver Hosking, 1981): *Seja* $\{X_t\}_{t\in\mathbb{Z}}$ *um processo estocástico* ARFIMA$(0, d, 0)$.

(a) *Quando* $d < 0,5$, $\{X_t\}_{t\in\mathbb{Z}}$ *é um processo estacionário e tem representação média móvel infinita dada por*

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

onde

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

Quando $k \to \infty$, $\psi_k \sim \frac{k^{d-1}}{\Gamma(d)}$.

(b) *Quando* $d > -0,5$, $\{X_t\}_{t\in\mathbb{Z}}$ *é um processo inversível e tem representação auto-regressiva infinita dada por*

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

onde

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

Quando $k \to \infty$, $\pi_k \sim \frac{k^{-d-1}}{\Gamma(-d)}$.

Nos itens (c), (d) *e* (e) *abaixo, assumimos que* $d \in (-0,5; 0,5)$.

(c) *A função densidade espectral de* $\{X_t\}_{t\in\mathbb{Z}}$ *é dada por*

$$f_X(w) = \frac{\sigma_\varepsilon^2}{2\pi}\left[2\,\mathrm{sen}\left(\frac{w}{2}\right)\right]^{-2d}, \qquad \textit{para todo } -\pi < w \leq \pi.$$

Quando $w \to 0$, $f_X(w) \sim \frac{\sigma_\varepsilon^2}{2\pi} w^{-2d}$.

(d) *A função de autocovariância de* $\{X_t\}_{t\in\mathbb{Z}}$ *é dada por*

$$\gamma_X(k) = \frac{\sigma_\varepsilon^2(-1)^k\Gamma(1-2d)}{\Gamma(k-d+1)\Gamma(1-k-d)}, \quad \text{para todo } k \in \mathbb{Z},$$

e a função de autocorrelação é dada por

$$\rho_X(k) = \frac{\Gamma(1-d)\Gamma(k+d)}{\Gamma(d)\Gamma(k-d+1)}, \text{ para todo } k \in \mathbb{Z}.$$

Quando $k \to \infty$, $\rho_X(k) \sim \frac{\Gamma(1-d)}{\Gamma(d)}k^{2d-1}$.

(e) *A função de autocorrelação parcial de* $\{X_t\}_{t\in\mathbb{Z}}$ *é dada por*

$$\phi_{kk} = \frac{d}{k-d}, \text{ para todo } k \in \mathbb{N}.$$

Demonstração:

(a) Quando $d < 0,5$, o processo $\{X_t\}_{t\in\mathbb{Z}}$ pode ser escrito na forma

$$X_t = (1-\mathcal{B})^{-d}\varepsilon_t.$$

Expandindo em séries de potências temos

$$(1-\mathcal{B})^{-d} = \sum_{k\geq 0}\binom{-d}{k}(-\mathcal{B})^k = \sum_{k\geq 0}\psi_k\mathcal{B}^k \equiv \Psi(\mathcal{B}),$$

onde

$$\psi_k = \frac{d(1+d)\cdots(k-1+d)}{\Gamma(k+1)} = \frac{\Gamma(k+d)}{\Gamma(k+1)\Gamma(d)}. \tag{2.29}$$

Assim

$$X_t = \Psi(\mathcal{B})\varepsilon_t = \sum_{k=0}^{\infty}\psi_k\varepsilon_{t-k}, \quad \forall t \in \mathbb{Z},$$

onde $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ é o processo ruído branco.

Para $d < 0,5$ e $|z| \leq 1$,

$$\sum_{j=0}^{n}\psi_j e^{-ij\cdot} \to (1-e^{-i\cdot})^{-d} = \Psi(e^{-i\cdot}), \quad \text{quando } n \to \infty, \tag{2.30}$$

onde a convergência em (2.30) é na norma $L^2(d\lambda)$, com $d\lambda$ denotando a medida de Lebesgue. Portanto, pelo Teorema 2.3, $\{X_t\}_{t\in\mathbb{Z}}$ é estacionário.

Utilizando a expressão (2.29) e a fórmula de Stirling dada por

$$\Gamma(x) \sim \sqrt{2\pi}e^{-x+1}(x-1)^{x-1/2}, \quad \text{quando } x \to \infty, \tag{2.31}$$

temos que

$$\frac{\Gamma(k+d)}{\Gamma(k+1)} \sim k^{d-1}, \quad \text{quando } k \to \infty.$$

Portanto,

$$\psi_k \sim \frac{k^{d-1}}{\Gamma(d)}, \quad \text{quando } k \to \infty.$$

(b) A demonstração para o item (b) é similar a prova do item (a), mas com $-d$ no lugar de d.

(c) Utilizando (2.13), a função densidade espectral de $\{X_t\}_{t\in\mathbb{Z}}$ é dada por

$$f_X(w) = \left|\Psi(e^{-iw})\right|^2 f_\varepsilon(w), \quad -\pi < w \leq \pi.$$

Logo, pela expressão (2.30) e pelo Teorema 2.3, segue-se que

$$f_X(w) = \left|1 - e^{-iw}\right|^{-2d} \frac{\sigma_\varepsilon^2}{2\pi} = \frac{\sigma_\varepsilon^2}{2\pi}\left(2\,\text{sen}\left(\frac{w}{2}\right)\right)^{-2d}, \quad -\pi < w \leq \pi.$$

Como $\text{sen}(w) \sim w$, quando $w \to 0$, então

$$f_X(w) \sim \frac{\sigma_\varepsilon^2}{2\pi} w^{-2d}.$$

(d) Utilizando a expressão

$$\int_0^\pi \cos(xk)\text{sen}^{\nu-1}(x)dx = \frac{\pi\cos(\frac{\pi}{2}k)\Gamma(\nu+1)2^{1-\nu}}{\nu\Gamma((\nu+k+1)/2)\Gamma((\nu-k+1)/2)} \tag{2.32}$$

(ver Gradshteyn e Ryzhik, 2000), temos que

$$\begin{aligned}
\gamma_X(k) &= \int_{-\pi}^{\pi} e^{iwk} f_X(w)dw \\
&= 2\int_0^\pi e^{iwk} f_X(w)dw = \frac{\sigma_\varepsilon^2}{\pi}\int_0^\pi \cos(wk)\left(2\,\text{sen}\left(\frac{w}{2}\right)\right)^{-2d} dw \\
&= \frac{\sigma_\varepsilon^2(-1)^k\Gamma(1-2d)}{\Gamma(k-d+1)\Gamma(1-k-d)}, \quad \text{para todo } k \in \mathbb{Z}.
\end{aligned}$$

Pelo item 3) da Observação 2.2 temos que

$$\rho_X(k) \equiv \frac{\gamma_X(k)}{\gamma_X(0)}$$
$$= \frac{\sigma_\varepsilon^2(-1)^k\Gamma(1-2d)}{\Gamma(k-d+1)\Gamma(1-d-k)}\frac{\Gamma(1-d)\Gamma(1-d)}{\sigma_\varepsilon^2\Gamma(1-2d)} = \frac{\Gamma(1-d)}{\Gamma(k-d+1)}\frac{(-1)^k\Gamma(1-d)}{\Gamma(1-d-k)}$$
$$= \frac{\Gamma(1-d)}{\Gamma(k-d+1)}\frac{(-1)^{d+k}d(d+1)(d+2)\cdots(d+(k-1))(d+k)\cdots 3\cdot 2\cdot 1}{(-1)^{d+k}(d+k)(d+k+1)\cdots 3\cdot 2\cdot 1}$$
$$= \frac{\Gamma(1-d)}{\Gamma(k-d+1)}d(d+1)(d+2)\cdots(k+d-1)$$
$$= \frac{\Gamma(1-d)}{\Gamma(k-d+1)}\frac{\Gamma(k+d)}{\Gamma(d)}, \quad \text{para todo } k \in \mathbb{Z}.$$

Utilizando a fórmula de Stirling dada pela expressão (2.31), temos que

$$\rho_X(k) \sim \frac{\Gamma(1-d)}{\Gamma(d)}k^{2d-1}, \quad \text{quando } k \to \infty.$$

(e) Para determinar a função de autocorrelação parcial escrevemos

$$\widehat{X}_{n+1} = \phi_{n1}X_n + \cdots + \phi_{nn}X_1,$$

ver Brockwell e Davis (1991), página 468.

Calculamos os coeficientes ϕ_{nj}, para $j \in \{1, 2, \cdots, n\}$, utilizando o algorítmo de Durbin-Levinson (ver Box et al., 1994) e indução matemática obtendo

$$\phi_{nj} = -\binom{n}{j}\frac{\Gamma(j-d)\Gamma(n-d-j+1)}{\Gamma(-d)\Gamma(n-d+1)}, \quad j = 1, \cdots, n. \tag{2.33}$$

Portanto, quando $n = k$ e $j = k$, temos

$$\phi_{kk} = -\frac{\Gamma(k-d)\Gamma(1-d)}{\Gamma(-d)\Gamma(k-d+1)}$$
$$= -\frac{\Gamma(k-d)(-d)\Gamma(-d)}{\Gamma(-d)(k-d)\Gamma(k-d)} = \frac{d}{k-d}, \quad \text{para todo } k \in \mathbb{N}.$$

□

Observação 2.11. Para $d > 0$, o decaimento ser do tipo hiperbólico para a função de autocorrelação, quando o valor de k aumenta, e a função densidade espectral ser ilimitada na freqüência zero demonstram a capacidade do modelo ARFIMA$(0, d, 0)$ exibir persistência.

O teorema a seguir (ver Hosking, 1981) apresenta várias propriedades dos processos $\{X_t\}_{t\in\mathbb{Z}}$ dados pela expressão (2.27).

Teorema 2.6. *Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo* ARFIMA(p,d,q) *dado pela expressão (2.27).*

(a) *Se $d < 0,5$ e todas as raízes da equação $\Phi(z) = 0$ estão fora do círculo unitário, então $\{X_t\}_{t\in\mathbb{Z}}$ é um processo estacionário.*

(b) *Se $d > -0,5$ e todas as raízes da equação $\Theta(z) = 0$ estão fora do círculo unitário, então $\{X_t\}_{t\in\mathbb{Z}}$ é inversível.*

Se $\{X_t\}_{t\in\mathbb{Z}}$ é estacionário e inversível, isto é, $d \in (-0,5; 0,5)$, com função densidade espectral $f_X(\cdot)$ e função de autocorrelação $\rho_X(\cdot)$, então

(c) $\lim_{\lambda\to 0}\lambda^{2d} f_X(\lambda)$ *existe e é finito;*

(d) $\lim_{k\to\infty} k^{1-2d}\rho_X(k)$ *existe e é finito.*

Observação 2.12.

1) Se o processo $\{X_t\}_{t\in\mathbb{Z}}$ é um processo ARFIMA(p,d,q), com $d \in (-0,5; 0,5)$, então $U_t \equiv \nabla^d X_t$ é um processo ARMA(p,q), e a recíproca também é verdadeira.

2) Se $\Theta(z) \neq 0$, para $|z| \leq 1$, então o processo $Y_t = \Phi(\mathcal{B})\Theta^{-1}(\mathcal{B})X_t$ satisfaz

$$\nabla^d(Y_t) = \varepsilon_t \quad \text{e} \quad \Phi(\mathcal{B})X_t = \Theta(\mathcal{B})Y_t,$$

onde $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ é o processo ruído branco. Então, o processo $\{Y_t\}_{t\in\mathbb{Z}}$ é um processo ARFIMA$(0,d,0)$ com função densidade espectral $f_Y(\cdot)$ dada pelo item (c) do Teorema 2.5.

3) Se $\{X_t\}_{t\in\mathbb{Z}}$ é um processo estacionário ARFIMA(p,d,q), $d \in (-0,5; 0,5)$, a função densidade espectral do processo é então dada por

$$f_X(w) = f_U(w)\left[2 \operatorname{sen}\left(\frac{w}{2}\right)\right]^{-2d}, \quad \text{para todo} \quad -\pi < w \leq \pi. \tag{2.34}$$

onde $f_U(\cdot)$ denota a função densidade espectral do processo ARMA(p,q), $\{U_t\}_{t\in\mathbb{Z}}$.

O teorema a seguir apresenta mais algumas propriedades do processo $\{X_t\}_{t\in\mathbb{Z}}$ dados pela expressão (2.27), e a demonstração pode ser encontrada em Brockwell e Davis (1991), página 469.

Teorema 2.7. *Seja* $\{X_t\}_{t\in\mathbb{Z}}$ *um processo* ARFIMA(p,d,q) *dado pela expressão (2.27). Suponha que* $d \in (-0,5; 0,5)$ *e que* $\Phi(\cdot)$ *e* $\Theta(\cdot)$*, os polinômios dados pela expressão (2.23), não têm raízes em comum.*

(a) *Se* $\Phi(z) \neq 0$*, para* $|z| = 1$*, então existe uma única solução estacionária de (2.27) dada por*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j \nabla^{-d}(\varepsilon_{t-j}), \tag{2.35}$$

onde $\Psi(z) \equiv \sum_{j=-\infty}^{\infty} \psi_j z^j = \frac{\Theta(z)}{\Phi(z)}$.

(b) *A solução* $\{X_t\}_{t\in\mathbb{Z}}$ *é causal se e somente se* $\Phi(z) \neq 0$*, para* $|z| \leq 1$.

(c) *A solução* $\{X_t\}_{t\in\mathbb{Z}}$ *é inversível se e somente se* $\Theta(z) \neq 0$*, para* $|z| \leq 1$.

(d) *Se a solução* $\{X_t\}_{t\in\mathbb{Z}}$ *é causal e inversível então, para* $d \neq 0$*, a função de autocorrelação* $\rho_X(\cdot)$ *e a função densidade espectral* $f_X(\cdot)$ *satisfazem*

$$\rho_X(k) \sim Ck^{2d-1}, \quad \text{quando } k \to \infty, \tag{2.36}$$

e

$$f_X(w) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{|\Theta(e^{-iw})|^2}{|\Phi(e^{-iw})|^2} |1 - e^{-iw}|^{-2d} \sim \frac{\sigma_\varepsilon^2}{2\pi} \left(\frac{\Theta(1)}{\Phi(1)}\right)^2 w^{-2d}, \tag{2.37}$$

onde $C > 0$, $\Phi(1) = 1 - \sum_{l=1}^{p} \phi_l$ *e* $\Theta(1) = 1 - \sum_{j=1}^{q} \theta_j$.

Observação 2.13. Observe que a função densidade espectral de um processo ARFIMA(p,d,q), dada pela expressão (2.37), tem o mesmo decaimento que a função densidade espectral dada em (2.19), quando a freqüência se aproxima de zero. Assim, relacionando os expoentes em (2.37) e (2.19), obtemos

$$d = H - \frac{1}{2}. \tag{2.38}$$

Para mais detalhes ver Beran (1994).

Capítulo 3

Estimação do Parâmetro de Longa Dependência

Neste capítulo apresentamos alguns métodos de estimação para o parâmetro fracionário d dos processos estocásticos ARFIMA(p, d, q): os métodos de regressão utilizando a função periodograma, em versão clássica e robusta; o método da máxima verossimilhança (ver Fox e Taqqu, 1986), utilizando a aproximação sugerida por Whittle (1953); o método R/S(n), proposto por Hurst (1951). Apresentamos, ainda, o método DFA ("*Detrended Fluctuation Analysis*"), proposto por Peng et al. (1994).

Os estimadores para o parâmetro de diferenciação d são classificados em três classes: paramétrica (ver Fox e Taqqu, 1986 e Sowell, 1992), semiparamétrica (ver Lopes e Mendes, 2006; Olbermann, 2002; Lopes e Nunes, 2006 e Lopes et al., 2004) e não-paramétrica (ver Lopes e Pinheiro, 2007; Olbermann et al., 2007). Neste capítulo tratamos das classes paramétrica e semiparamétrica.

Na classe paramétrica todos os parâmetros (auto-regressivos, médias móveis e de diferenciação) são estimados simultaneamente. Os métodos mais conhecidos desta classe, são os métodos propostos por Fox e Taqqu (1986) e Sowell (1992), os quais envolvem a função de máxima verossimilhança. No método proposto por Fox e Taqqu (1986), a função de máxima verossimilhança é obtida de forma aproximada enquanto que, no método proposto por Sowell (1992), ela é obtida de forma exata.

Na classe semiparamétrica, o parâmetro de diferenciação d é estimado em primeiro lugar. Os métodos desta classe consideram a estimação dos parâmetros em dois passos: apenas o parâmetro de diferenciação d é estimado no primeiro passo, e os demais parâmetros são estimados no segundo passo. O método mais popular desta classe, usualmente referido como método GPH, foi proposto por Geweke e Porter-Hudak (1983).

3.1 Métodos de Regressão Utilizando a Função Periodograma

Nesta seção apresentamos alguns estimadores para o parâmetro de diferenciação d, através do método da equação de regressão linear baseado na função periodograma.

Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo ARFIMA(p,d,q), com $d \in (-0,5; 0,5)$, dado pela expressão (2.27). Tomando o logaritmo da função densidade espectral de $\{X_t\}_{t\in\mathbb{Z}}$, dada pela expressão (2.34), temos

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

ou, escrevendo de outra maneira,

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

Substituindo w por $w_j = \frac{2\pi j}{n}$ e adicionando $\ln(I(w_j))$ em ambos os lados da expressão (3.1), onde $I(\cdot)$ é a função periodograma, dada pela expressão (2.5), obtemos

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

Considerando o limite máximo de j igual a $g(n)$, o qual é escolhido de tal forma que $\frac{g(n)}{n} \to 0$, quando $n \to \infty$, e $w_j \leqslant w_{g(n)}$, onde $w_{g(n)}$ é pequeno (neste trabalho consideramos $g(n) = [n^\beta]$, onde $\beta \in (0,1)$), o termo $\ln\left(\frac{f_U(w_j)}{f_U(0)}\right)$ é desprezível se comparado com os outros termos da equação (3.2) (ver Lopes e Mendes, 2006 e Lopes et al., 2004), obtemos então uma equação aproximada dada por

$$\ln(I(w_j)) \cong \ln(f_U(0)) - d\ln\left(4\,\mathrm{sen}^2\left(\frac{w_j}{2}\right)\right) + \ln\left(\frac{I(w_j)}{f_X(w_j)}\right). \tag{3.3}$$

Reescrevendo (3.3) na forma de uma equação de regressão linear simples

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

onde $b = -d$, $a = \ln(f_U(0))$, $y_j = \ln(I(w_j))$, $x_j = \ln\left(4\,\mathrm{sen}^2(\frac{w_j}{2})\right)$ e os erros $\epsilon_j = \ln\left(\frac{I(w_j)}{f_X(w_j)}\right)$.

Os estimadores semiparamétricos baseados no método de uma regressão linear podem ser obtidos minimizando algumas funções perda (ver Lopes e Mendes, 2006) dos resíduos

$$r_j = y_j - a - bx_j. \tag{3.5}$$

Utilizamos três tipos diferentes de função perda. Consideramos o método OLS ("*Ordinary Least Squares*"), o método LTS ("*Least Trimmed Squared*"), proposto por Rousseeuw (1984) e o método MM, proposto por Yohai (1987).

Definição 3.1 (Estimadores OLS). Os *estimadores* OLS são os valores $(\hat{a}, \hat{b})$ que minimizam a função perda dada por

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

onde r_j é dado pela expressão (3.5).

Definição 3.2 (Estimadores Robustos LTS). Os *estimadores robustos* LTS (ver Rousseeuw, 1984) são os valores $(\hat{a}, \hat{b})$ que minimizam a função perda

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

onde $(r^2)_{j:g(n)}$ são os quadrados dos resíduos ordenados, i.e., $(r^2)_{1:g(n)} \leqslant \cdots \leqslant (r^2)_{g^*(n):g(n)}$ e $g^*(n)$ é o número de valores usados no procedimento de otimização.

Definição 3.3 (Estimadores Robustos MM). Os *estimadores robustos* MM (ver Yohai, 1987) são os valores $(\hat{a}, \hat{b})$ que minimizam a função perda

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

sujeita à restrição

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

onde $\rho_2(\cdot)$ e $\rho_1(\cdot)$ são funções simétricas, limitadas, não decrescentes em $[0, \infty)$ com $\rho_j(0) = 0$ e $\lim_{u\to\infty} \rho_j(u) = 1$, para $j = 1, 2$; s é o parâmetro escala e C é uma constante ("*tuning constant*").

3.1.1 Estimadores GPH, GPH-LTS e GPH-MM

O primeiro método de estimação baseado na função periodograma dada por (2.5), foi proposto por Geweke e Porter-Hudak (1983). Para obter um estimador para d, estes autores aplicam o método OLS em (3.6) baseado na equação de regressão linear simples dada pela expressão (3.4), denotado aqui por GPH. Assim

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

onde

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

A variância do estimador GPH (ver Geweke e Porter-Hudak, 1983) é dada por

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

Para obter as versões robustas deste estimador, denotados por GPH-LTS e GPH-MM, aplicamos a metodologia LTS e MM, dadas pelas expressões (3.7) e (3.8), respectivamente, ao modelo de regressão linear dado por (3.4).

3.1.2 Estimadores R, R-LTS e R-MM

O estimador de regressão linear proposto por Robinson (1995), o qual denotamos por R, é obtido aplicando-se o método OLS em (3.6) baseado na equação de regressão linear dada pela expressão (3.4), mas considerando apenas as freqüências ω_j para $j \in \{l, l+1, \cdots, g(n)\}$, onde $l > 1$ é o valor do corte que tende para o infinito de forma mais lenta do que $g(n)$, quando $n \to \infty$ e ainda $g(n) \to \infty$, quando $n \to \infty$.

A variância assintótica do estimador R (ver Robinson, 1995) é dada por

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

Para obter as versões robustas deste estimador, denotados por R-LTS e R-MM, aplicamos a metodologia LTS e MM, dadas pelas expressões (3.7) e (3.8), respectivamente, no modelo de regressão linear dado por (3.4).

3.2 Método da Máxima Verossimilhança

Nesta seção apresentamos o estimador de máxima verossimilhança sugerido por Fox e Taqqu (1986), que será denotado por W pois ele é baseado em uma aproximação da função de verossimilhança sugerida por Whittle (1953). O estimador W, sob certas condições de regularidade, é consistente e tem distribuição assintótica normal. Além disso, é estimador não viciado e de variância mínima.

Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estacionário com média μ e variância σ_X^2. Seja $f_X(\cdot) = f_X(\cdot,\eta)$ a função densidade espectral caracterizada pelo vetor de parâmetros (desconhecidos) dado por

$$\eta = (\sigma_X^2, d, \phi_1, \phi_2, \cdots, \phi_p, \theta_1, \theta_2, \cdots, \theta_q). \tag{3.11}$$

Suponha que $\{X_t\}_{t\in\mathbb{Z}}$ é um processo Gaussiano. A estimação dos parâmetros no vetor η é obtida através da série temporal $\{X_t\}_{t=1}^n$, cuja função de distribuição conjunta, ou função de verossimilhança, é dada por

$$\mathrm{L}(x,\eta) = (2\pi)^{-\frac{n}{2}}|\Sigma_n(\eta)|^{-\frac{1}{2}}e^{-\frac{1}{2}x^t[\Sigma_n(\eta)]^{-1}x}, \tag{3.12}$$

onde $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$, x^t denota o vetor transposto do vetor x e $\Sigma_n(\eta)$ é a matriz quadrada $n \times n$, dada por

$$\Sigma_n(\eta) = [\gamma_X(k)]_{k=0}^n, \tag{3.13}$$

onde $\gamma_X(\cdot)$ é a *função de autocovariância* do processo $\{X_t\}_{t\in\mathbb{Z}}$, dada pela expressão (2.2).

O logaritmo da função de verossimilhança em (3.12) é dado por

$$\mathcal{L}_n(x;\eta) = \ln\big(\mathrm{L}(x;\eta)\big) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_n(\eta)|) - \frac{1}{2}x^t[\Sigma_n(\eta)]^{-1}x. \tag{3.14}$$

Assim obtemos o vetor m-dimensional

$$\begin{aligned}\mathcal{L}_n'(x;\eta) &= \Big(\frac{\partial}{\partial\eta_j}\mathcal{L}_n(x;\eta)\Big)_{j=1}^m \\ &= \Big(-\frac{1}{2}\frac{\partial}{\partial\eta_j}\ln|\Sigma_n(\eta)| - \frac{1}{2}x^t\Big[\frac{\partial}{\partial\eta_j}[\Sigma_n(\eta)]^{-1}\Big]x\Big)_{j=1}^m,\end{aligned} \tag{3.15}$$

onde $m = p + q + 2$. O estimador de máxima verossimilhança de η é obtido maximizando (3.14) com respeito ao vetor η. Isto pode ser feito (com algumas condições de suavidade) resolvendo o sistema de m equações

$$\mathcal{L}_n'(x;\widehat{\eta}) = 0, \tag{3.16}$$

dado pela expressão (3.15), onde $\widehat{\eta}$ é o estimador de máxima verossimilhança (EMV) de η.

Computacionalmente, é complicado obter a inversa da matriz de autocovariâncias $\Sigma_n(\eta)$. Whittle (1953) estabelece a forma aproximada para $[\Sigma_n(\eta)]^{-1}$. Fox e Taqqu (1986), fazendo uso desta aproximação, aplicam o método da máxima verossimilhança aproximada para estimar η maximizando a função

$$\mathrm{L}(x;\eta) = \left(\frac{1}{\sqrt{2\pi}\sigma_X}\right)^n \exp\left\{ -\frac{Z^t A_n(\eta) Z}{2n\sigma_X}\right\}, \tag{3.17}$$

ou seja, minimizando

$$\frac{Z^t A_n(\eta) Z}{n}, \tag{3.18}$$

onde $Z = (X_1 - \overline{X}, \cdots, X_n - \overline{X})$, $\overline{X}$ é a média amostral e

$$A_n(\eta) = [\alpha(k)]_{k=1}^n \tag{3.19}$$

é a matriz $n \times n$ proposta por Whittle (1953) para aproximar $\Sigma_n(\eta)$ com

$$\alpha(k) = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \frac{1}{f(w;\eta)} e^{iwk} dw. \tag{3.20}$$

Computacionalmente, o estimador W é obtido minimizando a forma discreta

$$\mathcal{L}_n(\eta) = \frac{1}{2\pi} \sum_{j=1}^{[\frac{m-1}{2}]} \left(\ln\left(f_X(w_j,\eta)\right) + \frac{I(w_j)}{f_X(w_j,\eta)}\right), \tag{3.21}$$

onde η é o vetor de parâmetros desconhecidos dado em (3.11), $[x]$ é a parte inteira de x e $w_j = \frac{2\pi j}{n}$, $j \in \{1, \cdots, [\frac{m-1}{2}]\}$ são as freqüências de Fourier.

A variância assintótica para o estimador W (ver Fox e Taqqu, 1986) é dada por

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

Ver Fox e Taqqu (1986) e Beran (1994), para um estudo mais completo do estimador W.

3.3 Estimador R/S(n)

A estatística R/S(n) foi introduzida por Hurst (1951) com o nome "*Rescaled Range*" (ou "*Range Over Standard Deviation*"), com o propósito de testar a existência de memória longa em uma série temporal. O método R/S(n) é um dos mais conhecidos para estimar o parâmetro *H de Hurst* (ver Observação 2.9). Foi mostrado que é viciado (ver Taqqu et al., 1995) e que

é afetado pela potencial presença de não estacionariedade nos dados (ver Bhattacharya et al., 1983).

Apresentamos, nesta seção, a *estatística* R/S(n) *clássica* e suas versões modificadas propostas por Lo (1991), Kwiatkowski et al. (1992) e Giraitis et al. (2003), respectivamente.

Definição 3.4 (Estatística R/S(n)). Dada uma série temporal $\{X_t\}_{t=1}^n$, *a estatística* R/S*(n) clássica* é definida por

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

onde $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ e $s_n = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2}$ é o *desvio padrão amostral*.

Para o processo ruído Gaussiano fracionário ou para o processo ARFIMA (ver Teverovsky et al., 1998),

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

onde C_{H} é uma constante positiva que não depende de n.

O método R/S(n) consiste em estimar o parâmetro de longa dependência H *de Hurst*. Para determinar H, utilizando a *estatística* R/S(n), seguimos os seguintes passos:

- Para cada $j \in \{1, \cdots, s\}$, divide-se a série temporal $\{X_t\}_{t=1}^n$ em $\left[\frac{n}{k_j}\right]$ blocos, onde k_j é o número de observações em cada bloco, é suficientemente grande e $k_j = \ell k_{j-1}$.
- Para cada bloco, computa-se a estatística R/S(k_j).
- Ajusta-se uma reta por mínimos quadrados, que relaciona $\ln(\mathrm{R/S}(k_j))$ e $\ln(k_j)$, $j = 1, \cdots, s$, obtendo o expoente H *de Hurst* que é o coeficiente de inclinação da reta ajustada.

A Figura 3.1 ilustra a aplicação do método R/S(n) na seqüência LAMCG com $n = 48.502$ pares de bases, vista no Exemplo 4.1.

Relação entre o valor de H e o comportamento do processo:

- Para processos de curta dependência, o gráfico de $\ln(\mathrm{R/S}(k_j))$ versus $\ln(k_j)$, mostra, para k_j suficientemente grande, os pontos dispostos aleatoriamente em torno de uma reta com inclinação $\mathrm{H} = 0,5$.

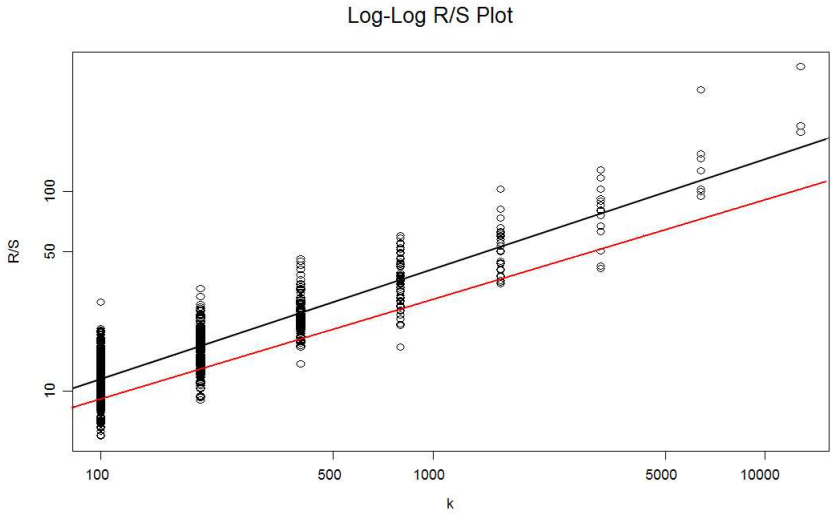

Figura 3.1: Gráfico $\ln(k_j)$ versus $\ln(R/S(k_j))$ para a seqüência LAMCG. A linha preta representa a reta ajustada com $\mathrm{H} = 0,5513$ e a linha vermelha representa o caso de não dependência quando $\mathrm{H} = 0,5$.

- Para processos de longa dependência, os pontos do gráfico $\ln(\mathrm{R/S}(k_j))$ versus $\ln(k_j)$ estão dispostos aleatoriamente em torno de uma reta com inclinação $\mathrm{H} > 0,5$, para k_j suficientemente grande.

Observação 3.1. *A estatística* $\mathrm{R/S}(n)$ apresenta diversas deficiências, como por exemplo, qual valor de k_j utilizar para desenhar a reta de melhor comportamento? Apesar de suas deficiências, plotar o $\mathrm{R/S}(n)$ é útil para se ter uma primeira idéia sobre o comportamento da dependência entre observações distintas dos dados.

A seguir definimos o estimador HAC ("*Heteroskedasticity and Autocorrelation Consistent*") da variância (ver Newey e West, 1987).

Definição 3.5 (Estimador HAC da Variância). O *estimador HAC* da variância com número de freqüências ("*bandwidth*") q, é definido por

$$\widehat{\sigma}_n^2(q) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2 + \frac{2}{n}\sum_{j=1}^{q}\omega_j(q)\left(\sum_{l=j+1}^{n}(X_l - \overline{X})(X_{l-j} - \overline{X})\right), \quad (3.22)$$

onde $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ e os pesos $\omega_j(q)$ são dados por

$$\omega_j(q) = 1 - \frac{j}{q+1}, \quad \text{para todo } q < n.$$

Observação 3.2. Não existe uma regra específica para a escolha da ordem q, mas q deverá satisfazer

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

Uma escolha razoável é $q = n^{0,5}$ (ver Giraitis et al., 2003).

O método $R/S(n)$ detecta memória curta sem diferenciá-la da memória de longo prazo. Lo (1991) propõe uma estatística modificada para solucionar este problema.

Definição 3.6 (Estatística R/S(q)). *A estatística* R/S(n) *modificada*, proposta por Lo (1991) e denotada por R/S(q), é definida por

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

onde $\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$ e $\widehat{\sigma}_n(q)$ é obtido pela raíz quadrada da expressão (3.22).

Observação 3.3. Se o processo estocástico não possui longa dependência e o conjunto de índices T é $\mathbb{R}^+$, Lo (1991) mostra que a partir de uma escolha certa para q, a distribuição de R/S(q) é assintótica para

$$W_1 = \max_{0 \le t \le 1} W_t^0 - \min_{0 \le t \le 1} W_t^0,$$

onde W_t^0 é dado pela Definição 2.19.

Definição 3.7 (Estatística KPSS(q)). *A estatística* R/S(n) *modificada*, proposta por Kwiatkowski et al. (1992), e denotada por KPSS(q), é definida por

$$\text{KPSS}(q) = \frac{1}{n^2 \widehat{\sigma}_n^2(q)} \left[\sum_{k=1}^{n} \left(\sum_{j=1}^{k} (X_j - \overline{X}) \right)^2 \right],$$

onde $\overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$ e $\widehat{\sigma}_n^2(q)$ é definida pela expressão (3.22).

Observação 3.4. Sob a hipótese nula de curta dependência (ver Giraitis et al., 2003), a estatística KPSS(q) tem convergência assintótica para

$$\int_0^1 (W_t^0)^2 dt,$$

onde W_t^0 é dado pela Definição 2.19.

Definição 3.8 (Estatística V/S(q)). *A estatística* R/S(n) *modificada*, proposta por Giraitis et al. (2003) e denotada por $V/S(q)$, é definida por

$$\mathrm{V/S}(q) = \frac{1}{n^2\widehat{\sigma}_n^2(q)}\left[\left(\sum_{k=1}^{n}\left(\sum_{j=1}^{k}(X_j - \overline{X})\right)^2\right) - \frac{1}{n}\left(\sum_{k=1}^{n}\sum_{j=1}^{k}(X_j - \overline{X})\right)^2\right],$$

onde $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ e $\widehat{\sigma}_n^2(q)$ é definida pela expressão (3.22).

Observação 3.5. Sob a hipótese nula de curta dependência (ver Giraitis et al., 2003), a estatística V/S(q) tem convergência assintótica para

$$\int_0^1 (W_t^0)^2 dt - \left(\int_0^1 W_t^0 dt\right)^2,$$

onde W_t^0 é dado pela Definição 2.19.

3.4 Método das Análises de Flutuações Destendenciadas (DFA)

Dada uma série temporal $\{X_t\}_{t=1}^n$, o método das análises de flutuações destendenciadas ("*Detrended Fluctuation Analysis*" - DFA), proposto por Peng et al. (1994), consiste de cinco passos. Primeiro, para cada $t \in \{1, 2, \cdots, n\}$, calcula-se

$$Y_t = \sum_{j=1}^{t} X_j. \tag{3.23}$$

Observe que o processo estocástico $\{Y_t\}_{t\in\mathbb{Z}}$ é não estacionário. No segundo passo, divide-se a série temporal $\{Y_t\}_{t=1}^n$ em $\left[\frac{n}{l}\right]$ blocos não sobrepostos, onde cada bloco contém exatamente l observações. No terceiro passo, para cada bloco ajusta-se, pelo método dos mínimos quadrados, uma reta aos dados (que representa a tendência no bloco).
No quarto passo, destendencia-se a série temporal $\{Y_t\}_{t=1}^n$, ou seja, em cada bloco calcula-se

$$Z_t = Y_t - Y_t^l, \tag{3.24}$$

onde Y_t^l denota a ordenada y dos segmentos de reta ajustados em cada bloco.

Exemplo 3.1. Para ilustrar o método DFA, mostramos, na Figura 3.2, a aplicação deste método com blocos de $l = 100$ observações, da seqüência *Enterobacteria phage lambda* do Exemplo 4.1, mas apenas considerando-se os 1.000 nucleotídeos iniciais.

Por fim, para cada $l \in \{4, 5, \cdots, g(n)\}$ calcula-se *a raíz da flutuação média quadrática* (ver Definição 3.9).

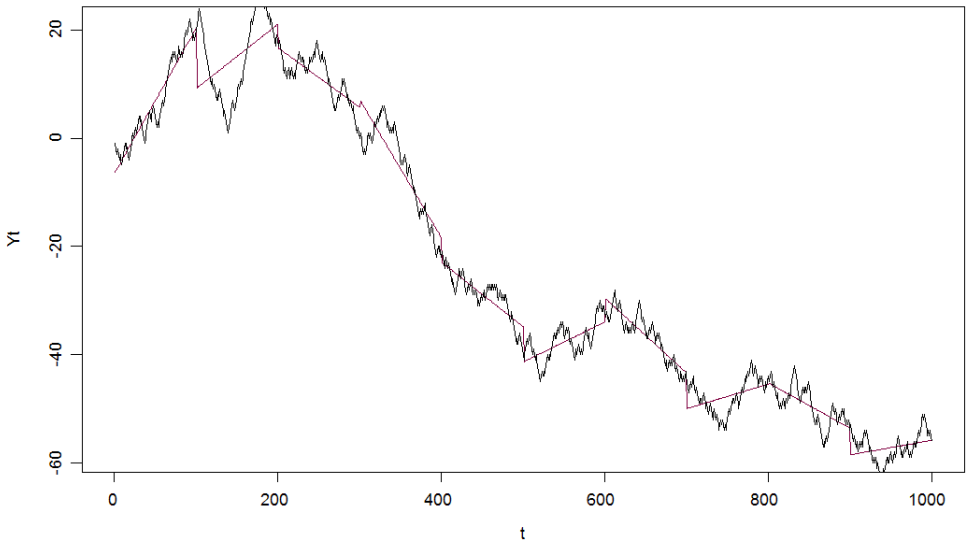

Figura 3.2: Aplicação do Método DFA na Seqüência LAMCG: foram utilizados 1.000 nucleotídeos iniciais e blocos com $l = 100$ observações.

Definição 3.9 (Raíz da Flutuação Média Quadrática). A *raíz da flutuação média quadrática* é definida por

$$\mathrm{F}(l) = \sqrt{\frac{1}{\widetilde{n}} \sum_{t=1}^{\widetilde{n}} Z_t^2}, \tag{3.25}$$

onde Z_t é dado pela expressão (3.24) e $\widetilde{n}$ é o maior múltiplo de l, inferior ou igual a n, isto é, $\widetilde{n} = [\mathrm{M} \cdot l] \leq n$.

Observação 3.6. Na literatura não existe uma regra específica para a escolha ótima de $g(n)$. Neste trabalho utilizamos $g(n) = [(0,03) \cdot n]$.

Observe que $\mathrm{F}(l)$, dada pela expressão (3.25), aumenta quando l cresce. Uma relação linear em um gráfico ln versus ln indica a presença de escala

$$\mathrm{F}(l) \sim \varphi l^{\alpha}. \tag{3.26}$$

Sob tais condições, as flutuações podem ser caracterizadas pelo expoente α onde

- $0 < \alpha < 0,5$ indica dependência intermediária.
- $\alpha = 0,5$ indica curta correlação ou curta dependência.

- $0,5 < \alpha < 1$ indica a presença de longa dependência.

Aplicando logaritmo em ambos os lados da expressão (3.26) obtemos

$$\ln(\mathrm{F}(l)) \sim \ln(\varphi) + \alpha \ln(l). \tag{3.27}$$

A equação (3.27) acima é da forma de uma equação de regressão linear simples, dada pela expressão (3.4), onde

$$y_j = \ln(\mathrm{F}(l)), \quad a = \ln(\varphi), \quad b = \alpha, \quad x_j = \ln(l), \quad l = j + 3, \tag{3.28}$$

com $l \in \{4, 5, \cdots, g(n)\}$ e $m = [g(n) - 3]$.

Na expressão (3.28), temos que $\alpha = b$. Logo, para estimar o *expoente* α, basta estimar b. Segue-se então que o estimador de α, pelo método dos mínimos quadrados da regressão linear de $y_1, \cdots, y_m$ em $x_1, \cdots, x_m$, é dado por

$$\begin{aligned} \widehat{\alpha} &= \frac{\displaystyle\sum_{j=1}^{m}(x_j - \overline{x})y_j}{\displaystyle\sum_{j=1}^{m}(x_j - \overline{x})^2} = \frac{\frac{1}{m}\displaystyle\sum_{j=1}^{m} x_j - \frac{\overline{x}}{m}\sum_{j=1}^{m} y_j}{\frac{1}{m}\displaystyle\sum_{j=1}^{m}(x_j - \overline{x})^2} \\ &= \frac{\overline{x} - \overline{xy}}{\frac{1}{m}\displaystyle\sum_{j=1}^{m}(x_j - \overline{x})^2} = \frac{\overline{x}(1 - \overline{y})}{\frac{1}{m}\displaystyle\sum_{j=1}^{m}(x_j - \overline{x})^2}, \end{aligned} \tag{3.29}$$

onde $y_j = \ln(\mathrm{F}(j+3))$, $x_j = \ln(j+3)$, $\overline{x} = \dfrac{1}{m}\displaystyle\sum_{j=1}^{m} x_j$ e $m = [g(n) - 3]$.

3.4.1 Propriedades Estatísticas do Método DFA

Definição 3.10 (Modelo Linear Geral). As variáveis dependentes $y_1, y_2, \cdots, y_m$ satisfazem o *modelo linear geral*, se podem ser expressas por

$$y_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_v x_{jv} + \epsilon_j, \quad j = 1, \cdots, m, \tag{3.30}$$

onde $x_{j\ell}$, para $1 \leq \ell \leq v$, são constantes conhecidas, β_ℓ, para $1 \leq \ell \leq v$, são parâmetros desconhecidos e ϵ_j são variáveis aleatórias independentes e identicamente distribuídas com função de distribuição $\mathcal{N}(0, \sigma^2)$.

Observação 3.7. Assumindo que no *modelo de regressão linear*, dado pela expressão (3.4), as variáveis aleatórias $\epsilon_j's$ são independentes e identicamente distribuídas com função de distribuição $\mathcal{N}(0, \sigma^2)$, obtemos o caso em que $v = 2$ na Definição 3.10.

A demonstração da Proposição 3.1, que apresentamos abaixo, pode ser encontrada em Bickel e Doksum (1977), página 265.

Proposição 3.1. *Sob o modelo linear geral com $v = r$, os estimadores de mínimos quadrados $\widehat{\beta}_1, \cdots, \widehat{\beta}_r$ são estimadores* U.M.V.U. *("Uniformly Minimum Variance Unbiased") de $\beta_1, \cdots, \beta_r$, dados na expressão (3.30). Mais ainda, qualquer função $\sum_{j=1}^{r} d_j\beta_j$ é estimada por $\sum_{j=1}^{r} d_j\widehat{\beta}_j$, que também é* U.M.V.U..

Teorema 3.1. *Se as variáveis aleatórias $\epsilon_j's$, que aparecem no modelo de regressão linear dado pela expressão (3.4), são variáveis aleatórias independentes e identicamente distribuídas com função de distribuição $\mathcal{N}(0, \sigma^2)$, então $\widehat{\alpha}$, dado pela expressão (3.29), é um estimador* U.M.V.U..

Demonstração: Vimos, pela expressão (3.29), que

$$\widehat{\alpha} = \frac{\sum_{j=1}^{m}(x_i - \overline{x})y_j}{\sum_{j=1}^{m}(x_j - \overline{x})^2}$$

é estimador de α e pela Observação 3.7 temos que $v = r = 2$, onde $m = [g(n) - 3]$. Logo, pela Proposição 3.1, este estimador é U.M.V.U.. Portanto, $\widehat{\alpha}$ é um estimador U.M.V.U. de α. □

Teorema 3.2. *Se as variáveis aleatórias $\epsilon_j's$, que aparecem no modelo de regressão linear dado pela expressão (3.4), são variáveis aleatórias independentes e identicamente distribuídas com função de distribuição $\mathcal{N}(0, \sigma^2)$, então $\widehat{\alpha}$, dado pela expressão (3.29), é um estimador consistente.*

Demonstração: Todo estimador U.M.V.U. é consistente (ver Bickel e Doksum, 1977). Logo, pelo Teorema 3.1, obtemos que $\widehat{\alpha}$ é um estimador consistente para α.

Observação 3.8.

1) Se as variáveis aleatórias $\epsilon_j's$, no modelo de regressão linear dado pela expressão (3.4), são independentes e identicamente distribuídas com função de distribuição $\mathcal{N}(0, \sigma^2)$, então pela Proposição 3.1 temos que $\widehat{\alpha}$ é um estimador U.M.V.U.. Portanto, $\widehat{\alpha}$ é estimador não viciado para α.

2) Se as variáveis aleatórias $\epsilon_j's$, no modelo de regressão linear dado pela expressão (3.4), são independentes e identicamente distribuídas com função

de distribuição $\mathcal{N}(0, \sigma^2)$, então a variância de $\widehat{\alpha}$ é dada por

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

com $m = [g(n) - 3]$.

Vimos que, para aplicar o método DFA, primeiro divide-se a série temporal em blocos com l observações. Em cada bloco, calcula-se as somas parciais $\{Y_t\}_{t=1}^{l}$, ajusta-se uma reta $Y_t^l = a + bt$ para estas somas parciais e então calcula-se a soma $\sum_{t=1}^{l}(Y_t - Y_t^l)^2$, para cada l observações. O Teorema 3.3 abaixo fornece uma aproximação para a esperança desta respectiva soma em cada bloco. O objetivo é que a esperança da variância amostral seja aproximadamente proporcional à l^{2H} (ver Teorema 3.4).

Teorema 3.3. (Taqqu et al., 1995). *Seja $\{X_t\}_{t\in\mathbb{R}^+}$ um processo ruído Gaussiano fracionário. Considere $\{X_t\}_{t=1}^{n}$ uma série temporal advinda deste processo. Então,*

$$\mathbb{E}\left(\sum_{t=1}^{l}(Y_t - Y_t^l)^2\right) \sim C_{\mathrm{H}}\, l^{2\mathrm{H}+1}, \quad \textit{quando } l \to \infty, \tag{3.31}$$

onde $Y_t = \displaystyle\sum_{j=1}^{t} X_j$ *e*

$$C_{\mathrm{H}} = \left(\frac{2}{2\mathrm{H}+1} + \frac{1}{\mathrm{H}+2} - \frac{2}{\mathrm{H}+1}\right). \tag{3.32}$$

Teorema 3.4. *Seja $\{X_t\}_{t\in\mathbb{R}^+}$ um processo ruído Gaussiano fracionário. Considere $\{X_t\}_{t=1}^{n}$ uma série temporal advinda deste processo. Então,*

$$\mathbb{E}(\mathrm{F}^2(l)) \sim C_{\mathrm{H}}\, l^{2\mathrm{H}}, \quad \textit{quando } l \to \infty, \tag{3.33}$$

onde $\mathrm{F}^2(l)$ é a flutuação média quadrática obtida pela expressão (3.25) e C_{H} é dada pela expressão (3.32).

Demonstração: Observe que

$$
\begin{aligned}
\mathbb{E}(\mathrm{F}^2(l)) &= \frac{1}{\widetilde{n}}\mathbb{E}\Big(\sum_{t=1}^{\widetilde{n}} Z_t^2\Big) = \frac{1}{\widetilde{n}}\mathbb{E}\Big(\sum_{t=1}^{\widetilde{n}}(Y_t - Y_t^l)^2\Big) \\
&= \frac{1}{\widetilde{n}}\mathbb{E}\Big(\sum_{t=1}^{l}(Y_t - Y_t^l)^2 + \sum_{t=l+1}^{2l}(Y_t - Y_t^l)^2 + \cdots + \\
&+ \sum_{t=[(n/l)-1]l+1}^{\widetilde{n}}(Y_t - Y_t^l)^2\Big) \\
&= \frac{1}{\widetilde{n}}\Big[\mathbb{E}\Big(\sum_{t=1}^{l}(Y_t - Y_t^l)^2\Big) + \mathbb{E}\Big(\sum_{t=l+1}^{2l}(Y_t - Y_t^l)^2\Big) + \cdots + \\
&+ \mathbb{E}\Big(\sum_{t=[(n/l)-1]l+1}^{\widetilde{n}}(Y_t - Y_t^l)^2\Big)\Big]. \qquad (3.34)
\end{aligned}
$$

Logo, pelo Teorema 3.3 e pela expressão (3.34) obtemos que

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

onde $\mathrm{F}^2(l)$ é a flutuação média quadrática dada pela expressão (3.25) e C_{H} é dada pela expressão (3.32).

□

Observação 3.9. Pela expressão (3.26) obtemos

$$
\mathbb{E}(F^2(l)) \sim \varphi^2 l^{2\alpha}. \qquad (3.35)
$$

Comparando as expressões (3.35) e (3.33), segue-se que $\alpha = \mathrm{H}$. Utilizando a expressão (2.38) obtemos então a seguinte relação

$$
\alpha = \mathrm{H} = d + \frac{1}{2}. \qquad (3.36)
$$

Teorema 3.5. *Suponha que as variáveis aleatórias $Z_1, Z_2, \cdots, Z_{\widetilde{n}}$, dadas pela expressão (3.24), são independentes e identicamente distribuídas com função de distribuição comum $\mathcal{N}(0, \sigma_l^2)$. Então, $\mathrm{F}^2(l)$, dada pela expressão (3.25), tem função de distribuição $\Gamma\big(\frac{\widetilde{n}}{2}, \frac{\widetilde{n}}{2\sigma_l^2}\big)$.*

Demonstração: Para todo $j \in \{1, 2, \cdots, \widetilde{n}\}$, a variável aleatória $\frac{Z_j}{\sigma_l}$ é a padronização de Z_j. Por hipótese, as variáveis aleatórias $Z_1, Z_2, \cdots, Z_{\widetilde{n}}$, são independentes e identicamente distribuídas com função de distribuição comum $\mathcal{N}(0, \sigma_l^2)$. Então, para cada $j \in \{1, 2, \cdots, \widetilde{n}\}$, a variável aleatória

$\frac{Z_j}{\sigma_l}$ tem função de distribuição $\mathcal{N}(0,1)$. Portanto, a variável aleatória $\displaystyle\sum_{j=1}^{\widetilde{n}} \frac{Z_j^2}{\sigma_l^2}$ tem função de distribuição $\chi^2(\widetilde{n}) = \Gamma(\frac{\widetilde{n}}{2}, \frac{1}{2})$, onde $\widetilde{n} = [M \cdot l] \leq n$.

Denotando $X \equiv \displaystyle\sum_{j=1}^{\widetilde{n}} \frac{Z_j^2}{\sigma_l^2}$, $Y \equiv \left(\frac{\sigma_l^2}{\widetilde{n}}\right) X$ e utilizando a expressão (3.25) obtemos

$$\mathrm{F}^2(l) = \frac{1}{\widetilde{n}} \sum_{j=1}^{\widetilde{n}} Z_j^2 = \frac{\sigma_l^2}{\widetilde{n}} \sum_{j=1}^{\widetilde{n}} \frac{Z_j^2}{\sigma_l^2} = \left(\frac{\sigma_l^2}{\widetilde{n}}\right) X = Y. \tag{3.37}$$

Sabemos que a função característica de uma variável aleatória qualquer determina a sua função de distribuição. A função característica da variável aleatória Y é dada por

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

$$= \left[\frac{1}{\frac{\widetilde{n} - 2it\sigma_l^2}{\widetilde{n}}}\right]^{\frac{\widetilde{n}}{2}} = \left[\frac{\frac{\widetilde{n}}{2\sigma_l^2}}{\frac{\widetilde{n}}{2\sigma_l^2} - it}\right]^{\frac{\widetilde{n}}{2}}, \quad \text{para todo } t < \frac{\widetilde{n}}{2\sigma_l^2}, \tag{3.38}$$

pois a variável aleatória X tem distribuição $\Gamma(\frac{\widetilde{n}}{2}, \frac{1}{2})$.

Observe que a função característica resultante na expressão (3.38) é a de uma variável aleatória com função de distribuição $\Gamma\big(\frac{\widetilde{n}}{2}, \frac{\widetilde{n}}{2\sigma_l^2}\big)$. Pela unicidade da função característica segue-se que Y tem função de distribuição $\Gamma\big(\frac{\widetilde{n}}{2}, \frac{\widetilde{n}}{2\sigma_l^2}\big)$, ou seja, $F^2(l)$ dada pela expressão (3.25), tem função de distribuição $\Gamma\big(\frac{\widetilde{n}}{2}, \frac{\widetilde{n}}{2\sigma_l^2}\big)$. □

Corolário 3.1. *Suponha que as variáveis aleatórias $Z_1, Z_2, \cdots, Z_{\widetilde{n}}$, dadas pela expressão (3.24), são independentes e identicamente distribuídas com função de distribuição comum $\mathcal{N}(0, \sigma_l^2)$. Então, $F^2(l)$, dada pela expressão (3.25), tem esperança e variância dadas, respectivamente, por*

$$\mathbb{E}(\mathrm{F}^2(l)) = \sigma_l^2 \quad e \quad \mathrm{Var}(\mathrm{F}^2(l)) = \frac{2\sigma_l^4}{\widetilde{n}}, \tag{3.39}$$

sempre que $0 < {\sigma_l}^4 < \infty$.

Demonstração: Pelo Teorema 3.5, temos que $\mathrm{F}^2(l)$, dada pela expressão (3.25), tem função de distribuição $\Gamma\big(\frac{\widetilde{n}}{2}, \frac{\widetilde{n}}{2\sigma_l^2}\big)$. Logo,

$$\mathbb{E}(F^2(l)) = \frac{\frac{\widetilde{n}}{2}}{\frac{\widetilde{n}}{2\sigma_l^2}} = \left(\frac{\widetilde{n}}{2}\right)\left(\frac{2\sigma_l^2}{\widetilde{n}}\right) = \sigma_l^2$$

e variância

$$\mathrm{Var}(\mathrm{F}^2(l)) = \frac{\frac{\widetilde{n}}{2}}{\left(\frac{\widetilde{n}}{2\sigma_l^2}\right)^2} = \left(\frac{\widetilde{n}}{2}\right)\left(\frac{4\sigma_l^4}{\widetilde{n}^2}\right) = \frac{2\sigma_l^4}{\widetilde{n}},$$

quando $0 < \sigma_l^4 < \infty$. □

Vimos, nesta seção, que podemos estimar o parâmetro de diferenciação d utilizando o método DFA (ver expressão (3.36)). Denotamos por DFA o estimador para o parâmetro de diferenciação d obtido pelo método DFA. Demonstramos nesta seção que se a equação (3.27) é da forma de uma equação de regressão linear simples dada pela expressão (3.4), onde

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

com $l \in \{4, 5, \cdots, g(n)\}$ e se as variáveis aleatórias $\epsilon_j's$ da expressão (3.4), são independentes e identicamente distribuídas com função de distribuição $\mathcal{N}(0, \sigma^2)$, então o estimador $\widehat{\alpha}$ é não viciado e consistente. Mostramos ainda, pela expressão (3.39) que, se as variáveis aleatórias $Z_1, Z_2, \cdots, Z_{\widetilde{n}}$, dadas pela expressão (3.24), são independentes e identicamente distribuídas com função de distribuição comum $\mathcal{N}(0, \sigma_l^2)$, então $F^2(l)$ é um estimador não viciado para a variância σ_l^2 e se $0 < \sigma_l^4 < \infty$, é consistente, quando $\widetilde{n} \to \infty$.

Capítulo 4

Conceitos de Biologia Molecular

Nosso objetivo, é estudar o parâmetro de *longa dependência* em diversas seqüências de DNA. Para isso, apresentamos neste capítulo uma breve introdução à *molécula de DNA*, a seguir abordamos aspectos da estrutura do DNA. Descrevemos também aqui o que é uma *seqüência* de DNA e expomos algumas de suas representações numéricas. Consideramos aqui funções que transformam *seqüências de DNA* em seqüências numéricas. Por fim, definimos a série temporal que utilizamos neste trabalho.

A partir da década de 40 do século XX, vários pesquisadores definiram algumas propriedades do DNA, tais como:

- A *molécula* de DNA, sendo uma substância orgânica, é formada por partículas menores denominadas *nucleotídeos*.
- Está relacionada à hereditariedade.
- Seu formato deve ser um fio em forma de dupla hélice.
- O açúcar do DNA é a pentose dessoxirribose.
- As bases nitrogenadas do DNA são *adenina, guanina, citosina* e *timina*.

Com base nestas informações, o americano James D. Watson e o inglês Francis H. C. Crick iniciaram um estudo com a finalidade de criar um modelo para a molécula de DNA. Em 1953 propuseram uma estrutura que ficou conhecida como *modelo de Watson e Crick* (que lhes valeu o Prêmio Nobel de Fisiologia e Medicina de 1962). Segundo o modelo proposto por Watson e Crick, a *molécula de DNA* é uma dupla-hélice, semelhante a uma escada espiral, formada por duas cadeias de *nucleotídeos*, lembrando duas fitas enroladas uma à outra. Cada corrimão da escada é constituído por uma cadeia que se sucedem, alternadamente, a desoxirribose de um *nucleotídeo* e o grupo

fosfato do seguinte. Cada degrau é um par de bases nitrogenadas, uma de cada cadeia, ligadas entre si por pontes de hidrogênio, sempre estabelecidas entre uma *adenina* e uma *timina* ou entre uma *guanina* e uma *citosina* (ver Figura 4.1).

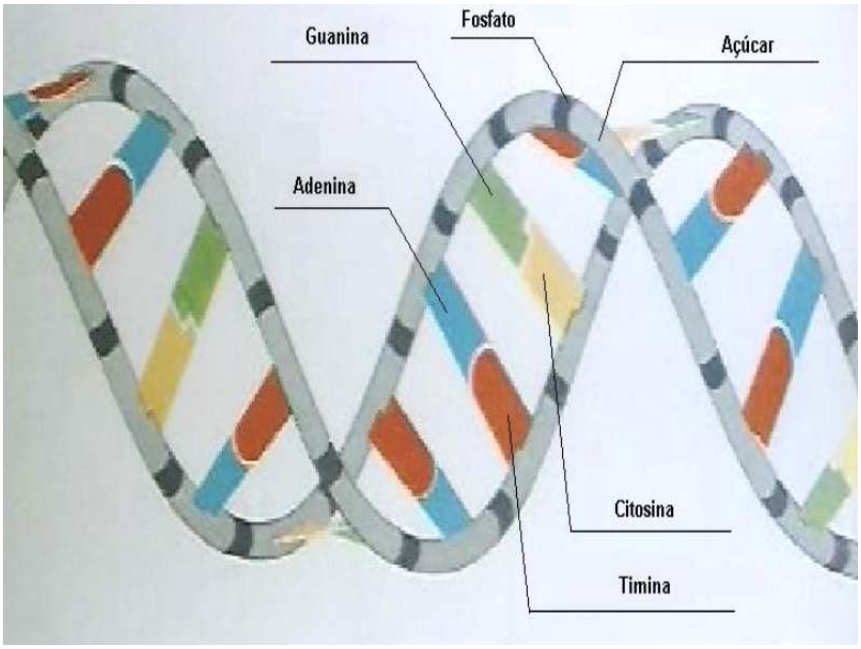

Figura 4.1: Seqüência de DNA. Adenina na cor azul; Citosina na cor amarela; Timina na cor vermelha e Guanina na cor verde.

4.1 Molécula DNA

Definição 4.1 (DNA). O *DNA* é uma molécula orgânica que reproduz o código genético. Quando transcrita em RNA, tem a capacidade de traduzir proteínas. É responsável pela transmissão das características hereditárias de cada espécie de todos os seres vivos. O DNA tem a forma parecida com uma escada espiral cuja disposição dos degraus se dá em quatro partes moleculares diferentes. Esta disposição constitui as chamadas quatro letras do código genético.

Observação 4.1. Na biologia, o ARN é a sigla que designa o ácido ribonucleico (ou, em inglês, RNA, *ribonucleic acid*). A composição do RNA é muito semelhante do DNA, contudo apresenta algumas diferenças. O RNA é um polímero de nucleotídeos, geralmente em cadeia simples, formado por moléculas de dimensões muito inferiores às do DNA.

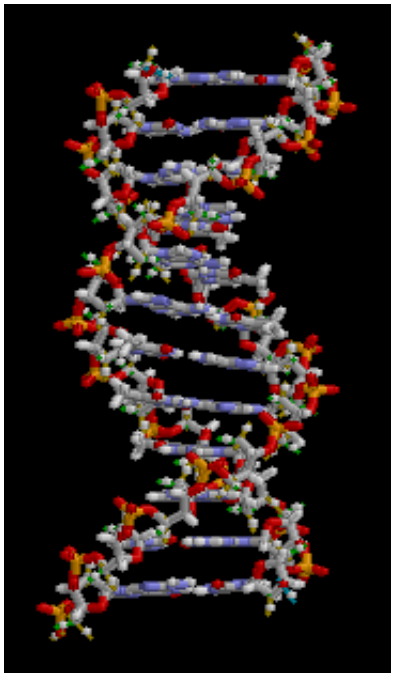

Figura 4.2: Seqüência de DNA.

O DNA é composto por açúcar (pentose), radicais fosfatos e por seqüências de quatro bases nitrogenadas, ligadas por pontes de hidrogênio, formando uma estrutura semelhante a uma escada em espiral (ver Figura 4.2). A seqüência de pares de bases se assemelha aos degraus, enquanto a desoxirribose e o agrupamento de fosfato se alternam, apresentando semelhança com o corrimão de uma escada em espiral. Devido a esta conformação, a cadeia de DNA fica com uma direção determinada, isto é, em uma extremidade temos livre a hidroxila do carbono-5 da primeira pentose e na outra temos livre a hidroxila do carbono-3 da última pentose. Isto determina que o crescimento da cadeia de DNA se faça na direção de 5 a 3 (ver Figura 4.3).

Definição 4.2 (Nucleotídeos). *Nucleotídeos* são compostos ricos em energia e que auxiliam os processos metabólicos, principalmente as biossínteses, na maioria das células. Funcionam ainda como sinais químicos, respondendo assim a hormônios e outros estímulos extracelulares; eles são também componentes estruturais de cofatores enzimáticos, intermediários metabólicos e ácidos nucleicos.

Observação 4.2. Os nucleotídeos são compostos por uma base nitrogenada, uma pentose e um grupo fosfato.

A seguir, definimos as quatro bases que compõem os nucleotídeos.

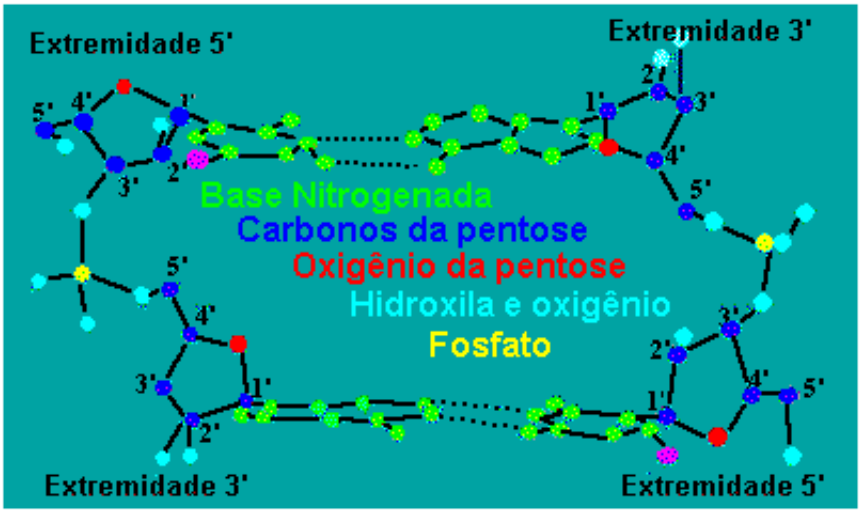

Figura 4.3: Crescimento da Seqüência de DNA na Direção 5' à 3'.

Definição 4.3 (Adenina). *Adenina* é uma das quatro bases nitrogenadas usadas na formação de nucleotídeos. No código genético é representada pela letra A. No DNA a *adenina* se emparelha com a *timina* através de duas ligações de hidrogênio. No RNA a *adenina* se emparelha com a *uracila* (U).

Definição 4.4 (Guanina). *Guanina* é uma base nitrogenada, orgânica, que se une com uma molécula de desoxirribose (pentose, monossacarídeo) e com um ácido fosfórico, geralmente o fosfato, para formar um nucleotídeo, principal base para formar cadeias polinucleotídeas que, por sua vez, formam o DNA (ácido desoxirribonucléico). No código genético é representada pela letra G.

Observação 4.3. *Adenina* e *guanina* são classificadas como *purinas* pois elas são moléculas compostas por dois anéis (ver Figura 4.4).

Definição 4.5 (Citosina). *Citosina* é uma fibra orgânica que constitui boa parte do citoplasma das células vivas, formando o chamado citoesqueleto. É uma substância cristalina, uma base nitrogenada, derivada do aminado da *pirimidina* cuja fórmula é a seguinte: C4H5N3O. É uma das quatro bases que compõem o código genético, e é representada pela letra C.

Definição 4.6 (Timina). *A timina* é uma base nitrogenada que compõe o nucleotídeo, a principal estrutura que forma o ácido desoxirubonucléico

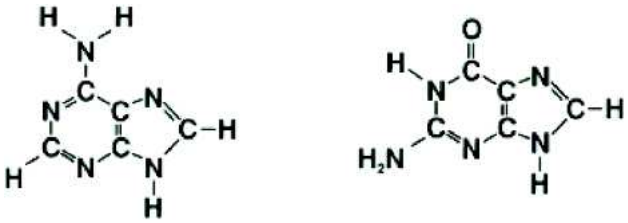

Figura 4.4: Estrutura Química da *Adenina* e da *Guanina.*

(DNA). A estrutura da *timina* é composta por substâncias químicas que formam uma molécula em um único anel. No código genético é representada pela letra T.

Observação 4.4. *Citosina* e *timina* são classificadas como *pirimidinas* pois elas são moléculas formadas por um único anel (ver Figura 4.5).

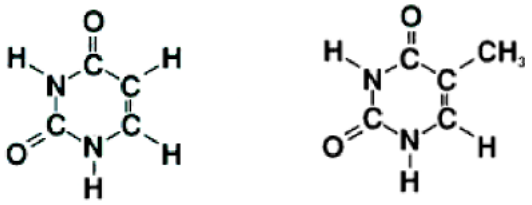

Figura 4.5: Estrutura Química da *Citosina* e da *Timina.*

Uma *purina* se liga a uma *pirimidina* no DNA para formar um par de base. *Adenina* e *timina* ligam-se uma à outra para formar um par de base A-T. Igualmente, *guanina* e *citosina* ligam-se uma à outra para formar um par de base G-C.

As bases permanecem unidas por fracas pontes de hidrogênio e são estas pontes de hidrogênio as responsáveis pela manutenção da estrutura do DNA (ver Figura 4.6).

A cadeia de DNA apresenta-se em uma estrutura de escada espiral que uma vez no núcleo recebe a ação de *histonas* e se enovela para formar a *cromatina* (ver Figura 4.7).

O DNA é encontrado em todos os seres vivos, incluindo os vírus, que ora possuem DNA, ora possuem RNA, porém, recentemente, foi encontrado um vírus raro que possui ao mesmo tempo, cadeia de DNA e RNA. O diâmetro de uma molécula de DNA é de cerca de $2,3nm$ (nanometros) ou $0,18\mu m$ (micra).

Observação 4.5. Neste trabalho, as palavras *nucleotídeo* e *base* serão usadas para representar a mesma coisa, isto é, um *nucleotídeo.*

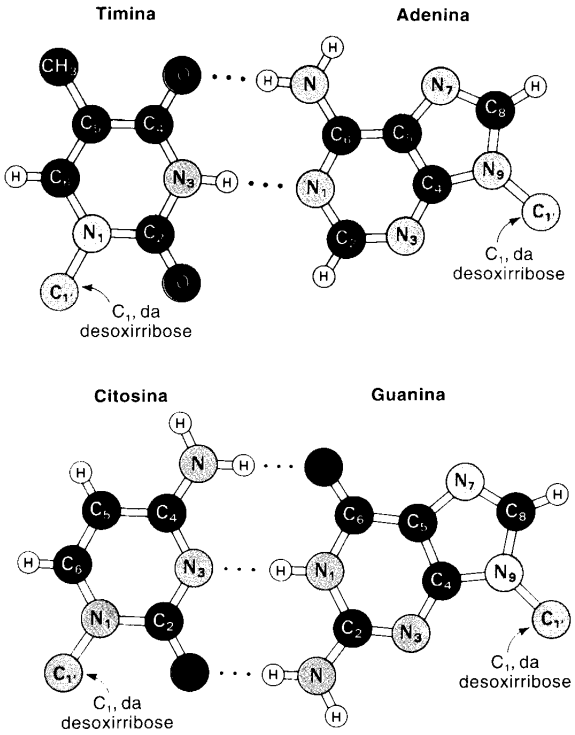

Figura 4.6: Ilustração de Pares de Base Unidas por Pontes de Hidrogênio.

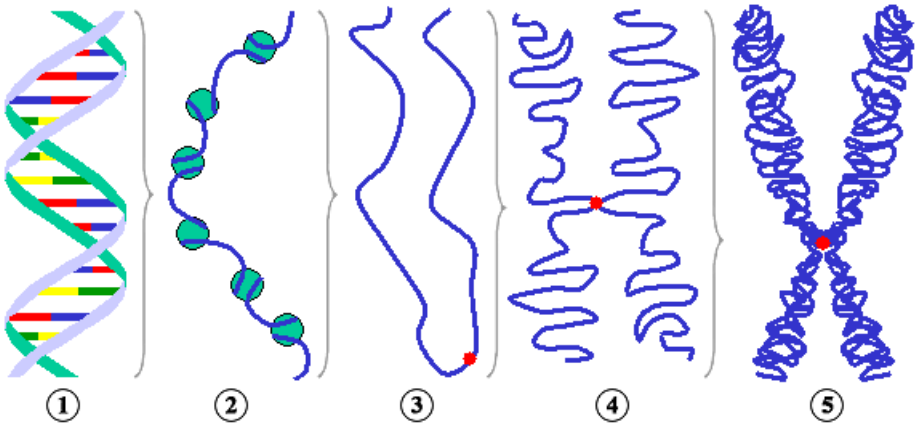

Figura 4.7: Diferentes Níveis de Condensação do DNA. (1) Cadeia simples de DNA . (2) Filamento de cromatina (DNA com histonas). (3) Cromatina condensada em interfase com centrômeros. (4) Cromatina condensada em prófase. (Existem agora duas cópias da molécula de DNA) (5) Cromossoma em metáfase.

A dupla cadeia polinucleotídica constitui a molécula de DNA, cuja seqüência de nucleotídeos codifica as instruções hereditárias, organizadas em genes,

que codificam as inúmeras proteínas existentes nas mais variadas células. As moléculas de DNA contém, portanto, a informação genética necessária para a codificação das características de um indivíduo, como a cor do cabelo em humanos, o formato da folha em Angiospermas e a sua morfologia.

Definição 4.7 (Mutações). *Mutações* são alterações na seqüência de nucleotídeos, que uma vez transmitida para a prole podem ocasionar mudanças nas características dos indivíduos.

Os tipos de mutações são: *mutação pontual* ou *modificação extensa*. As mutações pontuais podem ser por substituições de bases, inserções de bases ou deleções de bases. As *modificações extensas* incluem deleções, duplicações, inserções e rearranjos.

Denominamos de *transições* as substituições de bases nas quais uma purina substitui outra purina ou uma pirimidina substitui outra pirimidina. Por outro lado as *transversões* são mutações nas quais um tipo de base substitui o tipo oposto de base, por exemplo uma purina substituindo uma pirimidina.

As substituições de bases podem ser *silenciosas* ou *não. Silenciosas* nos casos em que a substituição de base não afeta o aminoácido que vai ser incorporado nesta posição.

Dependendo do tipo da alteração no código genético, a mesma poderá ser letal caso seja afetada a produção de enzimas e proteínas essenciais à sobrevida do organismo. Em outros casos, mutações podem conferir uma maior viabilidade ao indivíduo portador da alteração, favorecendo sua sobrevivência caso haja mudança na pressão seletiva do ambiente.

Geralmente, as mutações ocorrem ao acaso e são um dos principais fatores do processo evolutivo já que as mesmas contribuem para a existência de variação intra- e extra-específicas. Com a presença de variação, o processo seletivo pode conferir uma maior viabilidade para certos indivíduos de uma população.

4.2 Seqüência de DNA

Em vez de visualizarmos um diagrama molecular enorme de uma escada espiral de DNA, o que vemos freqüentemente é uma seqüência de letras, tais como “ATCTTAG”. Uma tal seqüência representa que bases estão em um determinado lado da escada de DNA. A seqüência do exemplo acima (“ATCTTAG”) representa o lado: “adenina-timina-citosina-timina-timina-adenina-guanina.”

As letras possíveis são A, C, G e T, representando os quatro nucleotídeos (subunidades) de uma cadeia de DNA - as bases adenina, citosina, guanina e timina. Uma sucessão de quaisquer nucleotídeos maior do que quatro está apta a ser considerada uma seqüência. Com respeito à sua função biológica, a qual pode depender do contexto, uma seqüência pode ou não “fazer sentido”,

e também ser ou não codificável. Seqüências de DNA também podem conter "DNA lixo". Seqüências podem ser obtidas de material biológico através de um processo denominado *seqüenciamento de DNA*.

A seguir, introduzimos uma definição formal para *seqüências de nucleotídeos*.

Definição 4.8 (Seqüência de DNA). Uma *seqüência de DNA* ou *seqüência de nucleotídeos* é uma série de letras representando a estrutura primária de uma molécula ou cadeia de DNA, real ou hipotética, com a capacidade de carregar informação.

Definição 4.9 (Gene). O *gene* é a unidade fundamental da hereditariedade. Cada *gene* é formado por uma seqüência de nucleotídeos. O *gene* é um dos fatores que determinam a forma ou função de uma ou várias características dos seres vivos, pois é por meio de genes que são determinadas as proteínas.

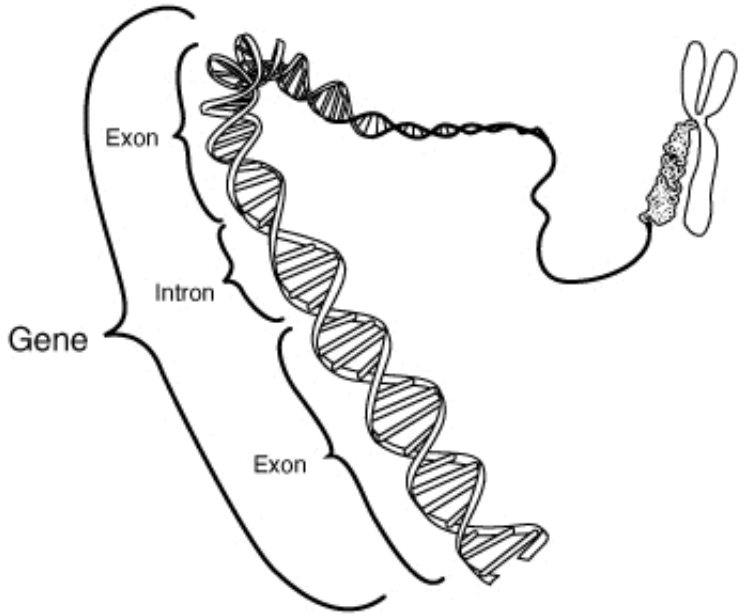

Figura 4.8: Esquema Ilustrando o Gene com Relação à Estrutura do DNA.

Uma das características marcantes dos *genes* eucarióticos é a presença de *seqüências de nucleotídeos* intervenientes interrompendo a região codificadora. Estas seqüências intervenientes, denominadas de *introns*, dividem o gene em várias *seqüências de nucleotídeos*, denominadas *exons* (ver Figura 4.8). A seguir definimos *introns* e *exons.*

Definição 4.10 (*Introns*). Os *introns* são seqüências de nucleotídeos que não geram proteínas. São também conhecidos pelo termo DNA-lixo. Os

introns são seções de DNA de um gene que não codificam qualquer parte da proteína produzida pelo gene e que separa a seqüência constituída pelos *exons*.

Definição 4.11 (*Exons*). Os *exons* são seqüências de nucleotídeos de um determinado gene eucariótico que geram proteínas. Geralmente adjacente a uma seqüência de DNA não codificante (*intron*).

Por razões desconhecidas, os *introns* sofrem mais mudanças que os *exons* durante a duplicação, fazendo com que a correlação de *longa dependência* seja mais evidente através deles.

4.3 Diferentes Transformações Aplicadas aos Nucleotídeos

Uma seqüência de nucleotídeos $\{n_i\}_{i=1}^n$ de tamanho n, é compreendida de bases A (adenina), C (citosina), T (timina) e G (guanina), ou seja, $n_i \in \{\text{A,C,T,G}\}$. A fim de estudar longa dependência em seqüências de DNA, é preciso transformar as seqüências de nucleotídeos, em seqüências numéricas. Apresentaremos aqui diversas transformações, que aplicam métodos numéricos a uma seqüência de nucleotídeos.

Dada uma seqüência de nucleotídeos $\{n_i\}_{i=1}^n \equiv \{n_1, n_2, \cdots, n_n\}$ de tamanho n, apresentamos, a seguir, funções que transformam a seqüência de nucleotídeos $\{n_i\}_{i=1}^n$ em uma seqüência numérica $\{g(n_i)\}_{i=1}^n$, tal que $g(n_i) \in \mathbb{R}$ (ver Garcia e José, 2005; Stanley et al., 1999; Buldyrev et al., 1995 e Chakravarthy et al., 2004).

Regra RY. Definimos a *transformação* $g_1 : \{n_1, n_2, \cdots, n_n\} \to \mathbb{R}$, considerando a seguinte regra

$$g_1(n_i) = \begin{cases} -1, & \text{se } n_i \in \{\text{A,G}\} \\ +1, & \text{se } n_i \in \{\text{C,T}\}. \end{cases} \tag{4.1}$$

Regra A$\bar{\text{A}}$. A *transformação* $g_2 : \{n_1, n_2, \cdots, n_n\} \to \mathbb{R}$, é definida por

$$g_2(n_i) = \begin{cases} +1, & \text{se } n_i = \text{A} \\ -1, & \text{caso contrário.} \end{cases}$$

Regra T$\bar{\text{T}}$. Definimos a *transformação* $g_3 : \{n_1, n_2, \cdots, n_n\} \to \mathbb{R}$, considerando a seguinte regra

$$g_3(n_i) = \begin{cases} +1, & \text{se } n_i = \text{T} \\ -1, & \text{caso contrário.} \end{cases}$$

Regra G$\bar{\text{G}}$. A *transformação* $g_4 : \{n_1, n_2, \cdots, n_n\} \to \mathbb{R}$, é definida por

$$g_4(n_i) = \begin{cases} +1, & \text{se } n_i = \text{G} \\ -1, & \text{caso contrário.} \end{cases}$$

Regra C$\bar{\text{C}}$. Definimos a *transformação* $g_5 : \{n_1, n_2, \cdots, n_n\} \to \mathbb{R}$, considerando a seguinte regra

$$g_5(n_i) = \begin{cases} +1, & \text{se } n_i = \text{C} \\ -1, & \text{caso contrário.} \end{cases}$$

Regra Real. O modelo Real, proposto por Chakravarthy el al. (2004), considera a *transformação* $g_6 : \{n_1, n_2, \cdots, n_n\} \to \mathbb{R}$, definida por

$$g_6(n_i) = \begin{cases} -1,5, & \text{se } n_i = \text{A} \\ +1,5, & \text{se } n_i = \text{T} \\ +0,5, & \text{se } n_i = \text{C} \\ -0,5, & \text{se } n_i = \text{G}. \end{cases}$$

As transformações abaixo tomam valores em $\mathbb{R}^N$, onde $N \geq 2$.

Regra Guharay et al. (2000). Guharay el al. (2000) considera as quatro bases A,C,G,T em um tetraedro onde cada letra é equidistante das demais e as componentes dos vetores somam zero. Este modelo é construído da seguinte maneira: tomar os vetores equidistantes $\mathbf{e_1} = (1,0,0,0), \mathbf{e_2} = (0,1,0,0), \mathbf{e_3} = (0,0,1,0)$ e $\mathbf{e_4} = (0,0,0,1)$, e subtrair $0,25$ de cada vetor $\mathbf{e}_i$, $i \in \{1,2,3,4\}$, para obter a transformação $g_7 : \{n_1, n_2, \cdots, n_n\} \to \mathbb{R}^4$, definida por

$$g_7(n_i) = \begin{cases} (0,75; -0,25; -0,25; -0,25), & \text{se } n_i = \text{A} \\ (-0,25; 0,75; -0,25; -0,25), & \text{se } n_i = \text{C} \\ (-0,25; -0,25; 0,75; -0,25), & \text{se } n_i = \text{G} \\ (-0,25; -0,25; -0,25; 0,75), & \text{se } n_i = \text{T}. \end{cases} \tag{4.2}$$

A vantagem da utilização da Regra Guharay et al. (2000) sobre as demais, é que esta transformação considera cada nucleotídeo com o mesmo peso.

Regra Cristea (2002). Cristea (2002) propõe a seguinte representação para uma seqüência de nucleotídeos: considera quatro vetores de mesmo tamanho, onde cada um é simétrico aos demais, i.e., orientado para os cantos de um tetraedro, e os vetores são colocados em correspondência com os nucleotídeos, como é mostrado na Figura 4.9 na qual destaca-se que os vértices de um tetraedro regular são subconjuntos dos vértices de um cubo.

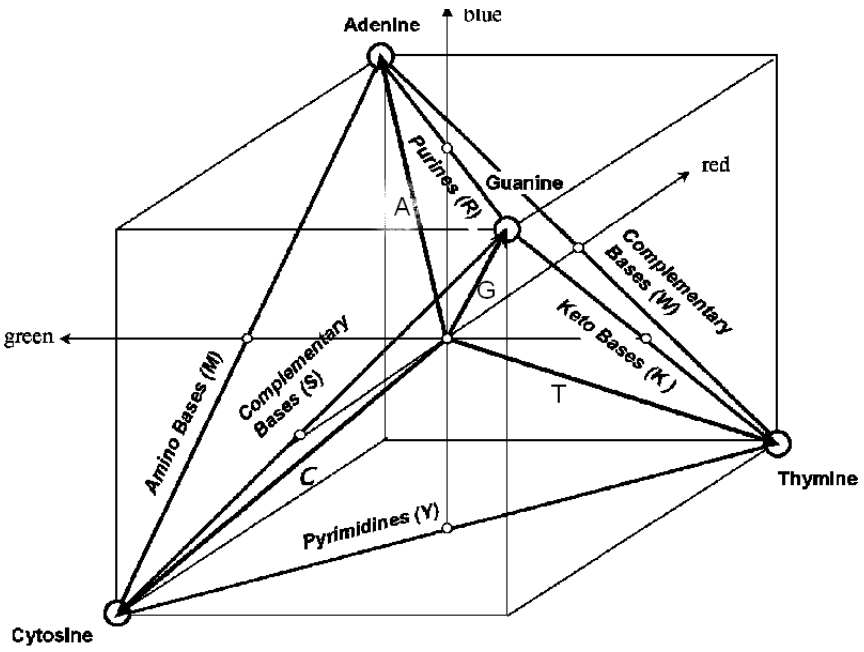

Figura 4.9: Representação através de um Tetraedro.

A descrição matemática do código resultante pode ser simplificada escolhendo coordenadas de número inteiros $\{\pm 1\}$ para os vértices do cubo, incluindo os pontos que representam as bases, sem impor uma condição de normalização. Obtemos a transformação $g_8 : \{n_1, n_2, \cdots, n_n\} \to \mathbb{R}^3$, definida por

$$g_8(n_i) = \begin{cases} (+1;+1;+1), & \text{se } n_i = \text{A} \\ (-1;+1;-1), & \text{se } n_i = \text{C} \\ (-1;-1;+1), & \text{se } n_i = \text{G} \\ (+1;-1;-1), & \text{se } n_i = \text{T}. \end{cases}$$

Representação de Jogo de Caos. As características do procedimento

CGR ("*Chaos Game Representation*") (ver Almeida et al., 2001) são ilustradas analisando a seqüência de *E.coli* threonine gene *thrA* (ver Figura 4.10). O espaço CGR gerado por seqüências genômicas é plano e é confinado pelos quatro possíveis nucleotídeos como vértices de um quadrado binário.

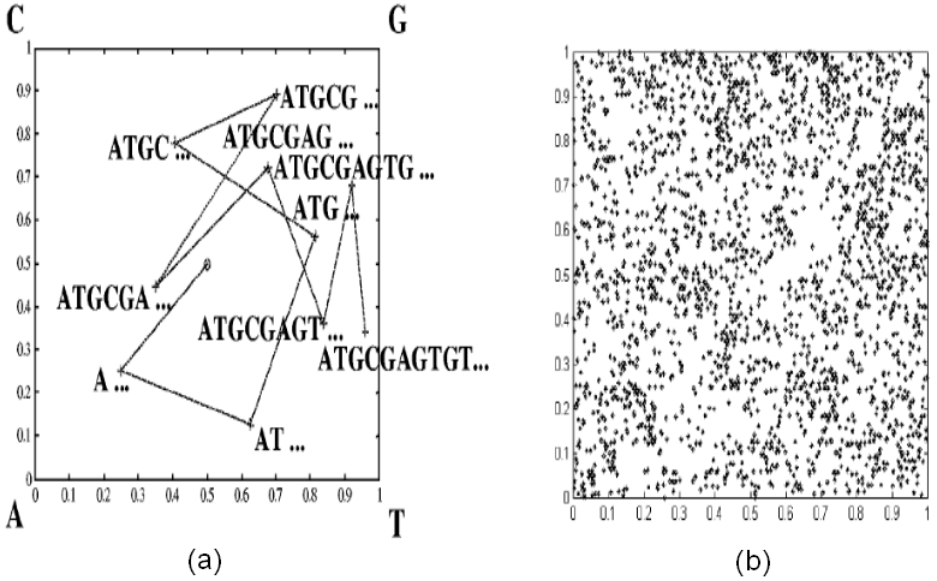

Figura 4.10: (a) CGR dos 10 primeiros nucleotídeos da seqüência *E. coli* gene *thrA*: ATGCGAGTGT. (b) CGR de toda a seqüência *E. coli* gene *thrA* totalizando 2.463 pares de bases.

A posição $g_9(n_i) = \mathrm{CGR}_{n_i}$, de cada nucleotídeo n_i, da seqüência $\{n_i\}_{i=1}^{n}$, de tamanho n, é calculada construindo um segmento, a partir do ponto anterior em direção à representação binária do nucleotídeo n_i, até a metade da distância deste ponto anterior e a representação binária do nucleotídeo n_i.

A representação binária dos vértices de CGR foram designados pela transformação $f : \{n_1, n_2, \cdots, n_n\} \to \mathbb{R}^2$, definida por

$$f(n_i) = \begin{cases} (0;0), & \text{se } n_i = \text{A} \\ (0;1), & \text{se } n_i = \text{C} \\ (1;1), & \text{se } n_i = \text{G} \\ (1;0), & \text{se } n_i = \text{T}. \end{cases} \tag{4.3}$$

Por fim, definimos a transformação $g_9 : \{n_1, n_2, \cdots, n_n\} \to \mathbb{R}^2$ que indica a posição de cada nucleotídeo

$$g_9(n_i) = \begin{cases} (0,5;0,5) + 0.5[(0,5;0,5) - f(n_i)], & \text{se } n_i = n_1 \\ g_9(n_{i-1}) + 0.5[g_9(n_{i-1}) - f(n_i)], & \text{se } n_i \neq n_1, \end{cases} \tag{4.4}$$

onde $f(\cdot)$ é dada pela expressão (4.3).

Regra Stofer e Rosen (2007). Stofer e Rosen (2007), consideram a transformação $g_{10} : \{n_1, n_2, \cdots, n_n\} \to \mathbb{R}^3$, definida por

$$g_{10}(n_i) = \begin{cases} (1;0;0), & \text{se } n_i = \text{A} \\ (0;0;0), & \text{se } n_i = \text{T} \\ (0;1;0), & \text{se } n_i = \text{C} \\ (0;0;1), & \text{se } n_i = \text{G}. \end{cases}$$

4.4 Série Temporal

Apresentamos aqui a série temporal que é utilizada no presente trabalho. Vimos, na Seção 3.3, que podemos considerar diferentes transformações para uma seqüência de DNA. Neste trabalho utilizamos a regra RY (veja a transformação $g_1(\cdot)$ dada na expressão (4.1)) em seqüências de DNA, no qual se desloca uma posição para outra em uma série de passos de mesmo tamanho. Cada passo pode ser dado na direção para frente ou para trás. Segue-se abaixo a definição da série temporal que representa uma seqüência de DNA neste trabalho.

Definição 4.12. Dada uma seqüência de nucleotídeos $\{n_i\}_{i=1}^n$, *a série temporal* $\{X_t\}_{t=1}^n$, é definida por

$$X_t = g_1(n_t), \tag{4.5}$$

onde $g_1(\cdot)$ é dada pela expressão (4.1).

Definição 4.13 (Deslocamento). Dada uma série temporal $\{X_t\}_{t=1}^n$ definida pela expressão (4.5), o *deslocamento* Y_t *da série temporal* $\{X_t\}_{t=1}^n$ é definido por

$$Y_t = \sum_{i=1}^{t} X_i, \quad t \in \{1, 2, \cdots, n\}. \tag{4.6}$$

Exemplo 4.1. Para exemplificar a construção do deslocamento $\{Y_t\}_{t=1}^n$, de uma série temporal $\{X_t\}_{t=1}^n$, conforme a Definição 4.13, consideramos a seqüência de DNA da *Enterobacteria phage lambda* (nome no GenBank: LAMCG). Esta seqüência é formada por 48.502 pares de bases. A Figura 3.11 mostra o deslocamento da série temporal $\{X_t\}_{t=1}^n$ para a seqüência de DNA da *Enterobacteria phage lambda*.

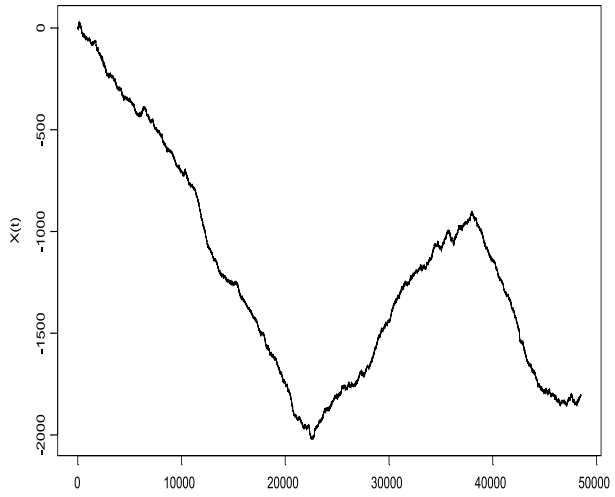

Figura 4.11: Deslocamento da Série Temporal $\{X_t\}_{t=1}^n$ para a Seqüência LAMCG, com $n = 48.502$.

Capítulo 5

Análise de Seqüências de DNA

Vimos no Cápítulo 3 que os processos ARFIMA(p, d, q) exibem a característica de *longa dependência* quando $d \in (0,0; 0,5)$, a de *curta dependência* quando $d = 0,0$ e a de *dependência intermediária* quando $d \in (-0,5; 0,0)$. Neste capítulo analisamos diversas seqüências de DNA, com o objetivo de detectar longa dependência. Para verificar a existência de longa dependência em cada seqüência de DNA, estimamos o parâmetro d através dos diversos métodos de estimação, apresentados no Capítulo 3. Consideramos os seguintes estimadores para o parâmetro de diferenciação d: GPH, GPH-LS, GPH-LTS, GPH-MM, R, R-LTS, R-MM, W, R/S(n), R/S(q) e DFA.

Para estimar o parâmetro de diferenciação d pelos métodos R/S(n) e o DFA, utilizamos a seguinte relação

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

em que α é o coeficiente de escala obtido pelo método DFA, para medir longa dependência e H é o parâmetro sugerido por Harold Edwin Hurst (1880-1978), também para medir longa dependência.

Utilizamos seqüências de nucleotídeos disponíveis no Instituto Europeu de Bioinformáticas (EBI, http://www.ebi.ac.uk/) e no Centro Nacional de Informação Biotecnológica (NCBI, http://www.ncbi.nlm.nih.gov/).

Cabe ressaltar que os métodos de estimação para o parâmetro de diferenciação d, utilizados neste trabalho, estão desenvolvidos no contexto dos softwares S-Plus e R-project.

Por muito tempo, os seres vivos foram classificados em dois grandes reinos: Animal e Vegetal. Posteriormente, outras classificações foram estipuladas, até 1969, quando o cientista americano R.H. Whittaker propôs uma nova classificação para os seres vivos, dividindo-os em cinco reinos, atualmente mais aceita: Monera, Protista, Fungi, Plantae ou Metaphyta e Animalia ou Metazoa.

Os *eucarióticos* variam desde organismos unicelulares até gigantescos organismos multicelulares, nos quais as células se diferenciam e desempenham

funções diversas, não sobrevivendo isoladamente. Fazem parte desta categoria, os Reinos: animalia, plantae, fungi e o protista. As formas vivas que não fazem parte do domínio Eukariota são as bactérias e as Archaea (anteriormente denomidadas arqueobactérias), ou seja, os seres vivos com células procarióticas. Também há os vírus, que são seres acelulados.

Neste capítulo selecionamos seqüências de nucleotídeos de vírus e de seres vivos que são dos reinos: Monera, Protista, Fungi, Plantae ou Metaphyta e Animalia ou Metazoa. Analisamos, a seguir, separadamente as seqüências de cada reino.

Em cada seqüência, para todos os estimadores propostos neste trabalho, testamos as hipóteses $H_0 : d = 0$ versus $H_1 : d \neq 0$, ou seja, testamos se as seqüências de DNA tem ou não curta dependência.

Para cada seqüência de DNA, representamos graficamente os intervalos de confiança para o parâmetro de diferenciação d ao nível de 95% de confiança, utilizando todos os estimadores propostos no Capítulo 3.

Observação 5.1.

1) A estatística do teste $H_0 : d = 0$ versus $H_1 : d \neq 0$ para cada estimador $\widehat{d}$ utilizado neste trabalho, é dada por

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

que tem função de distribuição $\mathcal{N}(0,1)$, onde $Var(\widehat{d})$ é a variância do estimador $\widehat{d}$ para cada método de estimação do parâmetro d, apresentados no Capítulo 3.

2) Em todas as tabelas deste Capítulo utilizamos a sequinte notação:

* : Rejeita-se H_0 ao nível de 10% de significância

** : Rejeita-se H_0 ao nível de 5% de significância.

3) Os limites para o intervalo de confiança do parâmetro de diferenciação d, baseados nas diversas estimativas $\widehat{d}$ utilizadas neste trabalho, são dados por

$$\text{limite inferior} = \widehat{d} - z_{\frac{\alpha}{2}} \cdot \sigma_{\widehat{d}}$$
$$\text{limite superior} = \widehat{d} + z_{\frac{\alpha}{2}} \cdot \sigma_{\widehat{d}},$$

onde $z_{\frac{\alpha}{2}}$=1,96 e $\sigma_{\widehat{d}} = \sqrt{Var(\widehat{d})}$.

5.1 Vírus

Nesta seção analisamos seqüências de nucleotídeos em vírus. Os vírus são seres muito simples, formados basicamente por uma cápsula protéica envolvendo o material genético (ver Figura 5.1), que, dependendo do tipo de vírus, pode ser o DNA ou o RNA, nunca os dois juntos.

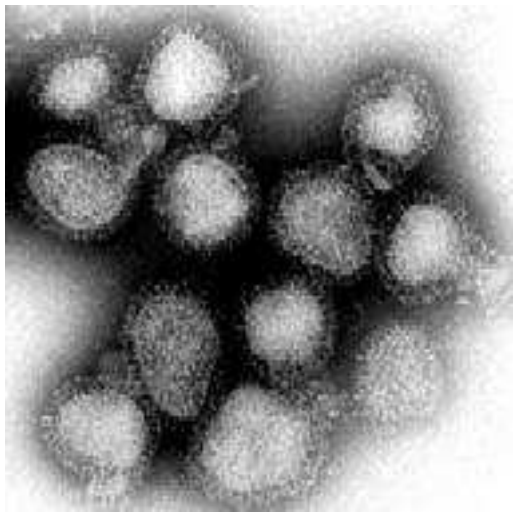

Figura 5.1: Vírus da Gripe.

Definição 5.1 (Vírus). *Vírus* é uma partícula basicamente protéica que pode infectar organismos vivos. *Vírus* são parasitas obrigatórios do interior celular e isso significa que eles somente reproduzem-se pela invasão e possessão do controle da maquinaria de auto-reprodução celular.

Na Tabela 5.1 encontramos códigos de acesso de seqüências de nucleotídeos em vírus, do Instituto Europeu de Bioinformáticas (EBI).

Tabela 5.1: Seqüências de Nucleotídeos no Reino Vírus.

Vírus		
Local	**Código de Acesso**	**Tamanho**
EBI	LAMCG	48.502
EBI	AF03812	5.894
EBI	AJ965540	2.745
EBI	J02057	2.779
EBI	M20036	4.801

A Tabela 5.2 apresenta a análise dos estimadores GPH, GPH-LS, GPH-LTS, GPH-MM, R, R-LTS, R-MM, W, R/S(n), R/S(q) e DFA para cada uma das cinco seqüências de DNA apresentadas na Tabela 5.1.

Tabela 5.2: Estimadores do Parâmetro de Diferenciação com seus respectivos níveis de significância em Seqüências de Nucleotídeos em Vírus.

	Vírus				
Seqüência	**LAMCG**	**AF03812**	**AJ965540**	**J02057**	**M20036**
GPH	$0,0468^{**}$	$0,0414^{**}$	0,0305	0,0112	$0,0588^{**}$
GPH-LTS	$0,0312^{**}$	$0,0657^{**}$	0,0041	0,0095	$0,0269^{**}$
GPH-MM	$0,0312^{**}$	$0,0801^{**}$	0,0154	0,0172	$0,0281^{**}$
R	$0,0439^{**}$	$0,0389^{**}$	0,0252	0,0033	$0,0573^{**}$
R-LTS	$0,0172^{**}$	$0,0693^{**}$	-0,0100	-0,0245	$0,0198^{**}$
R-MM	$0,0312^{**}$	$0,0834^{**}$	0,0058	0,0020	$0,0257^{**}$
W	$0,0212^{**}$	$0,1096^{**}$	0,0193	$0,0292^{**}$	0,0087
R/S(n)	$0,1004^{**}$	$0,1405^{**}$	0,1139	0,1182	0,0998
R/S(q)	$0,0401^{**}$	$0,0667^{**}$	0,0457	0,0450	0,0383
DFA	$0,0240^{**}$	$0,0562^{**}$	$0,0232^{**}$	$0,0340^{**}$	$0,0364^{**}$

Analisando a Tabela 5.2, podemos observar que a seqüência J02057, apresentou o menor estimador R-LTS $= -0,0245$ e a seqüência AF03812 obteve o maior estimador R/S(n) $= 0,1405$, para o parâmetro de diferenciação d. Nota-se que em todas as seqüências da Tabela 5.1, o estimador R/S(n) obteve a maior estimativa para o parâmetro de diferenciação d. Isto era esperado pois o estimador R/S(n) é viciado e super estima o parâmetro d. A existência de pequena longa dependência nas seqüências LAMCG e AF03812, é estatisticamente significante ao nível de 5%, para todos os métodos de estimação propostos neste trabalho. A existência de pequena longa dependência nas demais seqüências da Tabela 5.2, é estatisticamente significante ao nível de 5%, utilizando pelo menos um método de estimação proposto neste trabalho.

As Figuras 5.2 a 5.6 apresentam, graficamente, os intervalos à 95% de confiança para o parâmetro de diferenciação d, respectivamente, de cada seqüência da Tabela 5.1.

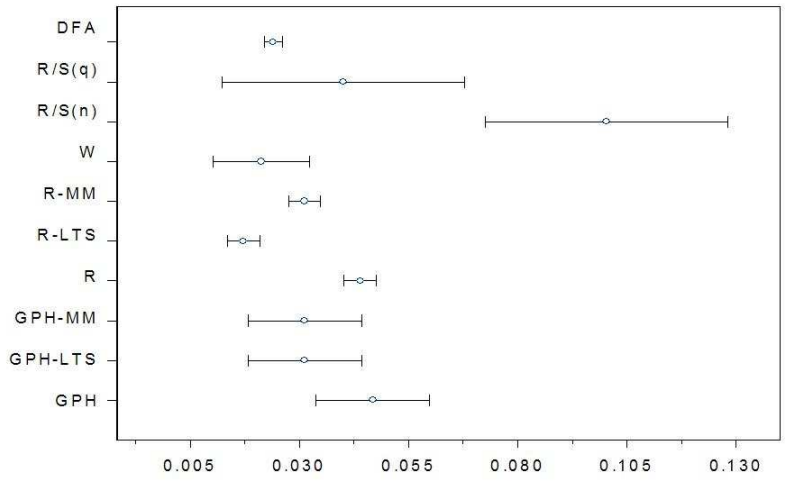

Figura 5.2: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência LAMCG.

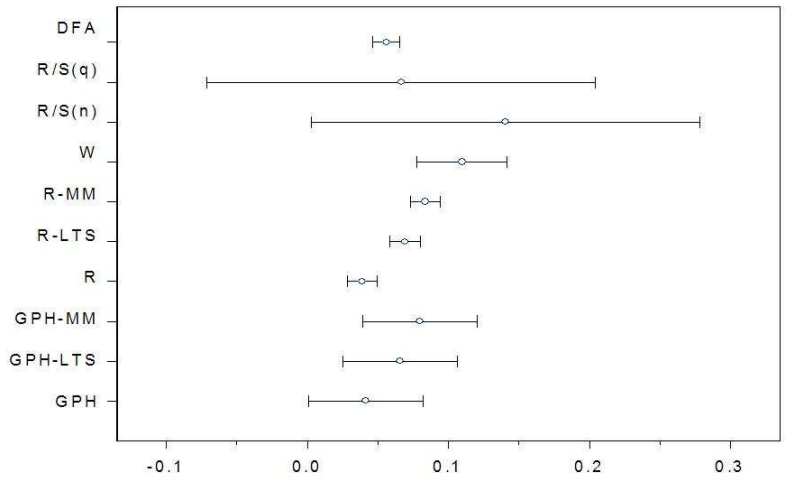

Figura 5.3: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência AF03812.

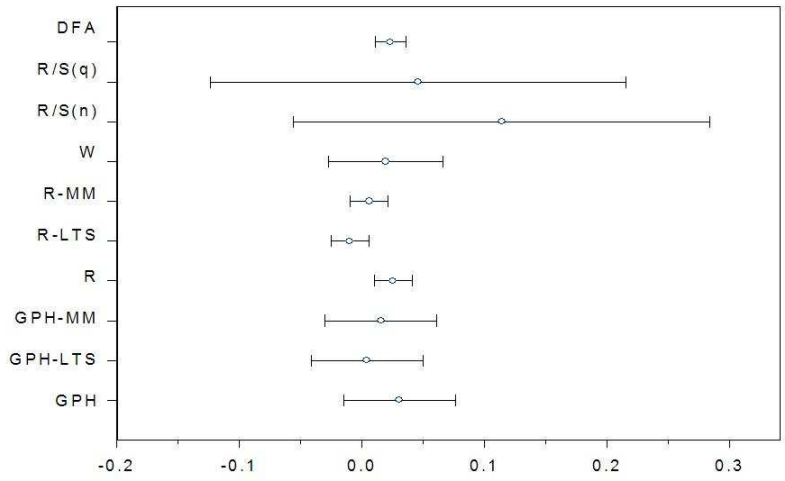

Figura 5.4: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência AJ965540.

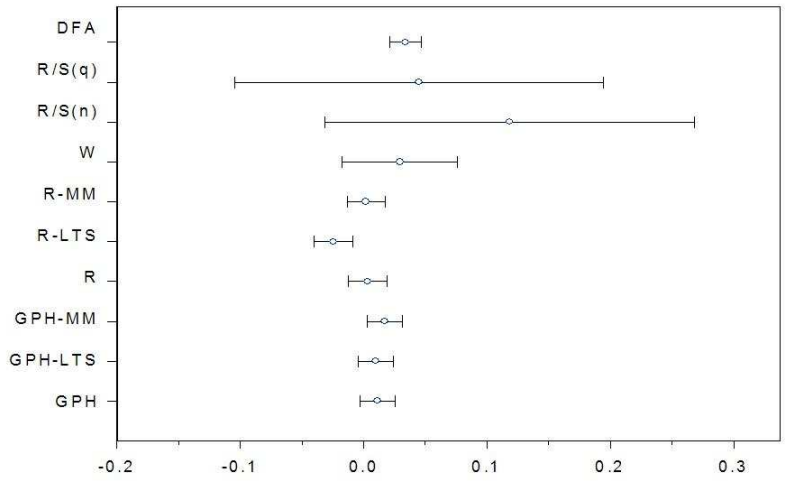

Figura 5.5: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência J02057.

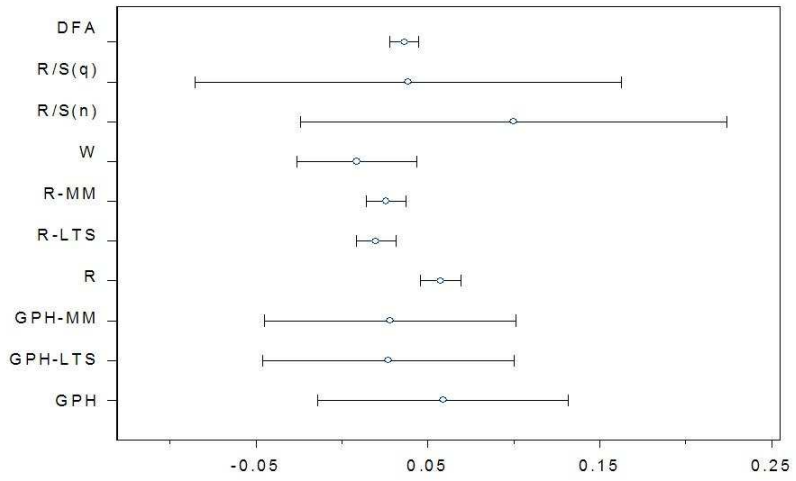

Figura 5.6: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência M20036.

5.2 Reino Monera

Definição 5.2 (Reino Monera). O *reino Monera* compreende todos os organismos unicelulares e procariontes, representados pelas bactérias e pelas algas azuis ou cianofíceas.

A Figura 5.7 ilustra a bactéria *Escherichia coli* e a Tabela 5.3 contém códigos de acesso, de seqüências de nucleotídeos no reino Monera, do Instituto Europeu de Bioinformáticas (EBI, http://www.ebi.ac.uk/).

Figura 5.7: Bactéria *Escherichia coli*.

Tabela 5.3: Seqüência de Nucleotídeos no Reino Monera.

Reino Monera		
Local	**Código de Acesso**	**Tamanho**
EBI	AF238307	10.276
EBI	U20550	1.440
EBI	AF110140	5.123
EBI	AJ311718	4.379

A Tabela 5.4 apresenta a análise dos estimadores GPH, GPH-LS, GPH-LTS, GPH-MM, R, R-LTS, R-MM, W, R/S(n), R/S(q) e DFA para cada uma das quatro seqüências de DNA apresentadas na Tabela 5.3.

Tabela 5.4: Estimadores do Parâmetro de Diferenciação com seus respectivos níveis de significância em Seqüências de Nucleotídeos no Reino Monera.

Reino Monera				
Seqüência	AF238307	U20550	AF110140	AJ311718
GPH	$0,1007^{**}$	0,0301	0,0071	$-0,0570^{**}$
GPH-LTS	$0,0875^{**}$	0,0380	-0,0080	$-0,0833^{**}$
GPH-MM	$0,0918^{**}$	0,0188	-0,0183	$-0,0242^{**}$
R	$0,1020^{**}$	0,0276	0,0003	$-0,0623^{**}$
R-LTS	$0,1043^{**}$	0,0175	0,0036	$-0,0867^{**}$
R-MM	$0,0976^{**}$	0,0903	-0,0181	$-0,0334^{**}$
W	$0,0711^{**}$	$0,0382^{*}$	-0,0037	$-0,0844^{**}$
R/S(n)	$0,1409^{*}$	0,1431	$0,0937^{*}$	0,0509
R/S(q)	$0,0699^{*}$	0,0615	$0,0348^{*}$	0,0050
DFA	$0,1026^{**}$	$0,0868^{**}$	$0,0350^{**}$	$-0,0570^{**}$

Analisando a Tabela 5.4, podemos observar que a seqüência AJ311718, apresentou o menor estimador R-LTS $= -0,0867$ e a seqüência U20550 obteve o maior estimador R/S(n) $= 0,1431$, para o parâmetro de diferenciação d. Nota-se que em todas as seqüências da Tabela 5.3, o estimador R/S(n) obteve a maior estimativa para o parâmetro de diferenciação d. A existência de pequena longa dependência na seqüência AF238307 é estatisticamente significante ao nível de 10%, para todos os métodos de estimação propostos neste trabalho. A existência de pequena longa dependência nas demais seqüências da Tabela 5.4 é estatisticamente significante ao nível de 10%, utilizando pelo menos um método de estimação proposto neste trabalho.

As Figuras 5.8 a 5.11 apresentam, graficamente, os intervalos à 95% de confiança para o parâmetro de diferenciação d, respectivamente, de cada seqüência da Tabela 5.3.

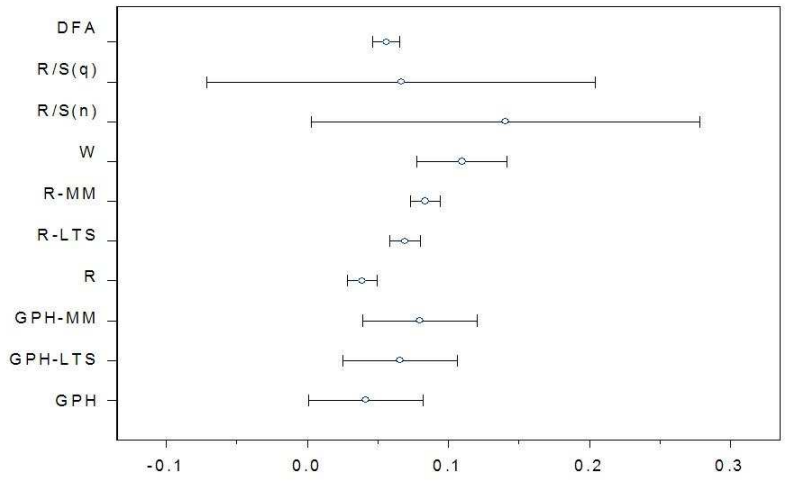

Figura 5.8: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência AF238307.

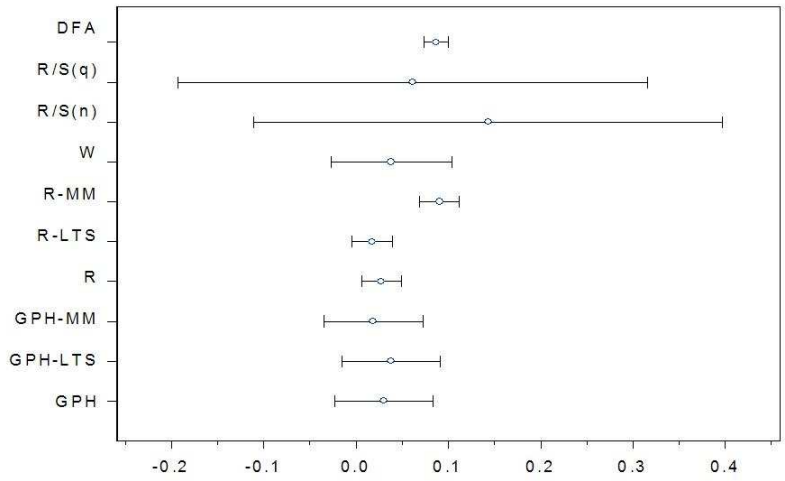

Figura 5.9: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência U20550.

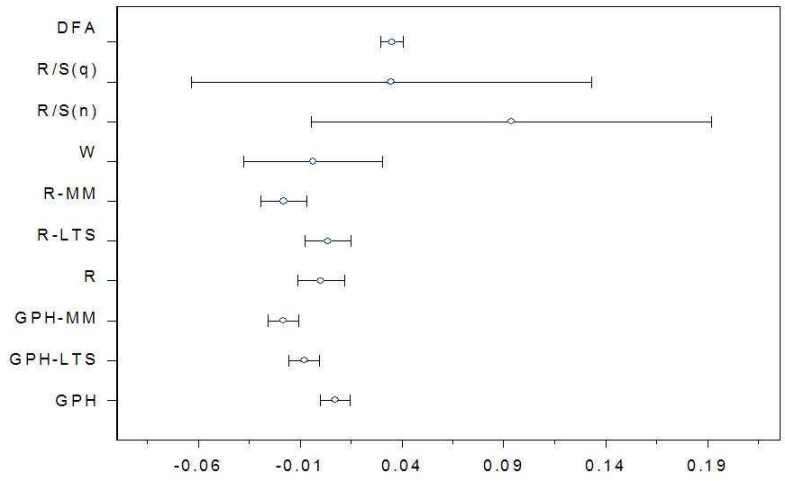

Figura 5.10: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência AF110140.

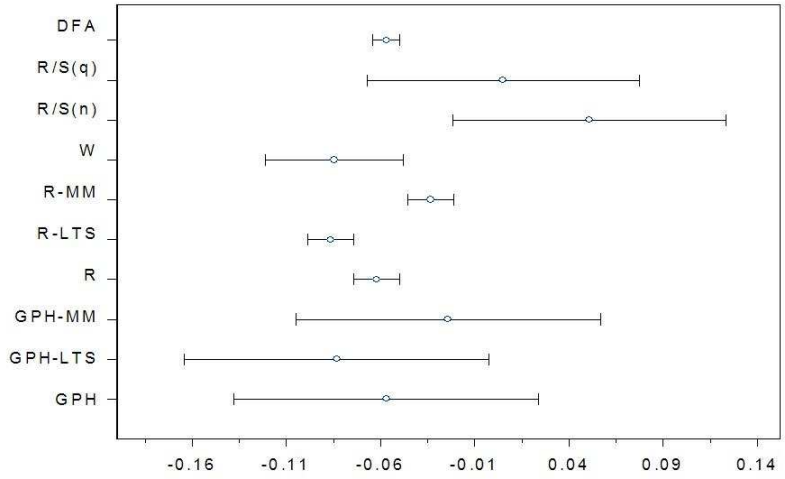

Figura 5.11: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência AJ311718.

5.3 Reino Animalia

Definição 5.3 (Reino Animalia). *O reino Animalia, Reino Animal* ou *Reino Metazoa* é composto por seres vivos multicelulares cujas células formem tecidos biológicos, com capacidade de responder ao ambiente que os envolve ou, por outras palavras, pelos animais.

A Figura 5.12 ilustra alguns seres vivos pertencentes ao reino Animalia e na Tabela 5.5 encontramos códigos de acesso, de seqüências de nucleotídeos no reino Animalia, do Instituto Europeu de Bioinformáticas (EBI) e do Centro Nacional de Informação Biotecnológica (NCBI).

A Tabela 5.6 apresenta a análise dos estimadores GPH, GPH-LS, GPH-LTS, GPH-MM, R, R-LTS, R-MM, W, R/S(n), R/S(q) e DFA para cada uma das sete seqüências de DNA apresentadas na Tabela 5.5.

Figura 5.12: Animais.

Tabela 5.5: Seqüência de Nucleotídeos no Reino Animalia.

Reino Animalia		
Local	**Código de Acesso**	**Tamanho**
EBI	AF200828	14.905
EBI	U37541	19.517
NCBI	M94081	97.630
NCBI	NM004023	7.048
EBI	U96639	16.727
EBI	X03240	16.019
EBI	X54252	13.794

As Figuras 5.13 a 5.19 apresentam, graficamente, os intervalos à 95% de confiança para o parâmetro de diferenciação d, respectivamente, de cada seqüência da Tabela 5.5.

Analisando a Tabela 5.6, podemos observar que a seqüência U3754, apresentou o menor estimador DFA $= -0,0115$ e a seqüência M94081 obteve

Tabela 5.6: Estimadores do Parâmetro de Diferenciação e os seus respectivos níveis de significância em Seqüências de Nucleotídeos no Reino Animalia.

Reino Animalia							
Seqüência	AF200828	U3754	M94081	NM004023	U96639	X03240	X54252
GPH	$0,0414^{**}$	$0,0334^{**}$	$0,1115^{**}$	$0,0543^{**}$	$0,0620^{**}$	$0,0478^{**}$	$0,0143$
GPH-LTS	$0,0194^{**}$	$0,0495^{**}$	$0,1048^{**}$	$0,0704^{**}$	$0,0424^{**}$	$0,0277^{**}$	$0,0177$
GPH-MM	$0,0492^{**}$	$0,0486^{**}$	$0,1289^{**}$	$0,0325^{**}$	$0,0580^{**}$	$0,0352^{**}$	$0,0231$
R	$0,0366^{**}$	$0,0305^{**}$	$0,1121^{**}$	$0,0488^{**}$	$0,0582^{**}$	$0,0441^{**}$	$0,0126$
R-LTS	$0,0223^{**}$	$0,0128^{**}$	$0,1302^{**}$	$0,0645^{**}$	$0,0577^{**}$	$0,0294^{**}$	$0,0055$
R-MM	$0,0488^{**}$	$0,0486^{**}$	$0,1287^{**}$	$0,0325^{**}$	$0,0577^{**}$	$0,0353^{**}$	$0,0176$
W	$0,0517^{**}$	$0,0503^{**}$	$0,1235^{**}$	$0,0933^{**}$	$0,0343^{**}$	$0,0503^{**}$	$0,0481^{**}$
R/S(n)	$0,1092^{**}$	$0,1092^{**}$	$0,1540^{*}$	$0,1276^{**}$	$0,1045^{**}$	$0,1099^{**}$	$0,1178^{*}$
R/S(q)	$0,0477^{**}$	$0,0458^{**}$	$0,1025^{*}$	$0,0537^{**}$	$0,0464^{**}$	$0,0481^{**}$	$0,0512^{*}$
DFA	$0,0321^{**}$	$0,0241^{**}$	$0,1094^{**}$	$0,0624^{**}$	$0,0436^{*}$	$0,0299^{**}$	$0,0401^{**}$

o maior estimador R/S(n) $= 0,1540$, para o parâmetro de diferenciação d. Observamos que a existência de pequena longa dependência nas seqüências da Tabela 5.6, é estatisticamente significante ao nível de 5% utilizando pelo menos um estimador proposto neste trabalho.

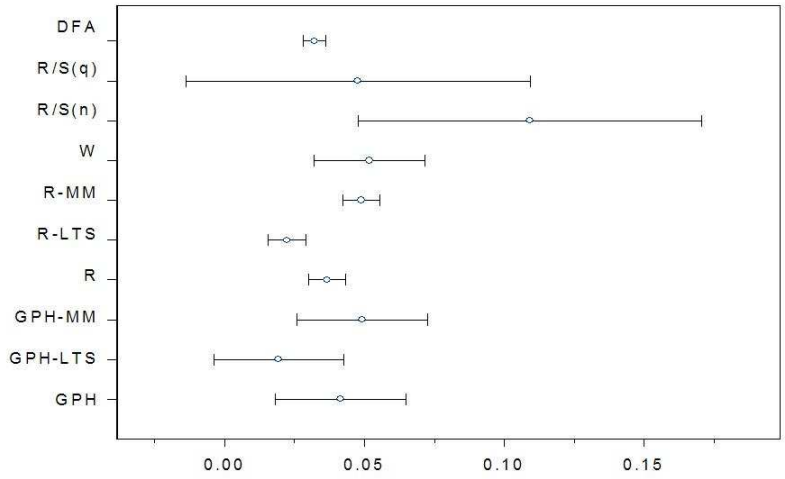

Figura 5.13: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência AF200828.

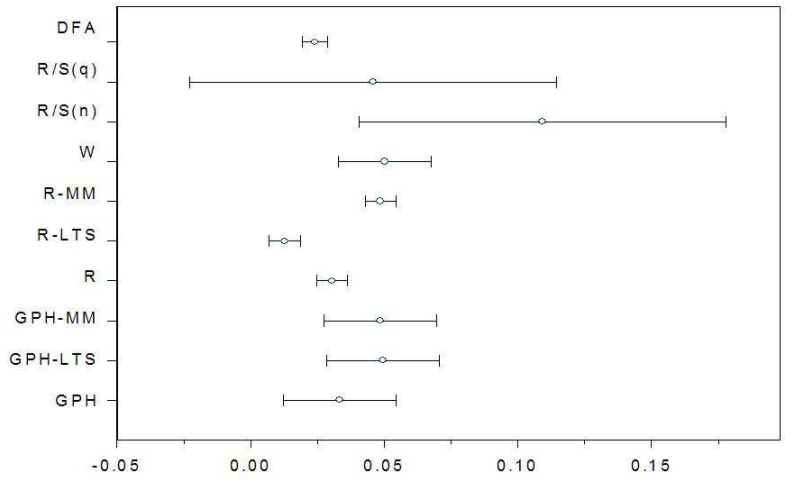

Figura 5.14: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência U3754.

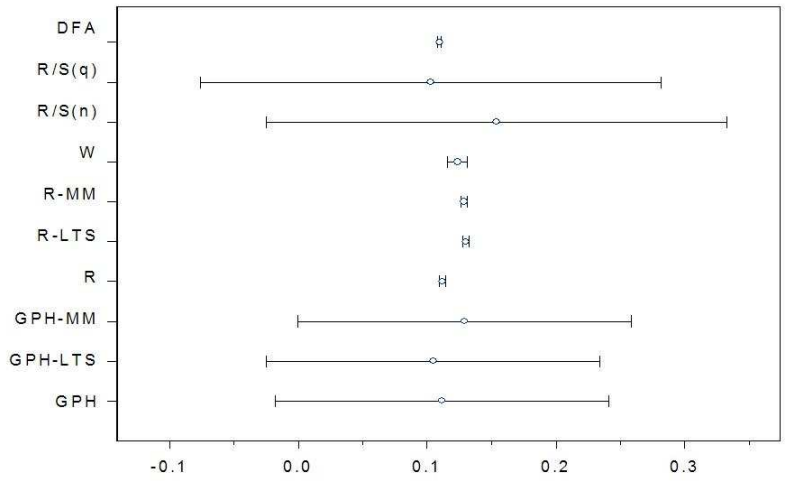

Figura 5.15: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência M94081.

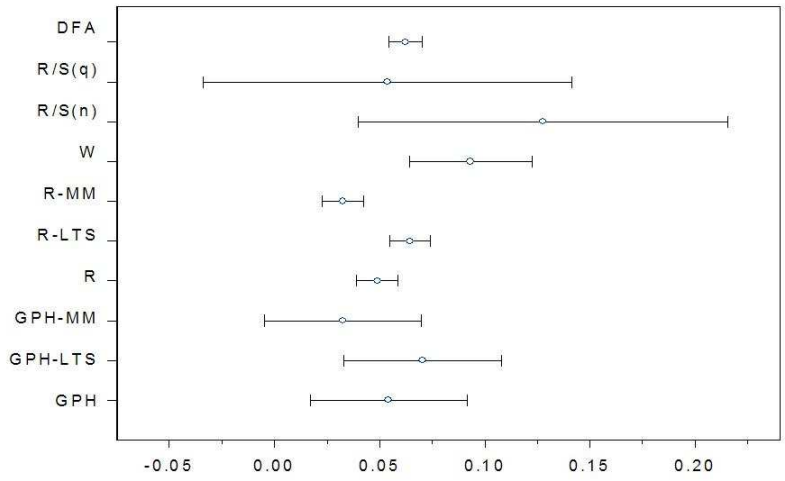

Figura 5.16: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência NM004023.

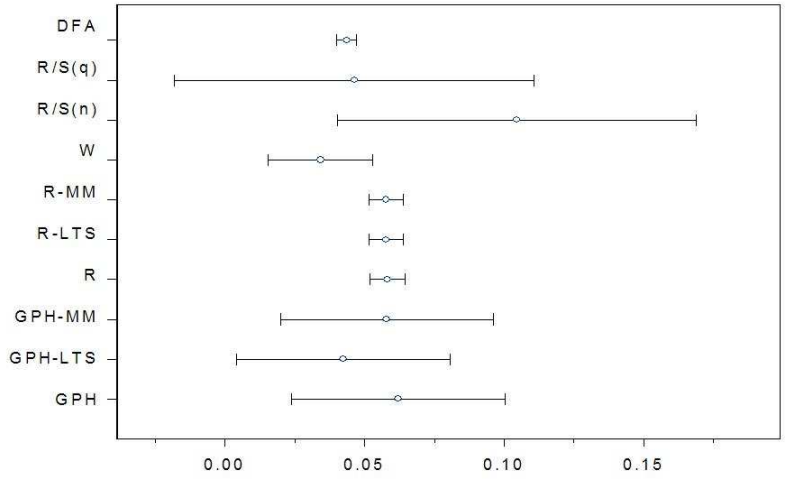

Figura 5.17: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência U96639.

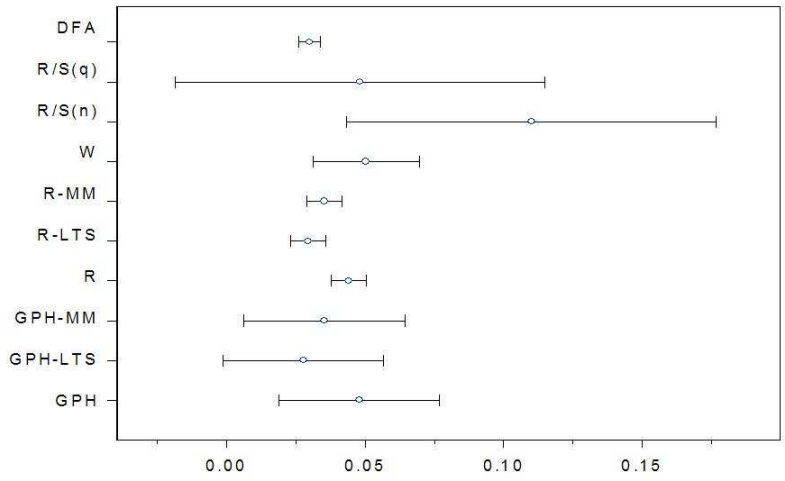

Figura 5.18: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência X03240.

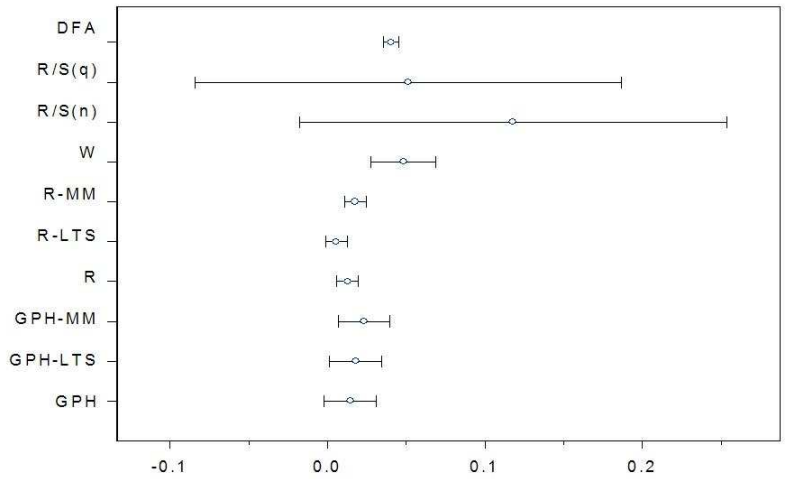

Figura 5.19: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência X54252.

5.4 Reino Plantae

Definição 5.4 (Reino Plantae). O *Reino Plantae, Metaphyta* ou *Vegetal* é um dos principais grupos em que se divide a vida na Terra (com cerca de 400.000 espécies conhecidas, incluindo uma grande variedade: ervas, árvores, arbustos, plantas microscópicas, etc). São, em geral, organismos autotróficos cujas células incluem um ou mais organelos especializados na produção de material orgânico a partir de material inorgânico e da energia solar: os cloroplastos.

A Figura 5.20 ilustra uma das 400.000 espécies de plantas conhecidas. Na Tabela 5.7 encontramos o código de acesso de uma seqüência de nucleotídeo no reino Plantae, do Instituto Europeu de Bioinformáticas (EBI).

A Tabela 5.8 apresenta a análise dos estimadores GPH, GPH-LS, GPH-LTS, GPH-MM, R, R-LTS, R-MM, W, R/S(n), R/S(q) e DFA para a seqüência de DNA apresentada na Tabela 5.7.

A Figura 5.21 apresenta, graficamente, os intervalos à 95% de confiança para o parâmetro de diferenciação d, da seqüência da Tabela 5.7.

Figura 5.20: *Tropaeolum majus.*

Tabela 5.7: Seqüência de Nucleotídeos no Reino Plantae.

Reino Plantae		
Local	**Código de Acesso**	**Tamanho**
EBI	BD006914	3.747

Na Tabela 5.8, observamos que o valor mínimo é obtido pelo estimador DFA e o valor máximo é obtido pelo estimador R/S(n). A existência de

Tabela 5.8: Estimadores do Parâmetro de Diferenciação e o seus respectivos níveis de significância em uma Seqüência de Nucleotídeos no Reino Plantae.

Reino Plantae	
Seqüência	BD006914
GPH	$0,0432^{*}$
GPH-LTS	$0,0635^{*}$
GPH-MM	$0,0467^{*}$
R	0,0349
R-LTS	0,0478
R-MM	0,0465
W	$0,0392^{**}$
R/S(n)	$0,1169^{*}$
R/S(q)	$0,0504^{*}$
DFA	$0,0475^{**}$

pequena longa dependência na seqüência BD006914 é estatisticamente significante ao nível de 10%, para os estimadores GPH, GPH-LTS, GPH-MM, W, DFA, R/S(n) e R/S(q).

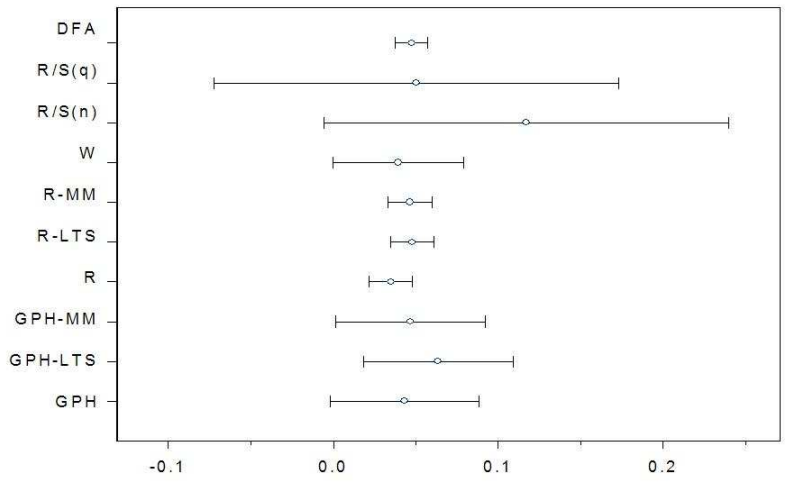

Figura 5.21: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência BD006914.

5.5 Reino Fungi

Definição 5.5 (Reino Fungi). O *Reino Fungi* é formado por seres uni e pluricelulares, eucariontes. Estão incluídos neste grupo organismos de dimensões consideráveis, como os cogumelos, o fungo amarelo, mas também muitas formas microscópicas, como bolores e leveduras.

A Figura 5.22 ilustra o fungo amarelo e na Tabela 5.9 encontramos códigos de acesso de seqüências de nucleotídeos no reino Fungi, do Instituto Europeu de Bioinformáticas (EBI, http://www.ebi.ac.uk/) e do Centro Nacional de Informação Biotecnológica (NCBI, http://www.ncbi.nlm.nih.gov/).

Figura 5.22: Fungo amarelo.

Tabela 5.9: Seqüência de Nucleotídeos no Reino Fungi.

Reino Fungi		
Local	**Código de Acesso**	**Tamanho**
NCBI	DQ449069	1620
NCBI	AY216992	1236
EBI	DQ207726	31103

A Tabela 5.10 apresenta a análise dos estimadores GPH, GPH-LS, GPH-LTS, GPH-MM, R, R-LTS, R-MM, W, R/S(n), R/S(q) e DFA para cada uma das três seqüências de DNA apresentadas na Tabela 5.9.

As Figuras 5.23 a 5.25 apresentam, graficamente, os intervalos à 95% de confiança para o parâmetro de diferenciação d, respectivamente, de cada seqüência da Tabela 5.9.

Tabela 5.10: Estimadores do Parâmetro de Diferenciação e os seus respectivos níveis de significância em Seqüências de Nucleotídeos no Reino Fungi.

Reino Fungi			
Seqüência	DQ449069	AY216992	DQ207726
GPH	$0,0753^{*}$	-0,0107	$0,0589^{**}$
GPH-LTS	$0,0497^{*}$	-0,0013	$0,0502^{**}$
GPH-MM	$0,0727^{*}$	-0,0665	$0,0455^{**}$
R	$0,0692^{**}$	0,0174	$0,0562^{**}$
R-LTS	$0,0506^{**}$	0,0196	$0,0504^{**}$
R-MM	$0,0659^{**}$	-0,0670	$0,0454^{**}$
W	0,0028	-0,0061	$0,0375^{**}$
R/S(n)	0,1084	0,0925	$0,1127^{**}$
R/S(q)	0,0461	0,0211	$0,0597^{**}$
DFA	$0,0242^{**}$	$0,0163^{*}$	$0,0402^{**}$

Analisando a Tabela 5.10 podemos observar que a seqüência AY216992, apresentou o menor estimador R-MM $= -0,067$ e a seqüência DQ207726 obteve o maior estimador R/S(n) $= 0,1127$, para o parâmetro de diferenciação d. Nota-se que para todas as seqüências da Tabela 5.9, o estimador R/S(n) obteve a maior estimativa para o parâmetro de diferenciação d. A existência de pequena longa dependência nas seqüências da Tabela 5.10, é estatisticamente significante ao nível de 5%, utilizando pelo menos um método de estimação proposto neste trabalho.

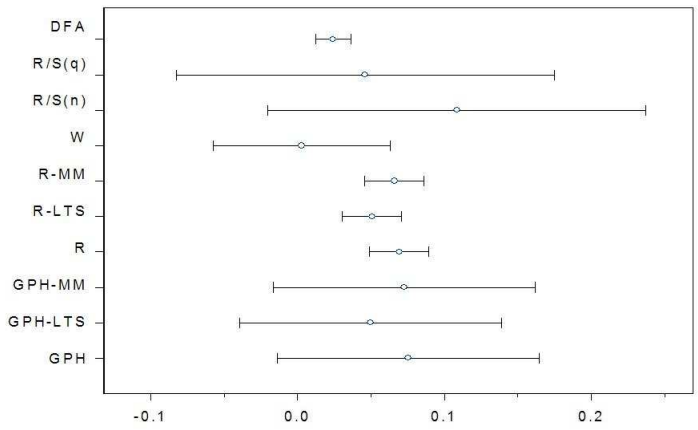

Figura 5.23: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência DQ449069.

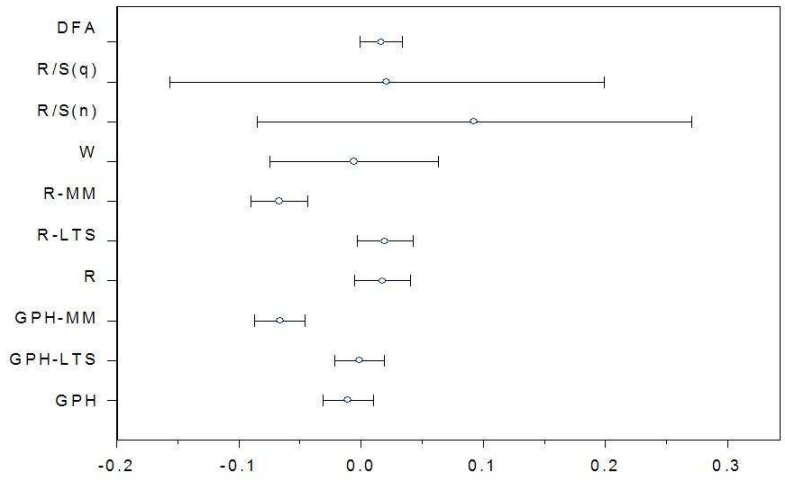

Figura 5.24: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência AY216992.

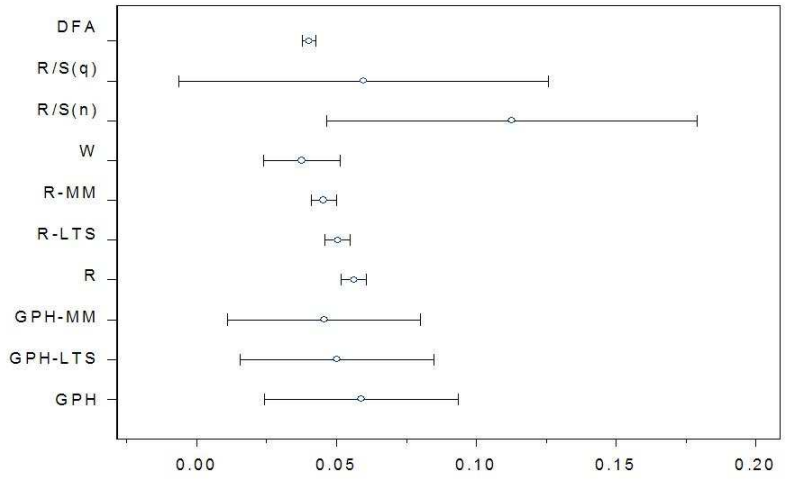

Figura 5.25: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência DQ207726.

5.6 Reino Protista

Definição 5.6 (Reino Protista). O *Reino Protista* ou *Protoctista* é um dos reinos biológicos comumente reconhecidos, inclui os seres unicelulares eucariontes, como é o caso dos protozoários e das algas unicelulares e pluricelulares que não possuem tecidos verdadeiros, como é o caso das algas multicelulares.

A Figura 5.26 ilustra um protozoário e na Tabela 5.11 encontramos códigos de acesso de seqüências de nucleotídeos no reino Protista, do Instituto Europeu de Bioinformáticas (EBI, http://www.ebi.ac.uk/) e do Centro Nacional de Informação Biotecnológica (NCBI, http://www.ncbi.nlm.nih.gov/).

A Tabela 5.12 apresenta a análise dos estimadores GPH, GPH-LS, GPH-LTS, GPH-MM, R, R-LTS, R-MM, W, R/S(n), R/S(q) e DFA para cada uma das três seqüências de DNA apresentadas na Tabela 5.11.

Figura 5.26: *Paramecium aurelia.*

Tabela 5.11: Seqüência de Nucleotídeos no Reino Protista.

Reino Protista		
Local	**Código de Acesso**	**Tamanho**
EBI	DQ851108	69066
NCBI	XM001025571	1043
NCBI	XM001019959	4822

Tabela 5.12: Estimadores do Parâmetro de Diferenciação e os seus respectivos níveis de significância em Seqüências de Nucleotídeos no Reino Protista.

Reino Protista			
Seqüência	DQ851108	XM001025571	XM001019959
GPH	$0,0303^{**}$	0,0256	0,0250
GPH-LTS	$0,0291^{**}$	0,0731	0,0071
GPH-MM	$0,0182^{**}$	0,0328	0,0040
R	$0,0289^{**}$	0,0602	0,0251
R-LTS	$0,0091^{**}$	0,0486	-0,0024
R-MM	$0,0182^{**}$	0,0322	0,0107
W	$0,0676^{**}$	$0,0667^{**}$	$0,0615^{**}$
R/S(n)	$0,1196^{**}$	0,1336	0,1216
R/S(q)	$0,0549^{**}$	0,0492	0,0495
DFA	$0,0205^{**}$	$0,1223^{**}$	$0,0508^{**}$

As Figuras 5.27 a 5.29 apresentam, graficamente, os intervalos à 95% de confiança para o parâmetro de diferenciação d, respectivamente, de cada seqüência da Tabela 5.11. Analisando a Tabela 5.12 podemos observar que a seqüência XM001019959, apresentou o menor estimador R-LTS $= -0,0024$ e a seqüência XM001025571 obteve o maior estimador R/S(n) $= 0,1336$, para o parâmetro de diferenciação d. A existência de pequena longa dependência nas seqüências da Tabela 5.12, é estatisticamente significante ao nível de 5%, para pelo menos um método de estimação proposto neste trabalho.

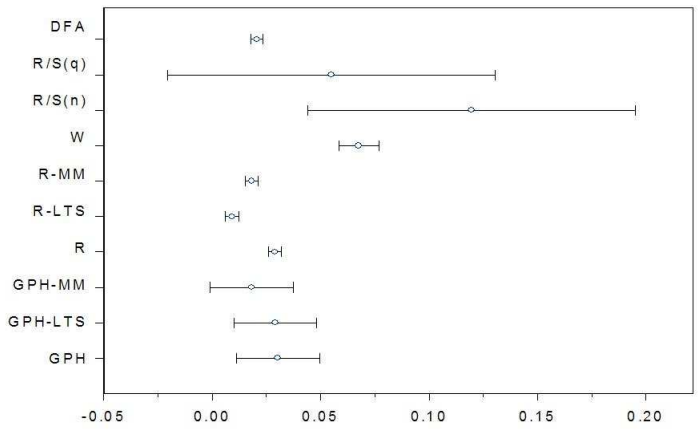

Figura 5.27: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência DQ851108.

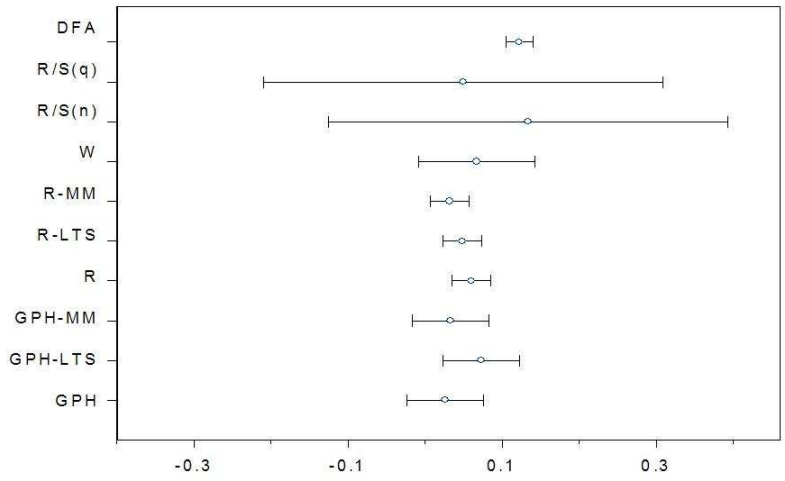

Figura 5.28: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência XM001025571.

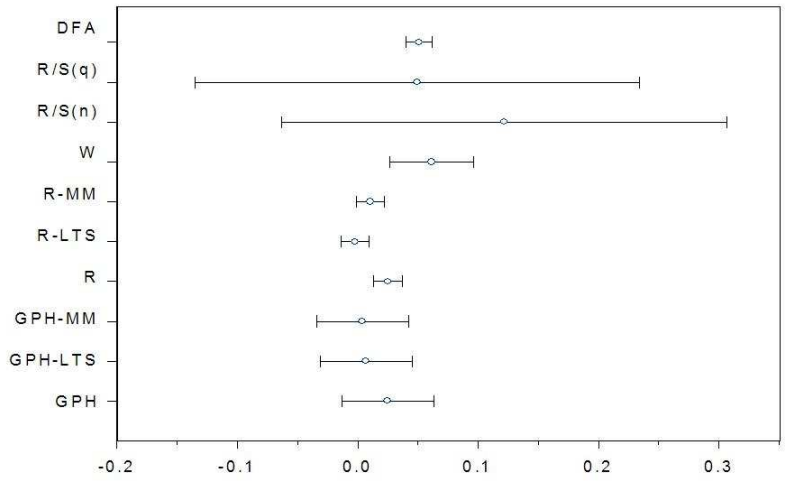

Figura 5.29: Intervalos à 95% de confiança para o parâmetro de diferenciação d da seqüência XM001019959.

Capítulo 6

Conclusão

Estudamos, neste trabalho, processos estocásticos com decaimento hiperbólico da função de autocorrelação, também denominados *processos com propriedade de longa dependência*.

Concentramos nosso estudo nas classes dos processos auto-regressivos médias móveis fracionariamente integráveis, denotados por ARFIMA(p, d, q).

Vimos que os processos ARFIMA(p, d, q) exibem *longa dependência* quando $d \in (0,0; 0,5)$, *curta dependência* quando $d = 0,0$ e *dependência intermediária* quando $d \in (-0,5; 0,0)$.

Estudamos diferentes métodos de estimação para o parâmetro de diferenciação d dentro das classes paramétrica e semiparamétrica.

Verificamos que podemos utilizar o método R/S(n) ("*Rescaled Range*"), proposto por Hurst (1951) e o método das análises de flutuações destendenciadas ("*Detrended Fluctuation Analysis*" - DFA), proposto por Peng et al. (1994) para estimar o parâmetro de diferenciação d, através da relação

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

em que α é o coeficiente de escala obtido pelo método DFA, para medir longa dependência e H é o parâmetro sugerido por Harold Edwin Hurst (1880-1978), também para medir longa dependência.

Descrevemos o método da análise de flutuações destendenciadas (DFA) e analisamos suas propriedades estatísticas. Mostramos que o método DFA tem como objetivo o cálculo de uma flutuação estatística $F(l)$, onde l representa o tamanho de uma janela, para mapear um conjunto de medidas. Variando o tamanho de l, as flutuações podem ser caracterizadas através de um expoente de escala obtido a partir da curva ajustada ao gráfico $\ln(F(l))$ versus $\ln(l)$. Mostramos que sob algumas suposições, o expoente de escala obtido pelo método DFA, é não viciado e consistente. Para aplicar o método DFA, divide-se a série temporal $\{X_t\}_{t=1}^{n}$ em blocos com l observações. Em cada bloco calcula-se as somas parciais $\{Y_t\}_{t=1}^{l}$, e então ajusta-se uma reta $Y_t^l = a + bt$. Mostramos que, se as variáveis aleatórias $Y_1 - Y_1^l, Y_2 - Y_2^l, \cdots, Y_{\widetilde{n}} - Y_{\widetilde{n}}^l$,

são independentes e identicamente distribuídas com função de distribuição comum $\mathcal{N}(0, \sigma_l^2)$, então $\mathrm{F}^2(l)$ tem função de distribuição $\Gamma\left(\frac{\widetilde{n}}{2}, \frac{\widetilde{n}}{2\sigma_l^2}\right)$, onde $\widetilde{n}$ é o maior múltiplo de l, inferior ou igual ao tamanho da amostra. Observamos que ${\sigma_l}^2$ é a variância teórica das variáveis aleatórias $Y_j - Y_j^l$, para $j = 1, \cdots, \widetilde{n}$. Provamos que $F^2(l)$ é não viciado para a variância σ_l^2 e se $0 < \sigma_l^4 < \infty$, a estatística $F^2(l)$ é consistente e de mínima variância quando $\widetilde{n}$ tende a infinito.

Para as seqüências de DNA em vírus analisadas neste trabalho, observamos que o estimador do parâmetro d é tal que $\widehat{d} \in (-0,03;\ 0,15)$. Para as seqüências de DNA no reino Monera, o estimador do parâmetro d é tal que $\widehat{d} \in (-0,09;\ 0,15)$. Para as seqüências de DNA no reino Animalia analisadas, o estimador do parâmetro d é tal que $\widehat{d} \in (-0,02;\ 0,16)$. Para a seqüência BD006914 no reino Plantae, todos os valores dos estimadores propostos neste trabalho pertencem ao intervalo $(0,0; 0,12)$. Para as seqüências de DNA no reino Protista analisadas, o estimador do parâmetro d é tal que $\widehat{d} \in (-0,003; 0,14)$. As vinte e duas seqüências de DNA analisadas, utilizando os métodos de estimação propostos neste trabalho, mostram a existência muito pequena de longa dependência. A existência de pequena longa dependência em cada uma destas seqüências, é estatisticamente significante ao nível de 10%, para pelo menos um estimador proposto neste trabalho.

6.1 Futuros Trabalhos

Vimos que podemos considerar diferentes transformações para obter uma série temporal a partir de uma seqüência de DNA (ver Seção 3.3). Neste trabalho utilizamos a regra RY (veja a transformação $g_1(\cdot)$ dada na expressão (4.1)) em seqüências de DNA, pois nosso maior objetivo era analisar o método das análises das flutuações destendenciadas (DFA). Sob algumas suposições, vimos que o estimador para o parâmetro de diferenciação d, obtido através do método DFA, é não viciado e consistente. Como futuros trabalhos pretendemos

- Utilizar a função dada pela expressão (4.2) para transformar uma seqüência de DNA em uma série temporal. Neste caso, será necessário estudar séries temporais multivariadas.

- Comparar os resultados dos estimadores para o parâmetro de diferenciação d, obtidos neste trabalho, com aqueles que serão obtidos utilizando outras transformações.

- Analisar as propriedades do método DFA, utilizando outras transformações para uma seqüência de DNA.

Referências

[1] Almeida, J.S.; J.A. Carriço, A. Maretzek, P.A. Noble e M. Fletcher (2001). "Analysis of Genomic Sequences by Chaos Game Representation". *Bioinformatics*, Vol. **17**(5), 429-437.

[2] Ben-Avrahan, D. e S. Havlin (2000). *Diffusion and Reactions in Fractals and Disordered Systems*. Cambridge: Cambridge University Press.

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

[4] Bhattacharya, R.N; V.K. Gupta e E. Waymire (1983). "The Hurst effect under trends". *J. Appl. Probab.*, Vol. **20**(3), 649-662.

[5] Bickel, P.J. e K.A. Doksum (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco: Holden-Day.

[6] Bisognin, C. e S.R.C. Lopes (2007). "Estimating and Forecasting the Long Memory Parameter in the Presence of Periodicity". *Journal of Forecasting*, Vol. **26**.

[7] Borstnik, B.; D. Pumpernik e D. Lukman (1993). "Analysis of Apparent $1/f^{\alpha}$ Spectrum in DNA Sequences". *Europhysics Letters*, Vol. **23**(6), 389-394.

[8] Box, G.E.P.; G.M. Jenkins e G.C. Reinsel (1994). *Time Series Analysis: Forecasting and Control*. New Jersey: Prentice Hall.

[9] Brockwell, P.J. e R.A. Davis (1991). *Time Series: Theory and Methods*. New York: Springer-Verlag.

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

[11] Buldyrev, S.V.; N.V. Dokholyan, A.L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley e G.M. Viswanathan (1998). "Analysis of DNA Sequences Using Methods of Statistical Physics". *Physica A*, Vol. **249**, 430-438.

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

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

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

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

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

[17] Fox, R. e 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.

[18] Garcia, J.A. e M.V. José (2005). "Mathematical properties of DNA sequences from coding and noncoding regions". *Revista Mexicana de Fisica*, Vol. **51**(2), 122-130.

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

[20] Giraitis, L; P. Kokoszka, R. Leipus e 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.

[21] Gradshteyn, I.S. e I.M. Ryzhik (2000). *Table of Integrals, Series e Products*. San Diego: Academic Press.

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

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

[24] Hosking, J. (1984). "Modelling Persistence in Hydrological Time Series using Fractional Differencing". *Water Resources Research*, Vol. **20**(12), 1898-1908.

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

[26] Hurst, H.E.; R.P. Black e Y.M. Simaika (1965). *Long-Term Storage: An Experimental Study*. London: Constable.

[27] Kantelhardt, J. W.; E. Koscielny-Bunde, H.H.A. Rego, S. Havlinb e A. Bunde (2001). "Detecting long-range correlations with detrended fluctuation analysis". *Physica A: Statistical Mechanics and its Applications*, Vol. **295**, 441-454.

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

[29] Kwiatkowski, D.; P.C.B. Phillips, P. Schmidt e Y. Shin (1992). "Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic series have a unit root?" *J. Econometrics*, Vol. **54**, 159-178.

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

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

[32] Lopes, S.R.C.; V.A. Reisen e B. Abraham (2001). "Estimation of Parameters in ARFIMA Processes: A Simulation Study". *Communications in Statistics: Simulation and Computation*, Vol. **30**(4), 787-803.

[33] Lopes, S.R.C.; B.P. Olbermann e V.A. Reisen (2004). "A Comparison of Estimation Methods in Non-stationary ARFIMA Processes". *Journal of Statistical Computation and Simulation*, Vol. **74**(5), 339-347.

[34] Lopes, S.R.C. e 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.

[35] Lopes, S.R.C. e A.S. Pinheiro (2006). "Wavelets for Estimating the Fractional Parameter in Non-Stationary ARFIMA Processes". Submetido.

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

[37] Lopes, S.R.C. (2007). "Topics on Long-Range Dependence". Em revisão.

[38] Lopes, S. (1996). *Biologia*. São Paulo: Editora Saraiva.

[39] Mandelbrot, B.B. e J.W. Van Ness (1968). "Fractional Brownian Motions, Fractional Noises and Applications". *SIAM Review*, Vol. **10**(4), 422-437.

[40] Newey, W. K. e K.D. West (1987). "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix". *Econometrica*, Vol. **55**, 703-708.

[41] Olbermann, B.P. (2002). *Estimação em Classes de Processos Estocásticos com Decaimento Hiperbólico da Função de Autocorrelação*. Tese de Doutorado, Programa de Pós-Graduação em Matemática, Instituto de Matemática, UFRGS, Porto Alegre.

[42] Olbermann, B.P.; S.R.C. Lopes e A.O. Lopes (2007). "Parameter Estimation in Manneville-Pomeau Processes". Submetido.

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

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

[45] Priestley, M.B. (1981). *Spectral Analysis in Time Series*. New York: Academic Press.

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

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

[48] Shlesinger, M.F.; B.J. West e J. Klafter (1987). "Lévy dynamics of enhanced diffusion: Application to turbulence ". *Phys. Rev. Lett*, Vol. **58**, 1100-1103.

[49] Sowell, F. (1992). "Maximum Likelihood Estimation of Stationary Univariate Fractionally Integrated Time Series Models". *Journal of Econometrics*, Vol. **53**, 165-188.

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

[51] Stoffer, D.S. e O. Rosen (2007). "Automatic Estimation of Multivariate Spectra via Smoothing Splines". *Biometrika*, Vol. **94**, 335-345.

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

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

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

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