Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Detectando longa dependência em sequências de DNA por mudança de regime

Raquel Romes Linhares [1]
Silvia Regina Costa Lopes [2]
Nuno Crato [3]

1 Introdução

Uma sequência de DNA é composta de bases A (adenina), C (citosina), T (timina) e G (guanina). Um dos principais resultados recentes é a existência de longa dependência em uma sequência de DNA (ver Crato et. al, 2011). Para aplicar métodos numéricos a uma sequência de DNA é necessário transformá-la em uma sequência numérica. Do ponto de vista biológico, é comum aplicar a regra SW "*strong-weak pairing*", a qual transforma C e G em 1 e A e T em 0. Na literatura existem diferentes transformações que aplicam métodos numéricos a uma sequência de DNA (ver Crato et al., 2011). Se considerarmos a abordagem de transformação numérica arbitrária para $\{A, C, T, G\}$ e depois utilizarmos a análise espectral, o resultado irá depender de cada transformação em particular. Portanto, não sabemos se a existência de longa dependência em sequência de DNA é real ou é induzida pela transformação. Nosso objetivo é analisar a longa dependência em sequência de DNA utilizando metodologia de mudança de regimes proposta por Liu (2000). Nesta metodologia, se a duração dos regimes de uma série temporal tem uma distribuição de cauda pesada com parâmetro $\alpha \in (1,2)$, então a série temporal apresenta a característica de longa dependência. Além disso, aplicando-se qualquer transformação linear que preserva a propriedade de variância finita na série temporal, igualmente preservará a propriedade de longa dependência. Portanto, nosso interesse é aplicar esta metodologia em sequências de DNA, para saber se a longa dependência nestas sequências é realmente uma propriedade destas sequências, ou ela é induzida pela transformação.

O trabalho está organizado da seguinte forma. Na Seção 2, apresentamos o o processo de mudança de regime e definimos a série temporal que representará uma sequência de DNA. Os resultados e discussões são apresentados na Seção 3. As conclusões finais são dadas na Seção 4.

[1]FAMAT - UFU. e-mail: *raquelromeslinhares@hotmail.com*
[2]PPGMAT - UFRGS.
[3]ISEG-Universidade Técnica de Lisboa.

2 Métodos

Nesta seção apresentamos o processo de mudança de regime, proposto por Liu (2000) com suas propriedades, e definimos a série temporal que representará uma sequência de DNA.

Definição 2.1 Sejam $w_0, w_1, \cdots$ variáveis aleatórias independentes e identicamente distribuídas, com média zero e variância finita σ^2, onde a variável aleatória w_k é constante no k-ésimo regime. Seja $I_k(\cdot)$ a função indicadora para o evento onde w_k sobrevive do período k até t, ou seja,

$$I_k(t) = \begin{cases} 1, & \text{se } k < t \leq k + T_k \\ 0, & \text{c.c,} \end{cases} \tag{1}$$

onde T_k é o tempo em que o k-ésimo regime termina, chamamos de *k-ésima duração de regime*. Assumimos que w_k é independente de T_k. Seja $\{W_t\}_{t\in N}$ um processo dado por

$$W_t = \sum_{k=0}^{\infty} w_k I_k(t), \quad t \in N, \tag{2}$$

então $\{W_t\}_{t\in N}$ é chamado *um processo com mudança de regime*.

Observação 2.1 A hipótese de média zero na Definição 2.1 é somente por conveniência analítica. Podemos eliminá-la sem fazer dano ao argumento.

O teorema a seguir mostra que, sob o mecanismo de mudança de regime com índice α, o processo com mudança de regime tem longa dependência com $d = 1 - \frac{\alpha}{2}$.

Teorema 2.1 *Seja $\{W_t\}_{t\in N}$ um processo com mudança de regime. Suponha que os tempos de mudança de regimes $\{T_k\}_{k\in N\cup\{0\}}$ são i.i.d. com distribuição de cauda pesada na forma de*

$$\lim_{t\to\infty} \frac{P(T_k > t)}{t^{-\alpha} h(t)} = 1, \tag{3}$$

onde $1 < \alpha < 2$ e $h(\cdot)$ é função com pouca variação. Então, o processo com mudança de regime tem longa dependência com parâmetro de longa dependência dado por $d = 1 - \frac{\alpha}{2}$.

O próximo teorema mostra que a longa dependência é preservada, quando é aplicada qualquer transformação linear que preserva a propriedade de variância finita no processo $\{W_t\}_{t\in N}$ com mudança de regimes.

Teorema 2.2 *Seja $\{W_t\}_{t=1}^{n}$ uma série temporal obtida de um processo com mudança de regimes $\{W_t\}_{t\in N}$. Suponha que as durações de regimes $\{T_k\}_{k\in N\cup\{0\}}$ são i.i.d. com função de distribuição com cauda pesada*

$$\lim_{t\to\infty} \frac{P(T_k > t)}{t^{-\alpha} h(t)} = 1, \tag{4}$$

onde $1 < \alpha < 2$ *e* $h(\cdot)$ *é função com pouca variação. Então, qualquer transformação linear que preserva a propriedade de variância finita no processo* $\{W_t\}_{t\in N}$ *com mudança de regimes, igualmente preserva a propriedade de longa dependência com mesma magnitude* $d = 1 - \frac{\alpha}{2}$.

Uma característica notável na distribuição de duração de regimes é o decaimento da cauda cuja taxa de decaimento é dada pelo parâmetro α. Diversos métodos vem sendo propostos para a estimação do parâmetro α. Geralmente, estão divididos em quatro categorias de métodos: baseados no estimador de Hill, baseados na máxima verossimilhança, baseados em quantis e baseados na função característica. Neste trabalho utilizamos os seguintes estimadores para o parâmetro α: o estimador de máxima verossimilhança $\hat{\alpha}_{\text{mle}}$ (ver Nolan, 2001); o estimador baseado na função característica empírica $\hat{\alpha}_{\text{fce}}$ (ver Press, 1972) e o estimador de regressão baseado na função característica empírica $\hat{\alpha}_{\text{reg}}$ (ver Koutrouvelis, 1981).

Definição 2.2 (Série Temporal de uma Sequência de DNA) Dada uma sequência de DNA $\{n_i\}_{i=1}^{n}$, a *série temporal* $\{W_t\}_{t=1}^{n}$, obtida desta sequência, é dada por

$$W_t = f(n_t), \tag{5}$$

onde $f(\cdot)$ é qualquer transformação linear que preserva a propriedade de variância finita.

Consideramos que a mudança de regime ocorre em uma sequência de DNA, quando há uma mudança na base da sequência. Seja $\{w_k\}_{k=1}^{m}$ os m diferentes regimes da sequência de DNA. Os regimes são relacionados a uma variável de estado latente w, que toma o valor w_k no k-ésimo regime. Seja $\{T_k\}_{k=1}^{m}$ *as durações de regimes* para a sequência de DNA. Na Seção 3, investigamos a propriedade de longa dependência, focalizando nossa atenção no parâmetro estável α da série temporal $\{T_k\}_{k=1}^{m}$, para várias sequências de DNA.

Exemplo 2.1 Seja uma parte de uma sequência de DNA de tamanho $n = 16$ dada por

$$\{n_i\}_{i=1}^{16} = \{\underbrace{\text{C}}_{w_1}\underbrace{\text{AAA}}_{w_2}\underbrace{\text{T}}_{w_3}\underbrace{\text{AAAAAA}}_{w_4}\underbrace{\text{C}}_{w_5}\underbrace{\text{A}}_{w_6}\underbrace{\text{C}}_{w_7}\underbrace{\text{A}}_{w_8}\underbrace{\text{T}}_{w_9}\}.$$

Considerando que a mudança de regime em uma sequência de DNA, ocorre quando há uma mudança na base nesta sequência, então os $\{w_k\}_{k=1}^{9}$ são os 9 diferentes regimes da sequência $\{n_i\}_{i=1}^{16}$. Observe que a série temporal das durações destes 9 regimes, para a sequência $\{n_i\}_{i=1}^{16}$, é dada por $\{T_k\}_{k=1}^{9} = \{1,3,1,6,1,1,1,1,1\}$.

3 Resultados e discussões

Neste capítulo analisamos a longa dependência em sequências de DNA utilizando a metodologia de mudança de regimes, proposta por Liu (2000).

Para analisar a longa dependência, consideramos sequências de DNA, apresentadas na Tabela 1, correspondentes a partes do cromossomo 21 (de AL163202 até AL163210 na Tabela 1) da espécie *Homo sapiens*, as partes do cromossomo X da espécie *Homo sapiens* (de AC004673 até AY286122 na Tabela 1) e a sequência completa do cromossomo 1 da espécie *Leishmania braziliensis* (AM494938). Estas sequências estão disponíveis em "*European Bioinformatics Institute*" (EBI, http://www.ebi.ac.uk/). Para algumas sequências apresentadas na Tabela 1, a Figura 1 mostra histogramas das durações de regimes $\{T_k\}_{k=1}^{m}$, das respectivas sequências de DNA. Podemos observar distribuição com cauda pesada em todos os gráficos da Figura 1. Pelo Teorema 2.1, dada uma série temporal $\{T_k\}_{k=1}^{m}$ de uma sequência de DNA, se $1 < \alpha < 2$, então a sequência de DNA apresenta a característica de longa dependência. E, pelo Teorema 2.2, qualquer transformação linear $f(\cdot)$ aplicada na sequência de DNA, que preserva variância finita do processo todo, igualmente preserva a propriedade de longa dependência. Para a Tabela 1, consideramos as seguintes notações: n o número de pares de bases da sequência de DNA; m o tamanho da série temporal das durações de regimes $\{T_k\}_{k=1}^{m}$.

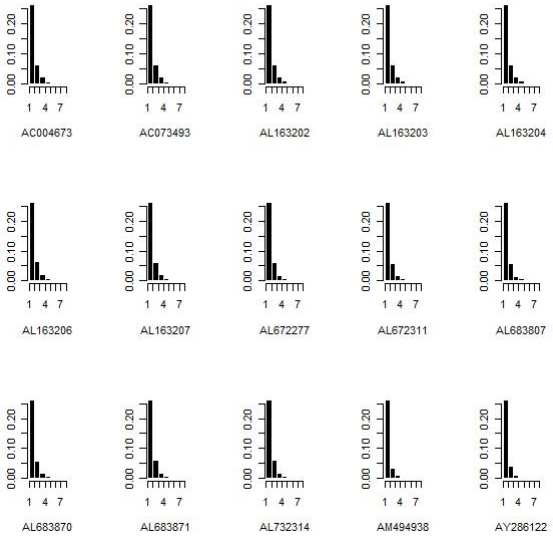

Figura 1: Histogramas das Séries Temporais de Durações de Regimes de Algumas Sequências de DNA Dadas na Tabela 1.

Podemos observar na Tabela 1, que $\hat{\alpha}_{\text{mle}} \in (1,2)$, $\hat{\alpha}_{\text{fce}} \in (1,2)$, $\hat{\alpha}_{\text{reg}} \in (1,2)$ e $\hat{\alpha}_{\text{wave}} \in (1,2)$ a nível de significância de 5%, para todas as sequências de DNA aqui estudadas. Portanto, pelo mecanismo de mudança de regime proposto por Liu (2000), as sequências de DNA, dadas na Tabela 1, apresentam a característica de longa dependência. Segue-se, pelo Teorema 2.2, que qualquer transformação linear $f(\cdot)$ aplicada na sequência de DNA, que preserva variância finita, igualmente preserva a propriedade de longa dependência da sequência de DNA. Portanto, a longa dependência nas vinte e quatro sequências de DNA da Tabela 1 é realmente propriedade da sequência e não da transformação utilizada.

Tabela 1: Estimadores $\hat{\alpha}_{mle}$, $\hat{\alpha}_{fce}$ e $\hat{\alpha}_{reg}$ para o Índice α, das Séries Temporais de Durações de Regimes, de Sequências de DNA.

Sequência	n	m	$\hat{\alpha}_{mle}$	$\hat{\alpha}_{fce}$	$\hat{\alpha}_{reg}$
AL163202	340000	236804	1.9459	1.7824	1.7814
AL163203	340000	237233	1.9440	1.7728	1.7731
AL163204	340000	238266	1.9597	1.8051	1.8035
AL163206	340000	236207	1.9514	1.7881	1.7881
AL163207	340000	237473	1.9530	1.7906	1.7908
AL163208	340000	237467	1.9545	1.7939	1.7940
AL163209	340000	237083	1.9537	1.7902	1.7900
AL163210	340000	235623	1.9434	1.7708	1.7704
AC004673	236281	164946	1.9521	1.7809	1.7797
AC073493	211422	148826	1.9482	1.7764	1.7744
AL672277	131682	93399	1.9461	1.7379	1.7390
AL672311	115998	83935	1.9526	1.7608	1.7546
AL683807	189825	135315	1.9464	1.7362	1.7336
AL683870	162377	117373	1.9595	1.7729	1.7723
AL683871	175765	126088	1.9478	1.7425	1.7423
AL732314	218723	156921	1.9505	1.7454	1.7431
AL445312	170984	123196	1.9687	1.8179	1.8187
AL450023	185532	131444	1.9510	1.7629	1.7617
AL591435	138038	95527	1.9432	1.7730	1.7726
AL929410	186649	130219	1.9457	1.7820	1.7836
Z98255	169998	118328	1.9476	1.7911	1.7915
AL732374	224187	156454	1.9381	1.7335	1.7316
AY286122	422021	327376	1.9906	1.8804	1.8805
AM494938	218723	180103	1.9576	1.7386	1.7392

4 Conclusões

Através da metodologia de mudança de regimes, proposta por Liu (2000), as sequências de DNA aqui analisadas, apresentaram a característica de longa dependência comprovando que esta é realmente propriedade da sequência e não da transformação utilizada.

Referências

[1] CRATO, N., LINHARES; R.R. e LOPES, S.R.C. α-Stable Laws for Noncoding Regions in DNA Sequences. **Journal of Applied Statistics**, v. 38(2), p. 261-271, 2011.

[2] KOUTROUVELIS, I.A. An Iterative Procedure for the Estimation of the Parameters of Stable Laws. **Communication in Statistics Simulation an Computation**, v. 10(1), p. 17-28, 1981.

[3] LIU, M. Modeling Long Memory in Stock Marcket Volatility". **Journal of Econometrics**, v. 99, p. 139-171, 2000.

[4] NOLAN, J.P. Maximum Likelihood Estimation of Stable Parameters. In O.E. Barndorff-Nielsen, T. Mikosch, and S.I. Resnick (Eds.), **Lévy Processes: Theory and Applications.** Boston: Birkhäuser, 2001.

[5] PRESS, S.J. "Estimation in Univariate and Multivariate Stable Distributions". **JASA**, v. 67, p. 842-846, 1972.