Universidade Estadual de Campinas

Bacharelado em Estatística

GILMAR FERNANDES VAZ PIRES

DANILO DE ATHAYDE ARFELLI

BRUNO DA COSTA PEIXOTO

JOSUÉ NIERI

CAMPINAS

2017

Sumário

1.Introdução

DNA é a sigla para ácido desoxirribonucleico, um aglomerado de moléculas que contém o material genético dos seres vivos. Toda a informação genética de um organismo está armazenada no DNA e estas informações são também transmitidas aos seus descendentes. Responsável por orientar as células na produção de proteínas, o DNA tem importância para toda a formação e funcionamento dos seres vivos.

A molécula de DNA é composta por uma fita dupla de nucleotídeos formado por 4 letras – A, T, C e G, que representam as quatro bases nitrogenadas: adenina, timina, citosina e guanina. As várias e diferentes combinações desses compostos (que podem alcançar mais de 3 bilhões em cada célula) produzem a variabilidade dos seres vivos.

Os cromossomos são longas sequências de DNA que contêm diversos genes e outras sequências de nucleotídeos. Em animais e na maior parte das plantas, organismos que se reproduzem sexuadamente, os cromossomos se apresentam aos pares, sendo estes organismos chamados de diplóides, também conhecidos como *2n*. Nos diplóides, os pares de cromossomos são denominados "cromossomos homólogos", pois estes têm sequências de DNA geralmente iguais, podendo exibir pequenas variações.

Anomalias cromossômicas podem aparecer por erros durante a segregação celular ocasionando o surgimento de síndromes genéticas. Tais síndromes podem ocorrer por alteração do número, do tamanho ou do ordenamento de partes dos cromossomos. Que será o nosso enfoque neste trabalho, descobrir deleções ou duplicações nos cromossomos.

2.Banco de dados

Os dados fornecidos apresentam informações sobre quantidade de moléculas presentes em 23 pares de cromossomos, separados por indivíduos e que são provenientes de amostras coletadas no ambulatório de neurologia da Universidade Estadual de Campinas (Unicamp).

Escolheu-se trabalhar apenas com um único cromossomo da amostra, para então reproduzir-se o mesmo procedimento para os outros 22, tal escolha é justificada assumindo independência entre os cromossomos, e o selecionado foi o número 20 (indicado pelo cliente como de menor tamanho).

Os 103 indivíduos foram renomeados para não serem expostos, e para a maior parte de nossas análises levamos em consideração o indivíduo "Array_Italia1_1.CEL.Log.R.Ratio" devido a qualidade visual de suas amostras para o nosso caso.

3.Objetivo

Objetiva-se desenvolver uma metodologia para identificação e remoção de padrões de onda. Com isso será possível diferir ocorrências de deleções ou duplicações cromossômicas. É desejável estabelecer critérios para definir a qualidade da amostra.

4.Metodologia

Para a identificação das ocorrências de delação será aplicado o modelo de Análise Bayesiana de Pontos de *Mudança* (BCP), que calcula a probabilidade de haver uma mudança de regime em cada posição. O método é uma implementação daquele descrito por Barry and Hartingan (1993) baseado no produto de partições. Este modelo foi escolhido pois modelos de séries temporais e não paramétricos supõem uma vizinhança densa, o que não ocorre no caso. Este modelo se adequa bem com o problema, "O bcp detecta com precisão os pontos de mudança e estima os segmentos das médias" (Chandra Erdman & John W. Emerson, *Bioinformatics* 2008)

O modelo assume que existe uma partição ρ desconhecida no conjunto dos dados em blocos contínuos tal que a média é constante dentro dos blocos, mas que pode diferir entre os blocos. É assumido que em cada bloco $X_1...X_n$, são independentes tendo densidade $N(\mu_i , \sigma^2)$ A distribuição a priori de μ_{ij} (a média do bloco começando na posição i + 1 e terminando na posição j) é escolhida como $N(\mu_0, \sigma^2_0 /(j - i))$.

O algoritmo usa uma partição $\rho = (U_1, U_2, ..., U_n)$, onde $U_i = 1$ indica um ponto de mudança na posição i + 1; Inicializa-se U_i em 0 para todos i <n, com $U_n \equiv 1$. Em cada etapa do MCMC em cada posição i um valor de U_i é extraído da distribuição condicional de U_i dados os dados e a partição atual. Seja b o número de blocos obtidos se $U_i = 0$, condicional em U_j, para $i \neq j$. A probabilidade de transição, p, para a probabilidade condicional de um ponto de mudança na posição i+1 pode ser obtida pela razão simplificada apresentada em Barry e Hartigan:

$$\frac{p_i}{1-p_i} = \frac{P(U_i=1|X,U_j,j\neq i)}{P(U_i=0|X,U_j,j\neq i)} = \frac{\left[\int_0^\gamma p^b(1-p)^{n-b-1}dp\right]\left[\int_0^\lambda \frac{w^{b/2}}{(W_1+B_1 w)^{(n-1)/2}}dw\right]}{\left[\int_0^\gamma p^{b-1}(1-p)^{n-b}dp\right]\left[\int_0^\lambda \frac{w^{(b-1)/2}}{(W_0+B_0 w)^{(n-1)/2}}dw\right]}$$

Onde W_0, B_0, W_1 e B_1 são as somas de quadrados, dentro e entre blocos, obtidas quando $U_i = 0$ e Ui=1, respectivamente, e X são os dados. Os parâmetros de ajuste γ e λ podem tomar valores em [0, 1], escolhidos para que este método "seja efetivo em situações em que não existam muitas mudanças (γ pequenos) e onde as alterações que ocorrem são de tamanho razoável (λ pequeno) "(Barry e Hartigan 1993, p. 312). Após cada iteração, as médias a posteriori são atualizadas de acordo com a partição atual. Para melhorar a eficiência numérica do método a razão do ponto de quebra pode ser re-expressada como:

$$\left(\frac{W_0}{W_1}\right)^{\frac{n-b-2}{2}}\left(\frac{B_0}{B_1}\right)^{\frac{b+1}{2}}\sqrt{\frac{W_1}{B_1}}\frac{\int_0^{\frac{B_1\lambda/W_1}{1+B_1\lambda/W_1}} p^{(b+2)/2}(1-p)^{(n-b-3)/2}dp \int_0^{\gamma} p^{b}(1-p)^{n-b-1}dp}{\int_0^{\frac{B_0\lambda/W_0}{1+B_0\lambda/W_0}} p^{(b+1)/2}(1-p)^{(n-b-2)/2}dp \int_0^{\gamma} p^{b-1}(1-p)^{n-b}dp}$$

A escolha dos parâmetros de ajuste é feita por inspeção visual, escolhidos de forma que o modelo atribui probabilidade alta para pontos de quebra e não para outliers.

5.Análise descritiva

No gráfico de pontos ordenado da “Figura 1”, pode-se notar grandes espaçamentos entre os dados, mesmo olhando as posições mais de perto, uma vez que são 43 mil dados dispersos em 62 milhões de posições. Sua maior concentração está no intervalo [-1,1] e na posição próxima a 1,5 milhão há uma região em que esse intervalo está entre [0,2], aparentando uma mudança de regime, ou seja, uma possível duplicação.

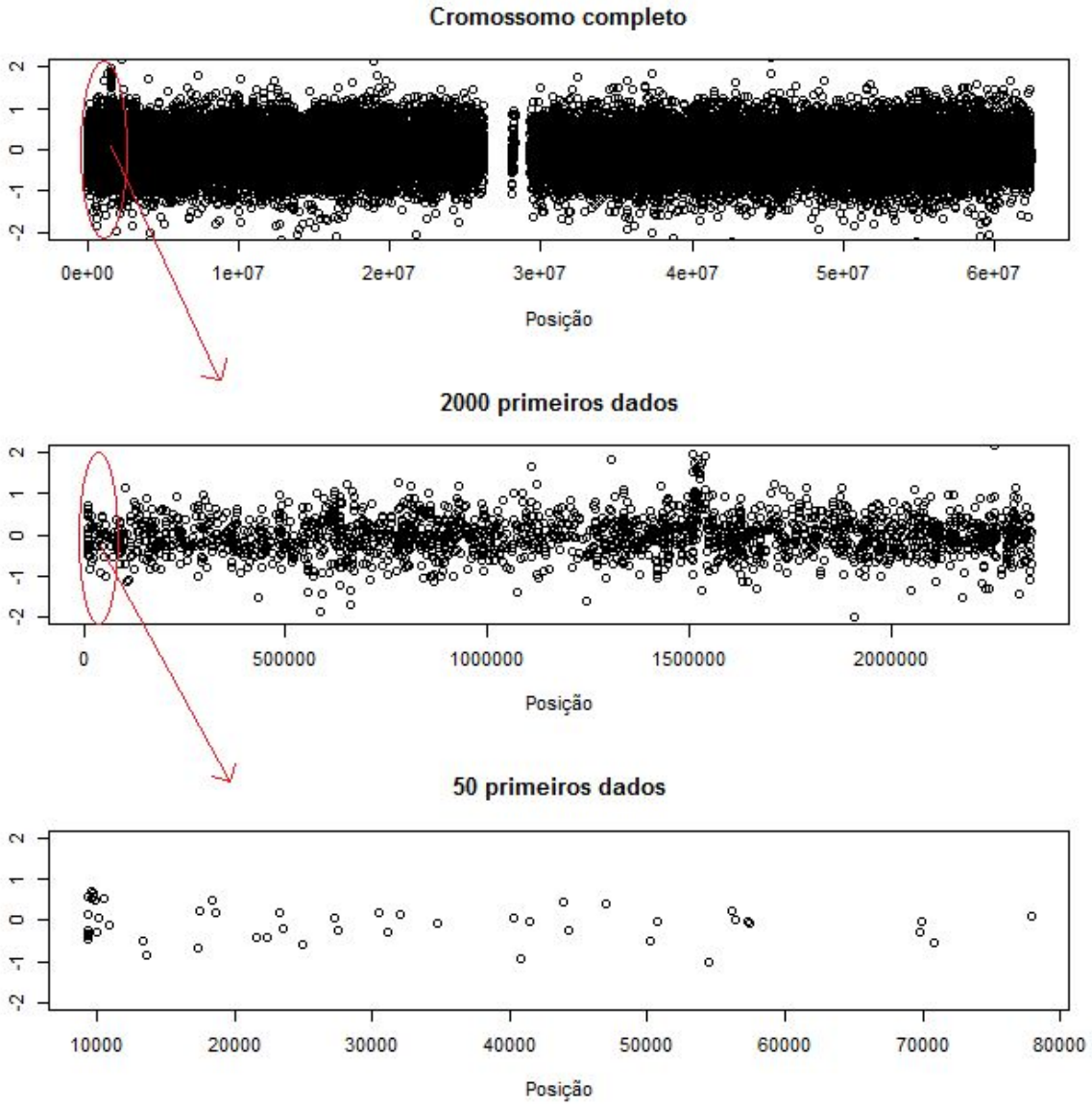


Figura 1: Gráficos de pontos em que são apresentados aproximadamente 43 mil dados do cromossomo ordenado sequencialmente em 62 milhões de posições, onde pontos estatisticamente diferentes de zero significam que houve duplicação ou deleção molecular.

Ignorando o espaçamento dos dados, na função de autocorrelação (Figura 2), há duas defasagens não nulas, porém seus valores são abaixo de 0,1, ou seja, possuem uma correlação muito baixa. Isso indica pouca dependência entre suas posições anteriores, além de que olhando as posições 3 à 3, os dados não são correlacionados entre si.

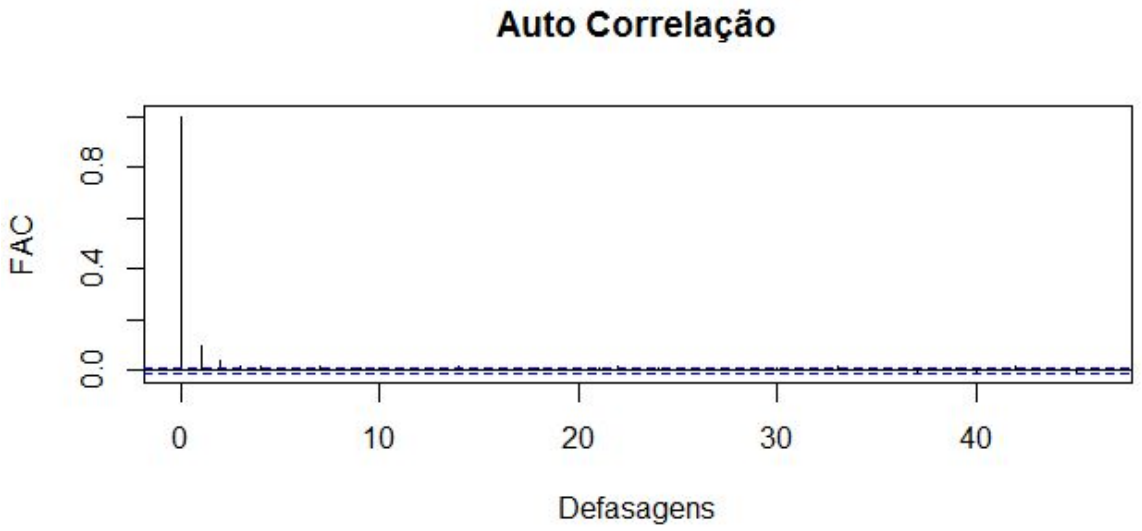


Figura 2: Função de autocorrelação ignorando o espaçamento dos dados. A função de autocorrelação mede o grau de correlação de uma dada posição com a posições posteriores.

Foi verificado o histograma de diversos indivíduos, e, na "Figura 3", pode-se observar o histograma de 8 deles, onde que para todos os casos há simetria e distribuição centrada em zero, dando indícios de um comportamento normal padrão $N(0, \sigma^2)$.

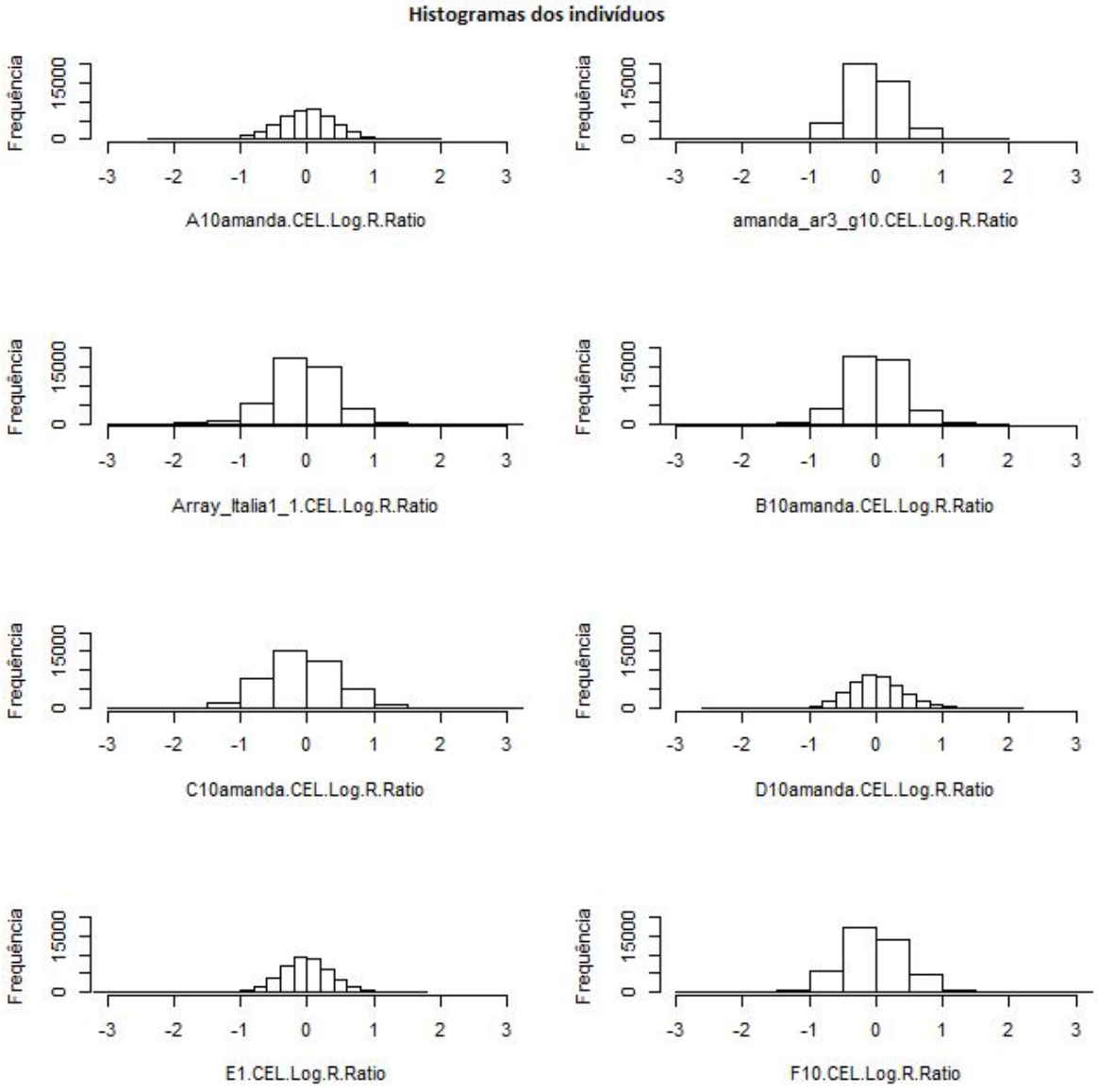


Figura 3: Histogramas de 8 amostras de indivíduos diferentes. Os dados aparentam ter distribuição normal centrada em zero ($N(0, \sigma^2)$).

6. Modelagem dos dados

A técnica é implementada a partir do pacote bcp, disponível no CRAN, como já citado na metodologia, uma implementação do modelo descrito por Barry and Hartingan (1993). A função principal *bpc,* retorna um vetor contendo a média, variância e probabilidade a posteriori de um ponto de quebra em cada posição.

Foi desconsiderado o espaçamento entre as posições, pois acredita-se que não há grandes perdas utilizando o método bcp. Também foi considerado independência e normalidade dos dados, por suas autocorrelações serem quase nulas (vide Figura 2) e os histogramas, simétricos e centrados em zero (vide Figura 3). O modelo foi ajustado para algumas amostras pré escolhidas, e pode ser estendida para outros cromossomos, desde que seja verificado a suposição de normalidade e independência.

O parâmetro de ajuste, p_0, foi escolhido levando em conta o baixo número esperado de ocorrências de pontos de quebra, assim também deve ser escolhido p_0 pequeno. Ao ajustar o modelo para o cromossomo completo, $p_0 = 0.0001$, trouxe resultados satisfatórios. O modelo captou mudanças de regime e deu probabilidade nula ou irrelevante a outliers, como pode ser visto na "Figura 4" e "Figura 5".

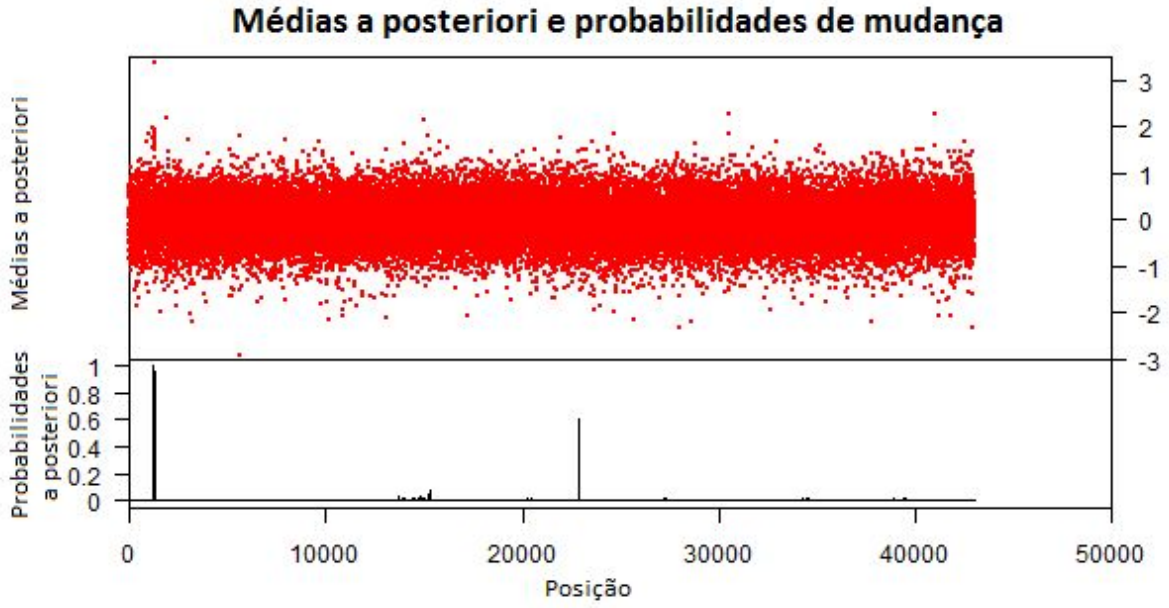


Figura 4: Médias a posteriori e probabilidade de mudança. Probabilidades altas mostram que existem mudanças na média naquela região. Médias superiores indicam duplicação e médias inferiores, deleção.

O gráfico de probabilidade indica possíveis mudanças na média, onde na região entre 1.000 e 1.500 há dois pontos de quebra com probabilidades acima de 95% e na região entre 22.850 e 22.900 as probabilidades são de 28% e 60%. Na "Figura 5", esses pontos de quebra podem ser visualizados com mais facilidade, onde na região entre 1.000 e 1.500 há claramente um aumento na média, já na região com menor probabilidade, não é tão evidente essa mudança.

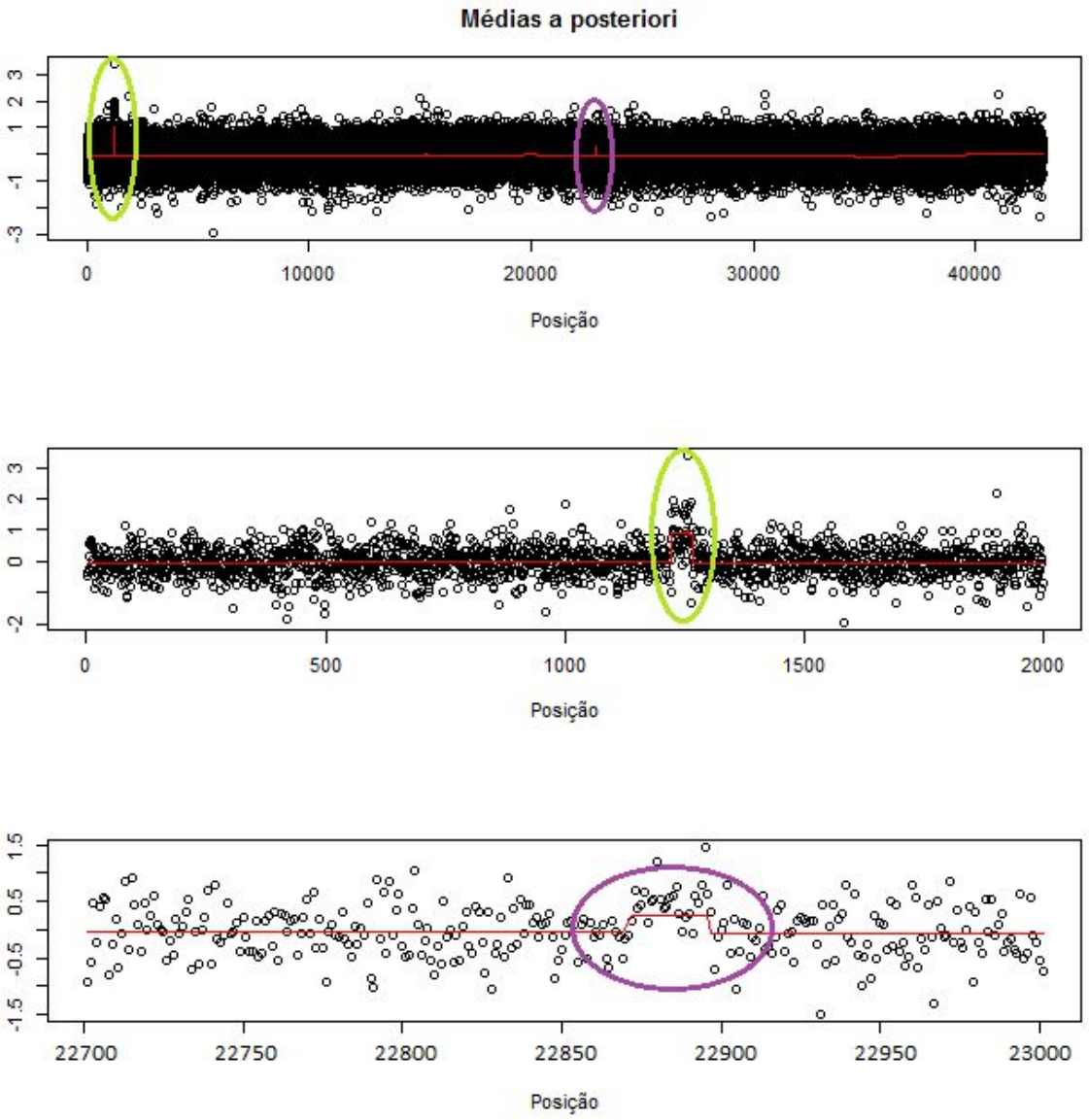
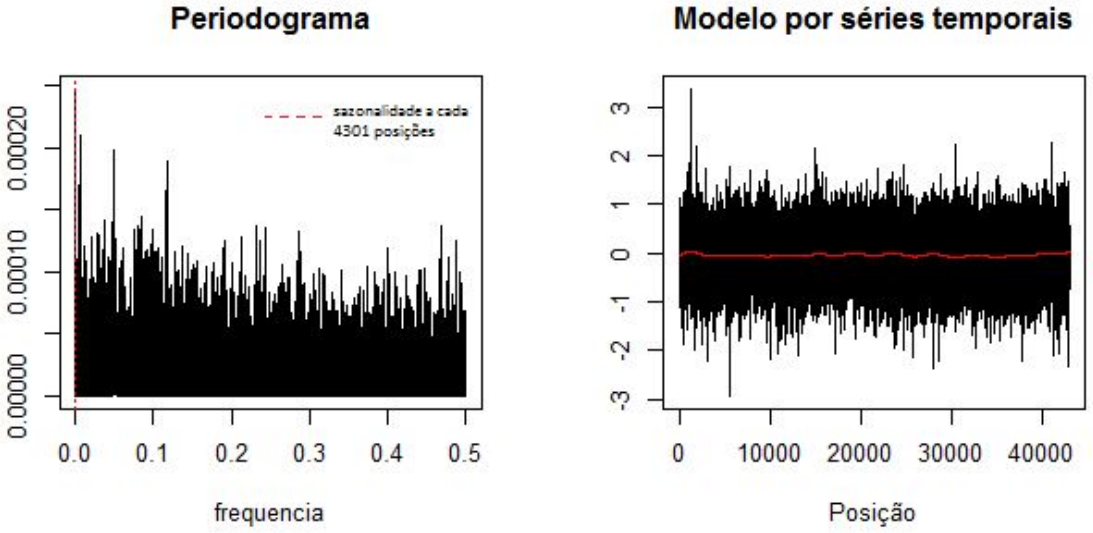


Figura 5: Médias a posteriori, com indicações onde houve mudanças de regime. O primeiro gráfico exibe todas as observações do indivíduo e destacam-se os dois picos de probabilidade, que ampliados resultam no segundo e terceiro gráfico.

7. Discussão

Inicialmente por ter sido desconsiderado os espaçamentos entre as posições, foram ajustados modelos de séries temporais. Analisando o indivíduo "Array_Italia1_1.CEL.Log.R.Ratio", foi detectado tendência e sazonalidade, porém pela "Figura 6" e "Figura 7", ao ajustar o modelo com essas características empíricas, embora todos os parâmetros fossem significativos, os resíduos não possuíam comportamento adequados.



Coeficientes:

	Estimativa	Erro padrão	valor t	p-valor	
t	-2.123e-05	1.760e-06	-12.062	< 2e-16	***
t.2	2.368e-09	2.249e-10	10.525	< 2e-16	***
t.3	-9.093e-14	8.959e-15	-10.150	< 2e-16	***
t.4	1.117e-18	1.117e-19	9.999	< 2e-16	***
cos(2 * pi * t * (1/4301))	-5.884e-02	1.513e-02	-3.890	0.000100	***
sin(2 * pi * t * (1/4301))	7.347e-02	1.582e-02	4.645	3.42e-06	***
t:cos(2 * pi * t * (1/4301))	1.795e-05	4.846e-06	3.705	0.000212	***
t:sin(2 * pi * t * (1/4301))	-2.419e-05	5.153e-06	-4.695	2.67e-06	***
t.2:cos(2 * pi * t * (1/4301))	-1.661e-09	4.577e-10	-3.628	0.000286	***
t.2:sin(2 * pi * t * (1/4301))	2.232e-09	4.912e-10	4.544	5.54e-06	***
t.3:cos(2 * pi * t * (1/4301))	5.345e-14	1.599e-14	3.342	0.000832	***
t.3:sin(2 * pi * t * (1/4301))	-7.786e-14	1.733e-14	-4.492	7.07e-06	***
t.4:cos(2 * pi * t * (1/4301))	-5.585e-19	1.845e-19	-3.027	0.002472	**
t.4:sin(2 * pi * t * (1/4301))	9.055e-19	2.021e-19	4.480	7.48e-06	***

Figura 6: Periodograma e modelo ajustado. No periodograma, foi detectado um pico alto que divide os dados em 10 partes, ou seja, há uma possível sazonalidade a cada 4301 posições. O modelo foi ajustado considerando tendência com polinômio de quarto grau e variáveis harmônicas que captam sua sazonalidade.

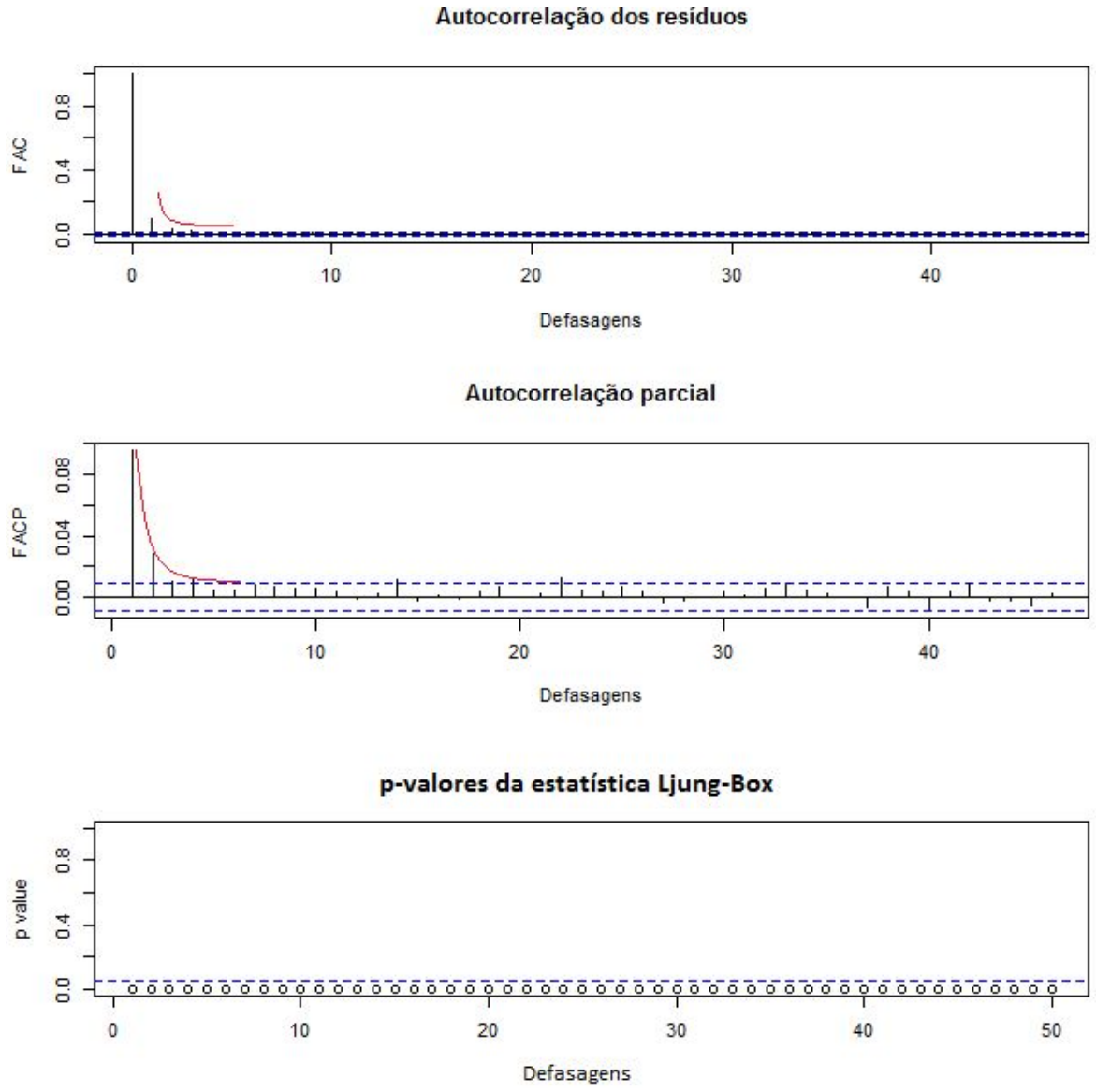


Figura 7: Análise dos resíduos. As autocorrelações e autocorrelações parciais são parecidas ao dos dados, indicando dependência entre os resíduos, mesmo ela sendo baixa. Pelo teste Ljung-Box conclui-se que os resíduos não possuem comportamento de ruído branco, por seus p-valores serem nulos, indicando que o modelo não é adequado.

Se levado em conta que as autocorrelações e/ou autocorrelações parciais terem um decaimento exponencial, dois modelos são candidatos para o ajuste dos dados, um sendo com comportamento auto regressivo AR(4), e outro tendo médias móveis com pelo menos 3 defasagens MA(3>). Porém ao tentar ajustar o modelo inicial considerando esses dois tipos de comportamento, houve problema na matriz hessiana por possuir valores muito próximos de zero. Outro problema encontrado ao usar séries temporais, foi ao tentar analisar outros indivíduos, onde a sazonalidade encontrada nunca era a mesma.

Em seguida, métodos não paramétricos (lowess e splines) foram candidatos para a modelagem dos dados, porém consultando especialistas com conhecimentos práticos na área,

foi alertado que se desconsiderado o espaçamento entre as posições, a penalidade na análise seria próxima a métodos de séries temporais, por depender de posições vizinhas.

Uma abordagem sugerida que não é tão penalizada, foi o método bayesiano de detecção de pontos de quebra (bcp). Esse método não identifica o padrão de onda, porém como o intuito de remoção da onda seria para detectar pontos de deleção e duplicação, e, o método bcp identifica pontos de quebra que definem essas mudanças de regime, não houve a necessidade de identificá-las.

8. Conclusão

Pela técnica de detecção de pontos de quebra, os espaçamentos entre as posições possuem menos influência do que técnicas de séries temporais e não paramétricas. Embora não tenha sido detectado o padrão de onda, através do bcp é possível identificar regiões de deleção e duplicação sem essa necessidade.

Analisando o indivíduo "Array_Italia1_1.CEL.Log.R.Ratio", na análise descritiva, foi identificado uma possível duplicação vista na "Figura 1". Já pelo método bcp, duas duplicações são destacadas, uma com probabilidade muito alta e outra com probabilidade de 60%. Um critério para estabelecer se houve realmente uma mudança de regime seria definir uma probabilidade mínima, onde valores superiores a essa probabilidade indicam que realmente houve mudança na média. Vale ressaltar que a análise pode ser estendida para outros cromossomos, desde que seja verificado a suposição de normalidade e independência.

9. Bibliografia

Alberts B., Bray D., Johnson A., Lewis J., Raff M., Roberts K., Walter P., Fundamentos da Biologia Celular. 3ª edição. Artmed,Porto Alegre, 2011.

Xiaofei Wang and John W. Emerson (2015)
Bayesian Change Point Analysis of Linear Models on General Graphs. Working Paper.

Chandra Erdman, John W. Emerson (2007), bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems, Journal of Statistical Software, 23(3), 1-13, URL http://www.jstatsoft.org/v23/i03/.

Chandra Erdman, John W. Emerson (2008)
A Fast Bayesian Change Point Analysis for the Segmentation of Microarray Data Bioinformatics, 24(19), 2143--2148, URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btn404.

Barry, Daniel; Hartigan, J. A. Product Partition Models for Change Point Problems. Ann. Statist. 20 (1992), no. 1, 260--279.doi:10.1214/aos/1176348521. http://projecteuclid.org/euclid.aos/1176348521.

Daniel Barry and J. A. Hartigan
Journal of the American Statistical Association
Vol. 88, No. 421 (Mar., 1993), pp. 309-319

Mauricio Zevallos, Séries Temporais: Notas de Aula (2015)