LEILA MARIA FERREIRA

EVALUATION OF GENOME SIMILARITIES: A WAVELET-DOMAIN APPROACH

LAVRAS – MG

2019

LEILA MARIA FERREIRA

EVALUATION OF GENOME SIMILARITIES: A WAVELET-DOMAIN APPROACH

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutora.

Profa. Dra. Thelma Sáfadi

Orientadora

LAVRAS – MG

2019

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Ferreira, Leila Maria
Evaluation of genome similarities: a wavelet-domain approach / Leila Maria Ferreira. – 2019.
89 p. : il.

Orientadora: Profa. Dra. Thelma Sáfadi.

Tese (doutorado) – Universidade Federal de Lavras, 2019.
Bibliografia.

1. Transformada não-decimada de ondaletas. 2. Genoma. 3. *Mycobacterium tuberculosis*. I. Sáfadi, Thelma. II. Título.

LEILA MARIA FERREIRA

EVALUATION OF GENOME SIMILARITIES: A WAVELET-DOMAIN APPROACH

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutora.

APROVADA em 07 de fevereiro de 2019.

Profa. Dra. Alessandra Querino da Silva	UFGD
Prof. Dr. Júlio Sílvio de Souza Bueno Filho	UFLA
Prof. Dr. Moysés Nascimento	UFV
Prof. Dr. Renato Ribeiro de Lima	UFLA

Profa. Dra. Thelma Sáfadi
Orientadora

LAVRAS – MG
2019

Aos meus pais Francisca e João,
pelo incentivo de sempre continuar nos estudos,
aos meus irmãos Juliano e Samuel,
pelo apoio e suporte nos momentos difíceis.

DEDICO

AGRADECIMENTOS

Primeiramente a Deus, socorro bem presente em todos os processos vividos durante esses 4 anos de doutorado. Muitas lutas foram enfrentadas, onde o medo e a insegurança eram muito presentes, mas em todas as adversidades eu contava com a presença de Deus, me sustentando e dando forças para seguir até o final.

Aos meus pais, minha mãe Francisca Borges Ferreira e meu pai João Batista Ferreira, por sempre acreditarem no meu potencial, pelo suporte financeiro nos estudos e pelas orações.

Ao meu irmão Samuel Lino Ferreira, pelo incentivo. A minha cunhada Mônica pela amizade e palavras de apoio. Ao meu sobrinho Matheus que nasceu no ano de 2016, trazendo muito alegria para nossa família.

Ao meu irmão Juliano Lino Ferreira, grande responsável por eu ter chegado até aqui. Ele foi o primeiro da família a conseguir chegar ao título de Doutor. Ele nunca media esforços para me ajudar a realizar meus objetivos. Em julho de 2017 enfrentamos uma grande provação, mas Deus por sua infinita misericórdia e bondade nos deu vitória.

A Professora Thelma pela orientação, sempre disponível para tirar dúvidas, pelos ensinamentos, compreensão e condução na estruturação da tese.

Ao Professor Renato Ribeiro de Lima, pelos ensinamentos e conselhos. Ele foi o responsável pela orientação na iniciação científica, onde tive o primeiro contato com a pesquisa.

Aos colegas do departamento em Estatística e Experimentação Agropecuária. Em especial a minha turma de doutorado: Laís, Thais, Michele, Tatiane, Paulo e Carlos, onde passamos por momentos juntos nas disciplinas.

Em especial a Kelly e Carlos por me ajudarem na otimização de alguns códigos no programa R.

A secretária Nádia do programa de pós-graduação em Estatística e Experimentação Agropecuária, ao secretário Fernando do Departamento de Estatística, as secretárias Josiane, Maria e Magali do Departamento de Ciências Exatas e as meninas da limpeza do prédio.

À Universidade Federal de Lavras (UFLA) pela oportunidade de realização do doutorado e aos professores do Departamento de Estatística (DES).

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo financiamento da bolsa de estudos.

"É possível mudar uma realidade,
basta correr atrás
e nunca desistir"

RESUMO

As ondaletas surgiram para solucionar os problemas quando se trabalha com dados não estacionários, sinais contaminados com ruídos, grande volume de dados, detecção de auto-semelhança, separação de componentes num sinal, entre outros. A técnica chamada "transformada de ondaleta" corresponde a uma das suas principais características, pois o dado (sinal, imagem ou função) pode ser decomposto tanto no domínio da frequência, quanto no domínio do tempo. As frequências baixas (escalas maiores) correspondem a uma informação global, que geralmente se estende por todo o dado analisado, enquanto que as frequências altas (escalas reduzidas) correspondem a uma informação mais detalhada, que dura um período de tempo relativamente curto. O presente trabalho foi dividido na apresentação de três técnicas distintas de análise de agrupamento de genomas utilizando ondaletas. Essas técnicas foram empregadas em dez sequências do genoma da *Mycobacterium tuberculosis*. A primeira técnica utilizada para o agrupamento dos genomas foi o uso da energia (variância). Essa energia foi obtida por meio da soma dos coeficientes de detalhes ao quadrado de cada nível de decomposição (cinco níveis) do sinal original por meio da ondaleta Daubechies com quatro momentos nulos. Como resultado, verificou-se a formação de 3 grupos distintos. A segunda técnica abordou a junção de ondaletas com a metodologia *Elastic net*. Nessa análise, depois de obtidos os níveis de decomposição utilizando ondaletas, o *Elastic net* foi aplicado em cada nível, onde pode-se verificar a formação dos grupos. Os resultados obtidos mostraram que os níveis 4 e 5 foram os que apresentaram a melhor formação dos grupos, sendo encontrados três grupos distintos. A terceira técnica abrangeu a combinação de ondaletas com o expoente de Hurst. A partir dos resultados obtidos dos níveis de decomposição por ondaletas, utilizando as mesmas configurações da primeira e segunda técnicas descritas anteriormente, foi feito o cálculo do expoente de Hurst para cada nível de decomposição, utilizando cinco métodos de estimação do expoente de Hurst. Cada método apresentou formações de grupos diferentes, mas o método que apresentou os resultados similares de acordo com as duas técnicas anteriores, foi o método de variância agregada.

Palavras-chave: Transformada não-decimada de ondaletas. Genoma. *Mycobacterium tuberculosis*.

ABSTRACT

The wavelets arised to solve the problems when you work with non-stationary data, signals contaminated with noise, large data volume, detection of self-similarity, separation of components in a signal, among others. The technique called the "wavelet transform" corresponds to one of its main characteristics, because the data (signal, image or function) can be decomposed in the frequency domain as well as in the time domain. The low frequencies (larger scales) correspond to a global information, which generally extends over the analyzed data, while the high frequencies (reduced scales) correspond to more detailed information, which lasts a relatively short period of time. The present work was divided in the presentation of three different genome cluster analysis techniques using wavelets. These techniques were employed in ten sequences of the *Mycobacterium tuberculosis* genome. The first technique used to grouping the of genomes was the use of energy (variance). This energy was obtained by summing the detail coefficients by the square of each level of decomposition (five levels) of the original signal by means of the Daubechies wavelet with four null moments. As a result, the formation of 3 distinct groups was found. The second technique approached the junction of wavelets with the methodology Elastic net. In this analysis, after obtaining the levels of decomposition using wavelets, the Elastic net was applied at each level, where it was possible to verify the formation of the groups. The results showed that levels 4 and 5 were the ones that presented the best formation of the groups, being found three different groups. The third technique involved the combination of wavelets with the Hurst exponent. From the results obtained of the levels of decomposition by wavelets, using the same configurations of the first and second techniques previously described, the Hurst exponent was calculated for each level of decomposition, using five methods of estimation of the Hurst exponent. Each method presented different group formations, but the method that presented the similar results according to the two previous techniques was the method of aggregate variance.

Keywords: Non-decimated wavelet transform. Genome. *Mycobacterium tuberculosis*.

LISTA DE FIGURAS

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO

Nas últimas décadas vem crescendo cada vez mais as análises empregando-se a técnica de ondaletas. Uma das grandes vantagens associadas a esse método corresponde ao ganho computacional, as análises são processadas quase que em tempo real. Diversas áreas da ciência vêm divulgando a sua aplicabilidade, entre elas estão a Física, Matemática, Engenharias, Genética, entre outras (OLIVEIRA, 2009).

A transformada de ondaleta é uma nova técnica de observar e representar um sinal. Matematicamente, ela é representada por uma função oscilante no tempo ou no espaço. Como característica possui janelas móveis que se dilatam ou se comprimem para capturar sinais de baixa e alta frequência respectivamente (BARBOSA; BLITZKOW, 2008). Sua origem ocorreu no campo do estudo sismológico, para descrever os distúrbios decorrentes de um impulso sísmico (OLIVEIRA, 2009).

As ondaletas possuem algoritmos que processam os dados em diferentes escalas ou resoluções e, independentemente de ser uma imagem, uma curva ou uma superfície. As ondaletas oferecem uma técnica refinada na representação dos níveis de detalhes presentes. Elas constituem uma ferramenta matemática para decompor funções hierarquicamente, permitindo que uma função seja descrita em termos de uma forma grosseira, sem muitas informações, como também numa forma mais detalhada, refinamento das informações (LIMA, 2002).

Dentre as técnicas de ondaletas, tem-se a transformada discreta não-decimada, cuja principal característica está no fato de se poder trabalhar com qualquer tamanho de sinais ou sequências. Nessa técnica os coeficientes de ondaletas são invariantes de translação, a escolha da origem é irrelevante, pois todas as observações são utilizadas na análise, situação que não ocorre na transformada discreta decimada de ondaleta.

Das características que a análise de ondaleta apresenta, temos a sua aplicabilidade na detecção de auto-similaridade, em que para determinados conjuntos de dados do mundo real, como por exemplo genomas de uma determinada espécie de bactéria, são estatisticamente auto-similares, onde parte desses genomas apresentam as mesmas propriedades estatísticas em várias escalas.

Uma ferramenta que vem ganhando destaque para detecção de dados similares é o *Elastic Net*. Essa ferramenta corresponde a uma regressão penalizada que combina linearmente as penalidades l_1 e l_2 dos métodos *Lasso* e *Ridge*. Como vantagens, o *Elastic Net* remove a limitação do número de variáveis selecionadas, incentiva o efeito de agrupamento (tendência de

que covariáveis altamente correlacionadas serem simultaneamente selecionadas) e estabiliza o caminho de regularização.

Outra metodologia também muito utilizada para análise de padrões de similaridades é o expoente de Hust, representado pela letra H. O valor desse expoente varia entre 0 e 1. Para $H = 0,5$ o sinal ou processo é aleatório. Para $0 < H < 0,5$, o sinal é caracterizado como anti-persistente, logo existe uma probabilidade maior do que cinquenta por cento de que um valor "negativo" seja seguido de um valor "positivo". E para $0,5 < H < 1$, o sinal é dito persistente, pois apresenta uma tendência, logo a probabilidade de repetição de um valor é maior do que cinquenta por cento.

O objetivo deste trabalho corresponde ao emprego de três técnicas distintas para a análise de agrupamento de genomas. Primeira técnica: uso da energia em cada nível de decomposição. Segunda técnica: interação de ondaletas e *Elastic net*. Terceira técnica: junção de ondaletas e expoente de Hurst.

A estrutura do texto foi dividida em duas partes. A primeira parte corresponde ao referencial teórico composto por 6 seções e na segunda parte são apresentados três artigos científicos, onde os dois primeiros artigos já foram publicados e o terceiro se encontra sob revisão.

Na primeira parte temos as seguintes seções: seção 2.1 são abordadas as técnicas referentes à análise com ondaletas, na seção 2.2 é apresentado a análise de agrupamento, na seção 2.3 é explorado a regressão penalizada, na seção 2.4 tem-se a descrição sobre memória longa e o expoente de Hurst, na seção 2.5 tem-se ondaletas aplicadas na genética e na seção 2.6 tem-se uma breve descrição referente a dados genômicos, tendo um direcionamento para o estudo do genoma de bactérias.

A segunda parte é constituída por três artigos científicos: o primeiro artigo com o título "Evaluation of genome similarities using the non-decimated wavelet transform" foi publicado na revista Genetics and Molecular Research; o segundo artigo "Wavelet-domain Elastic net for clustering on genomes strains" foi publicado na revista Genetics and Molecular Biology; e o terceiro "Evaluation of genome similarities: a wavelet-domain approach" se encontra sob revisão na revista Genetics and Molecular Biology.

Para finalizar são apresentadas as considerações gerais sobre o presente trabalho.

2 REFERENCIAL TEÓRICO

Nestas seções serão abordados temas sobre à análise com ondaletas, análise de agrupamento, regressão penalizada, memória longa e o expoente de Hurst, ondaletas aplicadas na genética e dados genômicos com direcionamento para o estudo do genoma de bactérias.

2.1 Ondaletas

Uma função ondaleta é a interpretação de uma onda de curta duração com crescimento e decrescimento rápidos. Sua teoria baseia-se na representação de funções em diferentes escalas e diferentes resoluções (tempo-escala), sendo considerada uma das suas principais características (DAUBECHIES, 1992).

Uma ondaleta é simplesmente uma função da onda cuidadosamente construída de modo a ter determinadas propriedades matemáticas. Todo o conjunto das ondaletas é construído a partir de uma única função “ondaleta mãe” e este conjunto fornece funções úteis de “blocos de construção” que podem ser usadas para descrever qualquer classe grande de funções. Várias possibilidades diferentes para funções ondaleta mãe tem sido desenvolvidas, cada uma com as suas vantagens e desvantagens associadas (OGDEN, 1997).

Ondaletas são uma forma relativamente nova de analisar por exemplo séries temporais, em que os dados formais remontam a 1980, mas em muitos aspectos as ondaletas são uma síntese de ideias mais antigas com novos resultados matemáticos elegantes e algoritmos computacionais eficientes. A análise de ondaletas é em alguns casos complementar às técnicas de análises existentes (correlação e análise espectral) e, em outros casos, capaz de solucionar problemas para os quais se tinha pouco progresso antes da introdução das ondaletas (PERCIVAL; WALDEN, 2000).

Ondaletas estão intrinsecamente ligadas à notação de “análise multirresolução", ou seja, objetos (sinais, funções, imagens) podem ser examinadas usando diferentes níveis de resolução (OGDEN, 1997; KAISER, 1994).

Na análise de sinais, as representações de ondaletas nos permitem ver uma evolução no domínio do tempo em termos de componentes de escala. Devido a isso, as transformações de ondaletas comportam-se similarmente às transformações de Fourier. Entretanto, na transformada de Fourier quando se extrai detalhes da frequência do sinal, toda a informação sobre a localização de uma determinada frequência dentro do sinal se perde. Na ondaleta, o tempo de localização pode ser alcançado por meio da primeira janela, pois as fatias do sinal processado

são de um comprimento fixo, o qual é determinado pela janela. Logo, as fatias do mesmo tamanho são usadas para resolver ambos os componentes de alta e baixa frequência. Para sinais não estacionários, essa falta de adaptabilidade pode levar a um local sub ou sobre apropriado. Portanto, em contraste com as janelas da transformada de Fourier, a largura das fatias da ondaleta selecionada são de acordo com o tempo e a frequência local do sinal. Essa propriedade de adaptabilidade das ondaletas é bastante importante (VIDAKOVIC, 1999).

Na análise de uma ondaleta, a janela é oscilante e é chamada de ondaleta mãe. Ocorrem-se translações arbitrárias e dilatações. Dessa maneira a ondaleta mãe gera outras ondaletas (HERNANDEZ; WEISS, 1996).

Por definição, uma ondaleta é uma função $\psi(t) \in L^2(\mathbb{R})$, tal que sua família de funções é dada por:

$$\psi_{j,k} = 2^{-j/2}\psi(2^{-j}t - k), \tag{2.1}$$

em que j e k são inteiros arbitrários em uma base ortonormal no espaço de Hilbert $L^2(\mathbb{R})$ (WOJTASZCZYK, 1997).

Existem duas principais transformadas de ondaletas. A primeira é conhecida como a transformada contínua de ondaletas (*Continuos Wavelet Transform* - CWT), que é projetada para trabalhar, por exemplo, com séries temporais definidas ao longo de todo o eixo real e a segunda é a transformada discreta de ondaletas (*Discrete Wavelet Transform* - DWT), que lida com os dados definidos essencialmente ao longo de um intervalo de números inteiros (geralmente $t = 0, 1, ..., N-1$, onde N representa o número de valores que podem ser de uma série temporal, sinal, imagem, entre outros) (PERCIVAL; WALDEN, 2000).

2.1.1 Transformada discreta decimada de ondaletas

Uma transformada de ondaleta decompõe um sinal em vários grupos (vetores) de coeficientes. Esses vetores de diferentes coeficientes contêm informações sobre as características do sinal em diferentes escalas. Os coeficientes das escalas mais grossas captam características globais do sinal, enquanto os coeficientes das escalas mais finas contêm os detalhes locais (LIÒ, 2003).

A versão decimada (Transformada Discreta de Ondaleta) é a mais conhecida. Ela é calculada por meio de um algoritmo eficiente, denominado algoritmo piramidal, que usa filtros discretos e decimação por 2 (MALLAT, 1989). Nesse processo, trabalha-se com uma série de

tamanho $N = 2^j$, onde no nível j tem-se metade dos coeficientes do nível anterior $j-1$. A transformada de ondaleta depende da escolha da decimação, que corresponde à escolha de uma origem (PERCIVAL; WALDEN, 2000).

Para analisar estruturas de sinais de tamanhos muito diferentes, é necessário o uso de energia (tempo-frequência) com diferentes suportes de tempo. A transformada de ondaleta decompõe sinais por ondas dilatadas e transladadas. Uma ondaleta é representada por uma função $\psi \in L^2(\mathbb{R})$, com média zero (MALLAT, 1999), tal que

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0, \tag{2.2}$$

em que $\|\psi\| = 1$ e centrado na vizinhança de $t = 0$.

De acordo com Mallat (1989) a Transformada Discreta de Ondaleta (DWT) pode ser representada na forma matricial, por meio de uma matriz ortogonal:

$$W = \left[W_1^T, W_2^T, ..., W_J^T, V_J\right]^T, \tag{2.3}$$

em que J é o maior nível da transformada e ^{T} corresponde à transposta. Uma DWT é aplicada para um vetor X de observações com $d = WX$ e decompõe os dados dentro de um conjunto de coeficientes de ondaletas:

$$d = \left[d_1^T, d_2^T, ..., d_J^T, c_J\right]^T, \tag{2.4}$$

com $d_j = W_j X$, $c_J = V_J X$. Na escala $\tau_J = 2^{j-1}$, ou nível j, existem $n/2^j$ coeficientes de d_j, que são associados com mudanças nas médias dos dados numa escala $\tau_j \Delta t$, onde Δt é o intervalo de tempo entre observações consecutivas. Cada coeficiente de ondaleta nesse nível nos indica quanto a média ponderada dos dados muda de um período de tempo particular de comprimento efetivo $\tau_j \Delta t$ para o próximo. Os coeficientes de escala c_J são em vez disso associados com as médias dos dados na escala $\tau_{J+1} \Delta t$ superior, com J sendo o maior nível da DWT. A transformada de ondaleta é uma medida cumulativa das variações dos dados em regiões proporcionais às escalas de ondaletas; para aumentar os valores de j, os coeficientes descrevem características dos intervalos de frequência mais baixos e mais altos.

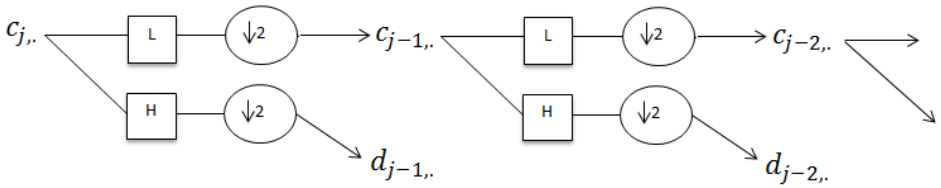
2.1.1.1 Algoritmo piramidal

O algoritmo piramidal desenvolvido por Mallat (1989), utiliza os filtros passa-baixo l_k, chamado de função escala ou ondaleta pai e passa-alto h_k, chamado de ondaleta mãe, com coeficientes dados por:

$$\begin{aligned} l_k &= \sqrt{2}\int_{-\infty}^{+\infty} \phi(t)\phi(2t-k)dt, \\ h_k &= \sqrt{2}\int_{-\infty}^{+\infty} \psi(t)\phi(2t-k)dt. \end{aligned} \tag{2.5}$$

Na Figura 2.1 mostra-se o procedimento do algoritmo.

Figura 2.1 – Algoritmo piramidal.



Fonte: Morettin (1999)

Na Figura 2.1, L indica o filtro passa-baixo, H o filtro passa-alto e $\downarrow 2$ indica a operação de decimação por 2, ou seja, a cada duas saídas do filtro, desprezamos uma (MORETTIN, 1999).

No j-ésimo passo, o algoritmo calcula $c_{j,k}$ e $d_{j,k}$ a partir dos coeficientes de ondaleta do nível $j+1$, $c_{j+1,k}$, por

$$\begin{aligned} c_{j,k} &= \sum_n l_{n-2k} c_{j+1,n}, \\ d_{j,k} &= \sum_n h_{n-2k} c_{j+1,n}, \end{aligned} \tag{2.6}$$

em que $c_{j,k}$ e $d_{j,k}$, são obtidos de $c_{j+1,n}$ por médias móveis amostradas nos inteiros pares, que é a decimação vista anteriormente. No nível j, teremos metade do número de coeficientes do nível $j+1$, daí vem o nome "piramidal", dado por Mallat (1989), ou "cascata", dado por Daubechies (1992).

Definindo os filtros L e H, teremos $L = (l_k)_{k\in\mathbb{Z}}$ e $H = (h_k)_{k\in\mathbb{Z}}$. O filtro L corresponde a tomar médias e H as diferenças. Se f for um sinal e $f' = Lf$ e $f^* = Hf$, teremos, o caso da ondaleta de Haar, ou seja,

$$\begin{aligned} f'_k &= \tfrac{1}{\sqrt{2}}(f_{2k} + f_{2k-1}), \\ f^*_k &= \tfrac{1}{\sqrt{2}}(f_{2k} - f_{2k-1}). \end{aligned} \tag{2.7}$$

Assumindo $c_{j,.}$ os coeficientes de aproximação no nível j e por $d_{j,.}$ os coeficientes de detalhes no nível j, teremos

$$\begin{aligned} c_{j,.} &= Lc_{j+1,.}, \\ d_{j,.} &= Hc_{j+1,.}, \end{aligned} \tag{2.8}$$

de modo que

$$d_{j-m,.} = HL^{m-1}c_{j,.}. \tag{2.9}$$

Os filtros são normalizados de tal forma que $\sum_k l_k = 2$ e $\sum_k h_k = 0$.

2.1.2 Transformada discreta não-decimada de ondaletas

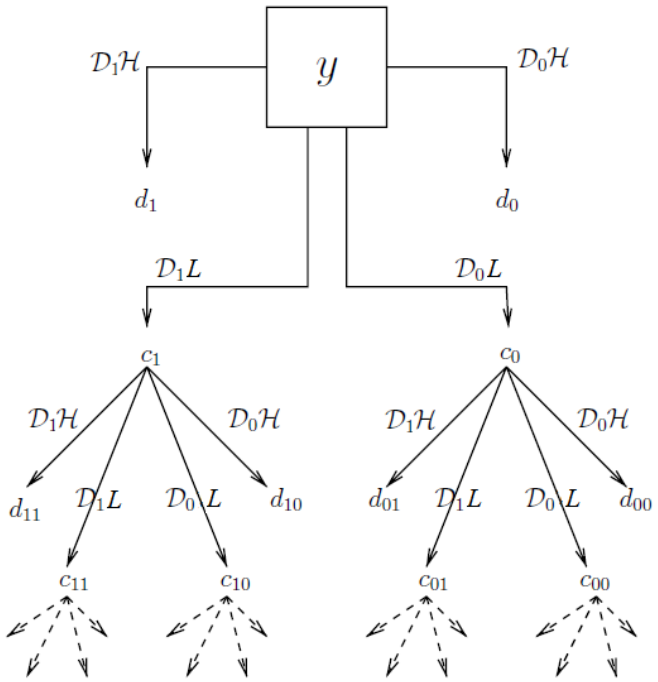
A característica da transformada discreta não-decimada de ondaletas (*Non-Decimated Wavelet Transform*: NDWT) é manter a mesma quantidade de dados nas decimações pares e ímpares em cada escala e continuar a fazer o mesmo em cada escala subsequente. Seja D_0 a decimação par, D_1 decimação ímpar, H o filtro passa-alto e L o filtro passa-baixo. Considere como exemplo, um vetor de entrada $(y_1, \cdots, y_n)$. Em seguida aplica-se e mantêm-se ambos D_0H_y e D_1H_y, pares e ímpares indexados das observações de ondaletas filtradas. Cada uma dessas sequências são de comprimento $n/2$. Logo, no total, o número de coeficientes de ondaletas em ambas decimações na escala mais fina é $2 \times n/2 = n$ (NASON, 2008).

Na Figura 2.2 tem-se que na escala mais fina os coeficientes da ondaleta são d_0 e d_1. Na próxima escala mais fina são $d_{00}, d_{01}, d_{10}, d_{11}$.

Outra maneira similar de se obter a escala mais fina dos coeficientes da ondaleta pai é calcular D_0L_y ($n/2$ números) e D_1L_y ($n/2$ números). Assim, para o próximo nível dos coeficientes de ondaleta aplica-se ambos D_0H e D_1H para ambos D_0L_y e D_1L_y. O resultado de cada um desses é $n/4$ coeficientes de ondaleta na escala $J-2$. Portanto, existem 4 conjuntos, onde o total do número de coeficientes é n (NASON, 2008).

Na transformada discreta não-decimada de ondaleta os coeficientes são invariantes de translação, ou seja, significa que o deslocamento circular dos dados é refletido na mesma direção dos coeficientes. Uma outra característica, é a capacidade de lidar com dados de tamanho arbitrário que não requer que o tamanho da amostra n seja uma potência de dois. A principal vantagem desse método está associada com os filtros de passa zero, o que significa que opera

Figura 2.2 – Diagrama da transformada discreta não-decimada de ondaleta.



Fonte: Nason (2008)

circularmente aos dados permitindo funcionalidades em diferentes escalas para serem alinhados com a sequência dos dados originais (VANNUCCI; LIÒ, 2001).

2.1.3 Escalograma

O escalograma é uma ferramenta bastante útil para a interpretação do sinal da ondaleta representada. É definido como um gráfico da soma de quadrados dos coeficientes da ondaleta nos diferentes níveis. No contexto da transformação discreta, representa uma decomposição da energia de uma função no tempo-frequência (escala). Como uma das suas características está a capacidade de detecção de componentes periódicos do sinal, ou seja, diferentes componentes resultarão em visíveis picos no escalograma. Esses componentes podem ser extraídos do sinal por divisão dos coeficientes da ondaleta dentro de diferentes conjuntos, onde cada um desses conjuntos pertencem ao mesmo pico. Componentes de frequência alta e baixa de um sinal podem ser reconstruídos aplicando a transformação de ondaleta inversa para conjuntos separados (LIÒ; VANNUCCI, 2000).

A energia $E(j)$ para os coeficientes da ondaleta d em cada nível j, que corresponde ao escalograma é representado por

$$E(j) = \sum_{k=0}^{n} d_{j,k}^2 \qquad \text{para } \; j = 1, \cdots, J. \tag{2.10}$$

A equação 2.10 corresponde ao cálculo da energia na transformada discreta não-decimada de ondaleta (GENÇAY; SELÇUK; WHITCHER, 2002).

2.1.4 Ondaletas Daubechies

As ondaletas de Daubechies são uma família de ondaletas ortogonais que definem uma transformação de ondaletas discretas e são caracterizadas por um número máximo de momentos nulos para algum suporte dado. Com cada tipo de ondaleta dessa classe, existe uma função de escalonamento (chamada de ondaleta pai), que gera uma análise de multirresolução ortogonal.

Na Figura 2.3 tem-se a esquerda as ondaletas Daubechies pai e a direita as ondaletas Daubechies mãe, sendo que as letras (a) e (b) correspondem a 2 momentos nulos, as letras (c) e (d) a 4 momentos nulos e nas letras (e) e (f) a 8 momentos nulos.

Segundo Daubechies (1992), para cada inteiro r, a base ortonormal para $L^2(\mathbb{R})$ está definida como

$$\phi_{r,j,k}(x) = 2^{-j/2}\phi_r(2^{-j}x - k), \quad j,k \in \mathbb{Z}, \tag{2.11}$$

em que a função $\phi_r(x)$ em $L^2(\mathbb{R})$ tem a propriedade que $\phi_r(x-k) \mid k \in \mathbb{Z}$ é uma base sequencial ortonormal em $L^2(\mathbb{R})$. Aqui j é o índice de escala, k é o índice de translação e r é o índice de filtragem.

A tendência f^j na escala 2^{-j} de uma função $f \in L^2(\mathbb{R})$ está definida como

$$f_j(x) = \sum_k \left\langle f, \phi_{r,j,k} \right\rangle \phi_{r,j,k}(x). \tag{2.12}$$

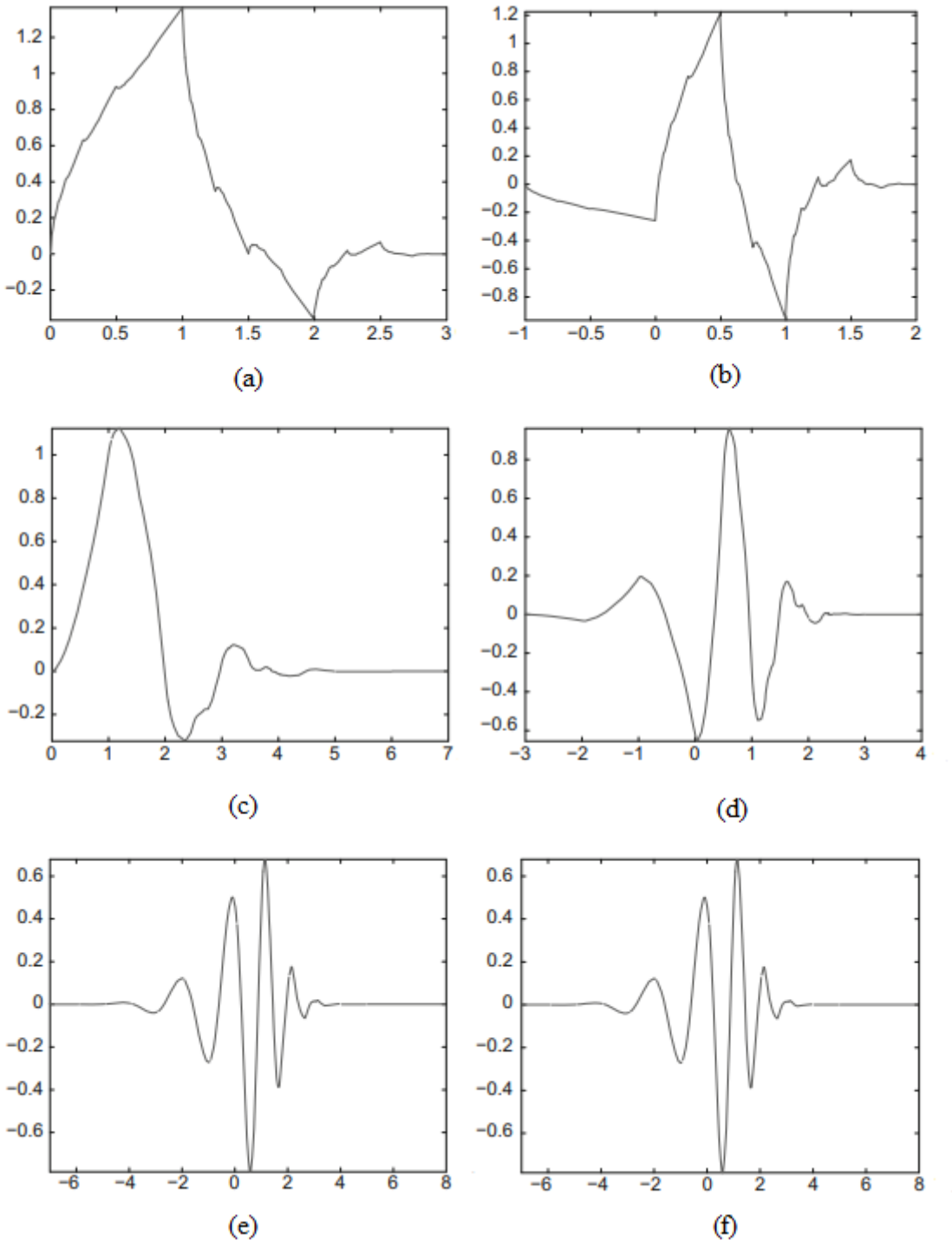
Os detalhes e flutuações são definidos por

$$d_j(x) = f_{j+1}(x) - f_j(x). \tag{2.13}$$

Para analisar esses detalhes em uma dada escala, define-se uma base ortonormal $\psi_r(x)$ com propriedade similares às de $\phi_r(x)$ descritas anteriormente.

As funções $\phi_r(x)$ e $\psi_r(x)$, denominadas ondaleta pai (função escala) e ondaleta mãe (função ondaleta), respectivamente, são as funções protótipas necessárias para a análise de on-

Figura 2.3 – Ondaletas da família Daubechies. A direita tem-se a função ondaleta (ondaleta mãe) e a esquerda tem-se a função escala (ondaleta pai).



Fonte: Morettin, Pinheiro e Vidakovic (2017)

daletas. As famílias de ondaletas, como aquelas descritas na equação 2.11, são geradas a partir da ondaleta pai ou da mãe mudando a escala e translação no tempo (ou espaço em processamento de imagens).

A base ortonormal de Daubechies tem as seguintes propriedades:

1) $\psi_r(x)$ tem o intervalo de suporte compacto $[0, 2r+1]$;

2) $\int_{-\infty}^{+\infty} \psi_r(x)dx = ... = \int_{-\infty}^{+\infty} x^r \psi_r(x)dx = 0$.

Uma característica da ondaleta de Daubechies está no fornecimento de excelentes resultados no processamento de imagens.

2.1.5 Momentos nulos

As ondaletas possuem um número de momentos nulos, ou seja, uma função $\psi \in L^2(\mathbb{R})$ tem m momento nulos se satisfaz:

$$\int x^l \psi(x)dx = 0 \tag{2.14}$$

para $l = 0, ..., m-1$.

Se as ondaletas tem m momentos nulos, então todos os coeficientes da ondaleta de qualquer polinômio de grau m ou menor serão exatamente zero (NASON, 2008).

Momentos nulos e suavidade estão matematicamente relacionados, pois quanto maior o número de momentos nulos de uma ondaleta, mais suave ela será. E quanto maior a suavidade da ondaleta, maior é a probabilidade de reconstrução perfeita do sinal decomposto pela transformada de ondaleta (UZINSKI, 2013).

2.2 Análise de agrupamento

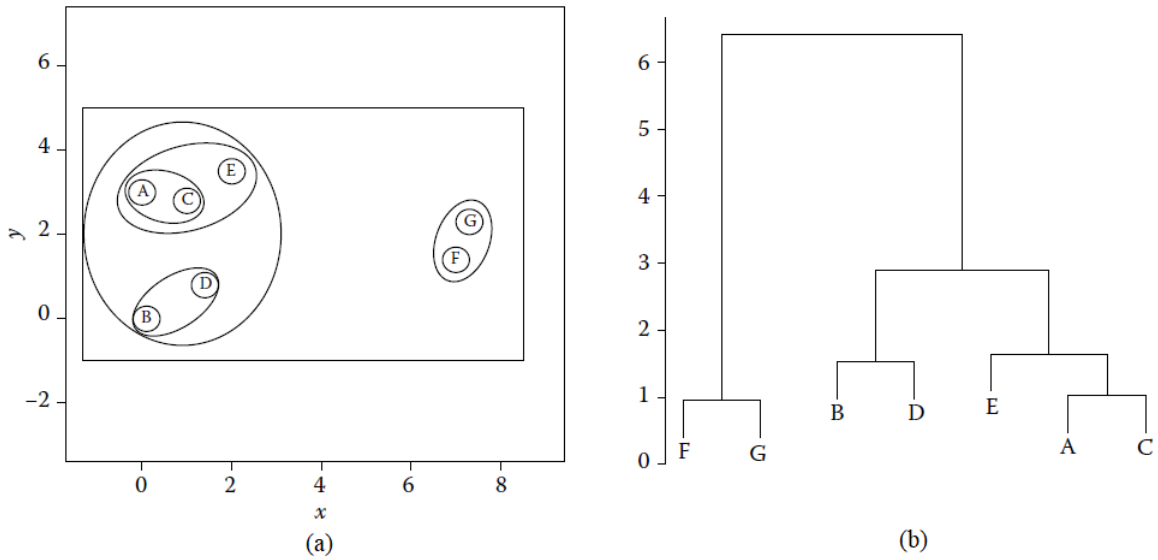
Conceitualmente a análise de agrupamento é a classificação de objetos em diferentes grupos, cada um dos quais deve conter os objetos semelhantes segundo alguma função de distância estatística. De maneira geral, os elementos dentro de um grupo tem um alto grau de "associação natural" entre si, enquanto que os grupos são "relativamente distintos" um do outro. A abordagem do problema e os resultados alcançados dependem principalmente de como o investigador escolhe dar significado operacional às expressões "associação natural" e "relativamente distinto" (ANDERBERG, 1973).

As técnicas de agrupamentos vem crescendo na análise de conjuntos de dados biológicos. Embora muitos métodos genéricos de agrupamento tenham sido usados com sucesso para analisar conjuntos de dados biológicos, muitas propriedades específicas desses conjuntos de dados exigem métodos personalizados que são especificamente projetados para atender a essas propriedades. Portanto, ambos os aspectos sobre o projeto e a aplicação de métodos

de agrupamento neste campo têm estado sob investigações e considerações ativas por muitos grupos de pesquisa em todo o mundo. De fato, tem-se feito muitas atividades para assimilar o trabalho neste campo, especialmente projetando métodos de última geração que abordem de maneira única, várias questões que são particularmente relevantes em dados biológicos (ABU-JAMOUS; FA; NANDI, 2015).

Segundo Abu-Jamous, Fa e Nandi (2015) o agrupamento hierárquico (Figura 2.4) é um dos métodos de agrupamento mais populares na literatura. Em contraste com o particionamento parcial, que tenta decompor diretamente o conjunto de dados em um conjunto de grupos desarticulados, um método de agrupamento hierárquico é um procedimento para transformar uma matriz de proximidade em uma partição aninhada, que pode ser representada graficamente por uma árvore, chamada dendrograma. Para obter o número de grupos e a partição correspondente, temos que cortar o dendrograma em um determinado nível. Cortá-lo em diferentes níveis levará a diferentes resultados de agrupamento com diferentes níveis de resolução.

Figura 2.4 – (a) Hierarquia de conjuntos, (b) Dendrograma.



Fonte: Hennig et al. (2016)

Os algoritmos hierárquicos de agrupamento são principalmente classificados em métodos aglomerativos (métodos de baixo para cima) e divisivos (métodos de cima para baixo), baseados em como o dendrograma hierárquico é formado. Os métodos de aglomeração começam com N grupos inicialmente (basicamente consideram cada objeto como um grupo). Gradualmente, eles mesclam o par de grupos mais próximo em termos de diferentes métodos de vinculação, até que todos os grupos sejam mesclados em um grupo e um dendrograma seja for-

mado. Existem muitos métodos de ligação, nomeadamente ligação simples, ligação completa, ligação média, ligação de Ward e entre outros (ABU-JAMOUS; FA; NANDI, 2015).

O agrupamento de ligação média foi desenvolvido como uma alternativa aos extremos de ambos ligação simples e completa. Embora existam algumas variantes do método, cada uma essencialmente calcula uma média da similaridade de um caso em consideração, com todos os casos no grupo existente e, subsequentemente, associa o caso a esse grupo se um determinado nível de similaridade for obtido usando esse valor médio. A variante mais comumente utilizada de ligação média calcula a média aritmética de semelhanças entre os casos. Outras variantes da ligação média são projetadas para calcular a similaridade entre os centróides de dois grupos que podem ser mesclados. A ligação média tem sido usada extensivamente nas ciências biológicas, mas só recentemente começou a se ver muito uso nas ciências sociais (ALDENDERFER; BLASHFIELD, 1984).

Dentre as funções de distâncias utilizadas para a análise de agrupamento, abordaremos sobre a distância de Mahalanobis, que é baseada nas correlações entre variáveis com as quais distintos padrões podem ser identificados e analisados.

A distância de Mahalanobis aumenta com o aumento das distâncias entre dois grupos e diminuiu com a variação dentro do grupo. Ao empregar correlações dentro do grupo, a distância de Mahalanobis leva em conta a forma (possivelmente não esférica) dos grupos (EVERITT et al., 2011).

O uso da distância de Mahalanobis (D^2) de acordo com a Equação 2.15 implica que o pesquisador está disposto a assumir que as matrizes de covariância são aproximadamente as mesmas nos dois grupos. Quando isso não acontece, D^2 é uma medida de intergrupo inadequada e, para tais casos, várias alternativas foram propostas. A distância de Mahalanobis é dada por:

$$D^2 = (\bar{x}_A - \bar{x}_B)' W^{-1} (x_A - x_B), \tag{2.15}$$

em que $\bar{x}_A{}'$ significa o vetor do grupo A, $\bar{x}_B{}'$ significa o vetor do grupo B e W é a matriz de covariância dentro do grupo.

Segundo McLachlan (1999) a distância de Mahalanobis leva em conta a variabilidade. Em vez de tratar todos os valores da mesma maneira quando calcula a distância ao ponto central, pondera-os pela diferença à amplitude de variação na direção do ponto de teste. A fronteira de Mahalanobis torna-se assim clara. Esta função constrói um espaço ao longo do eixo de alongamento elíptico que for detectado.

Exemplificando em termos das medidas de Mahalanobis, uma amostra"A" terá um valor substancialmente menor de distância à média que uma amostra "B", se distribuir ao longo do eixo do grupo com maior variabilidade. Assim, a amostra "A" é mais provavelmente classificada como relacionada com o grupo. A distância de Mahalanobis permite observar não apenas as variações (variância) mas também a covariância. O grupo com as distâncias de Mahalanobis como medida, define um espaço multidimensional, cujas fronteiras determinam o intervalo de variação tido por aceitável, para que amostras desconhecidas possam ser classificadas como relacionadas com uma distribuição (MANLY; ALBERTO, 2016).

Outra vantagem de usar a medida de Mahalanobis para discriminar, é que as distâncias são calculadas em unidades de desvio-padrão a partir da média do grupo, o que faz com que a elipse englobante calculada, formada à volta do agrupamento, defina a zona de um desvio-padrão. Isto permite ao analista atribuir uma probabilidade estatística a essa medida. Em teoria, amostras com uma distância de Mahalanobis de três ou mais, têm probabilidade de 0,01 ou menos e podem ser classificadas como não-membros do grupo em causa (MANLY; ALBERTO, 2016).

2.3 Regressão penalizada

De acordo com Matthew e Yahaya (2015), considere um modelo padrão de regressão linear múltipla dado por:

$$y = X\beta + \varepsilon, \tag{2.16}$$

em que $y = (y_1, y_2, ..., y_n)^T$ um vetor de resposta e $X = [X_1 |, ..., | X_p]$ o modelo matricial, onde $X_i = (x_{1i}, ..., x_{ni})^T$, e $i = 1, ..., p$ são as variáveis preditoras, $\beta = (\beta_0, ..., \beta_p)$ é um vetor coluna que contém os coeficientes de regressão e ε é um vetor de termos de erro que são considerados normalmente distribuídos com média 0 e variância σ_ε^2, isto é $\varepsilon \sim N\left(0, I\sigma_\varepsilon^2\right)$. Para modelos onde $n > p$, os valores dos parâmetros desconhecidos de β podem ser estimados minimizando a soma de quadrados residual (SQR), dada por:

$$SQR = \left(y - X\hat{\beta}\right)^T \left(y - X\hat{\beta}\right). \tag{2.17}$$

Após uma transformação de localização e escala, podemos assumir que a resposta é centrada e que os preditores são padronizados. Em geral, mínimos quadrados penalizados (MQP) é

um problema de otimização destinado a minimizar a soma de quadrados devido ao erro (SQE) sujeito a alguma penalidade sobre os valores dos parâmetros desconhecidos. Em suma, pode-se escrever MQP como:

$$\text{Minimizar} \quad SQR = \left(y - X\hat{\beta}\right)^T \left(y - X\hat{\beta}\right) \qquad \text{sujeito a} \qquad \text{Pen}\,(\beta) \leq t, \tag{2.18}$$

em que Pen (β) é uma penalidade específica, uma função de $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p\right)^T$, enquanto t é um parâmetro de ajuste.

O problema de otimização restrito formulado anteriormente pode ser resolvido usando a formulação Lagrangeana equivalente, que pode ser obtida minimizando:

$$\left(y - X\hat{\beta}\right)^T \left(y - X\hat{\beta}\right) + \lambda \text{Pen}\,(\beta)\,, \tag{2.19}$$

em que λ é um parâmetro de ajuste que controla a força do encolhimento. Por exemplo, quando $\lambda = 0$, ou seja, quando nenhuma penalização é aplicada, acabamos em uma clássica regressão de mínimos quadrados ordinários. Entretanto, quando λ fica maior, mais peso é dado ao termo de penalização.

2.3.1 *Ridge regression*

De acordo com Hoerl e Kennard (1970), a *Ridge regression* é uma maneira de criar um modelo parcimonioso, quando o número de variáveis preditores em um conjunto excede o número de observações ou quando um conjunto de dados tem multicolinearidade (correlações entre variáveis preditoras)

O estimador *Ridge regression*, segundo Ogutu, Schulz-Streeck e Piepho (2012) é ideal quando existem muitos preditores, todos com coeficientes não nulos e com comportamento de uma distribuição normal. Em particular, ele funciona bem com muitos preditores, cada um com efeito pequeno e evita que os coeficientes de modelos de regressão linear com muitas variáveis correlacionadas sejam pouco determinados e exibam alta variação. O *Ridge regression* encolhe os coeficientes de preditores correlacionados igualmente para zero. Assim, por exemplo, dados k preditores idênticos, cada um teria coeficientes idênticos igual a $1/kth$, do tamanho que qualquer um preditor poderia ser ajustado sozinho. Portanto, *Rigde regression* não força os coeficientes a desaparecer, logo não pode selecionar um modelo com apenas o subconjunto de preditores mais relevante e preditivo (FRIEDMAN; HASTIE; TIBSHIRANI, 2010).

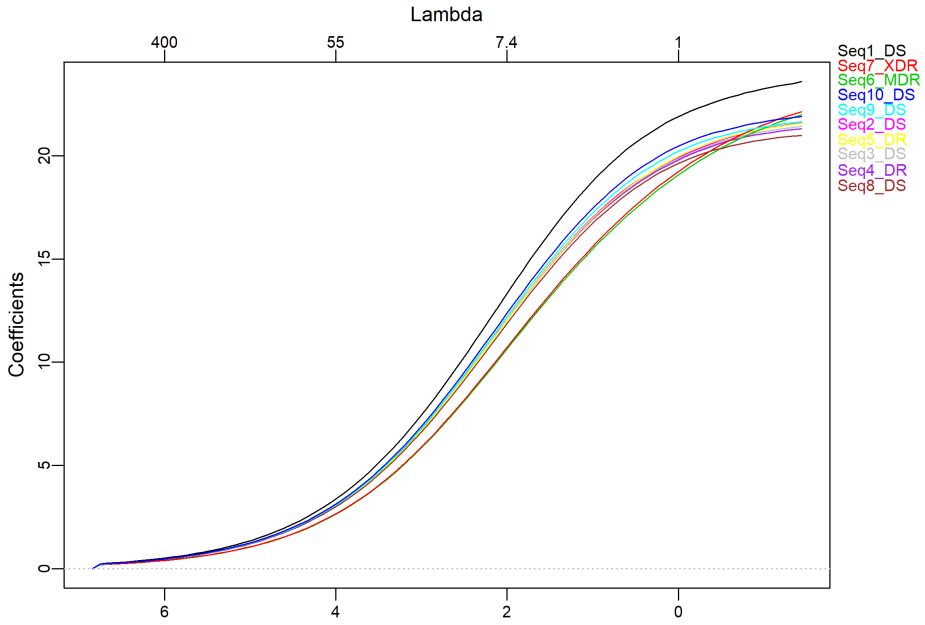
Na Figura 2.5 temos o exemplo referente as dez sequências do genoma da *mycobacterium tuberculosis* (conjunto de dados utilizado na tese) aplicando a metodologia *Ridge regression*. Pode-se verificar que todas as dez sequências saem juntas, olhando o gráfico no eixo x, da esquerda para direita. As sequências mais significativas correspondem as primeiras na legenda, que são as sequências Seq1_DS, Seq7_XDR e Seq6_MDR.

O estimador *Ridge regression* utiliza a penalização de quadrados mínimos l_2:

$$\hat{\beta}(ridge) = \arg\min_{\beta} |y - X\beta|_2^2 + \lambda |\beta|_2^2, \tag{2.20}$$

em que $|y - X\beta|_2^2 = \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2$ é a norma l_2 (quadrática) função perda (soma de quadrado residual), x_i^T é a i-ésima linha de X, $|\beta|_2^2 = \sum_{i=1}^{p} \beta_j^2$ é a norma l_2 penalizada sobre β e $\lambda \geq 0$ é o ajuste (penalidade, regularização ou complexidade) do parâmetro que regula a força da penalidade (contração linear), determinando a importância relativa do erro empírico, dependente de dados e o prazo de penalidade.

Figura 2.5 – *Ridge regression*.



2.3.2 *Lasso*

De acordo com Tibshirani (1996) o estimador *Lasso* estima os coeficientes de regressão por meio de um critério de mínimos quadrados penalizados por uma norma l_1. Isso equivale a minimizar a soma de quadrado residual, mais uma penalidade l_1 nos coeficientes de regressão.

Devido à natureza da penalidade l_1, o *Lasso* (Figura 2.6) realiza o encolhimento contínuo e a seleção de variáveis simultaneamente.

A Figura 2.6 exemplifica o método *Lasso*, também aplicado as dez sequências do genoma da *mycobacterium tuberculosis*. Analisando o gráfico, podemos ver que o método seleciona as variáveis mais significativas e exclui as não significativas. A interpretação desse gráfico corresponde ao fato de que as primeiras variáveis que entraram (olhando o eixo x, da esquerda para a direita) são as mais significativas, logo nesse caso são as sequências Seq5_DR, Seq10_DS e Seq6_MDR.

Definição do estimador *Lasso*:

Suponha os dados $\left(x^i, y_i\right)$, $i = 1,2,3,...,N$, em que $x^i = (x_{i1},...,x_{ip})^T$ são as variáveis preditoras e y_i são as variáveis respostas. Como na configuração usual de regressão, assumimos que as observações são independentes ou que os $y_i^{'}s$ são condicionalmente independentes, dados $x_{ij}^{'}s$. Assumimos que x_{ij} são padronizados, tal que $\sum_i x_{ij}/N = 0$, $\sum_i x_{ij}^2/N = 1$.

Seja $\hat{\beta} = \left(\hat{\beta}_1,...,\hat{\beta}_p\right)^T$, o estimador *Lasso* $\left(\hat{\alpha},\hat{\beta}\right)$ é definido por

$$\left(\hat{\alpha},\hat{\beta}\right) = \arg\min \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \quad \text{sujeito a} \quad \sum_j \left|\beta_j\right| \leq t, \tag{2.21}$$

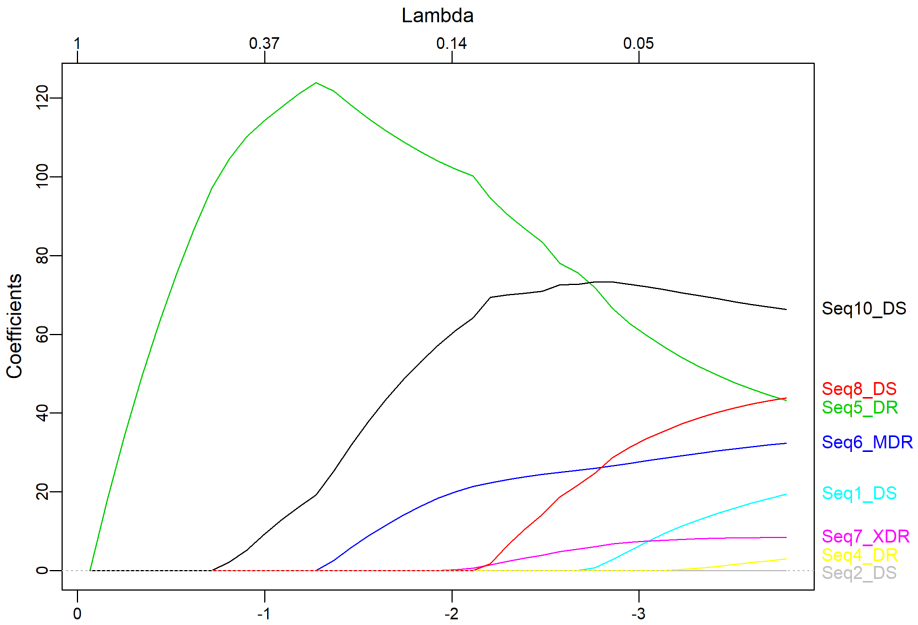
em que $t \geq 0$ é um parâmetro de ajuste. Para todo t, a solução para α é $\hat{\alpha} = \bar{y}$. Assuma-se sem perda de generalidade que $\bar{y} = 0$, logo omite-se α.

O parâmetro $t \geq 0$ controla a quantidade de encolhimento que é aplicada à estimativa. Seja $\hat{\beta}_j^o$ a estimativa dos mínimos quadrados completos e seja $t_0 = \sum \left|\beta_j^o\right|$. Valores de $t < t_0$ irá causar encolhimento das soluções para 0 e alguns coeficientes podem ser exatamente iguais a 0. Por exemplo, se $t = t_0/2$ o efeito será mais ou menos semelhante a encontrar o melhor subconjunto de tamanho $p/2$. Note também que a matriz de delineamento não precisa ser de posto completo.

2.3.3 *Naive Elastic net*

Segundo Zou e Hastie (2005), suponha que um conjunto de dados tem n observações com p preditores. Seja $y = (y_1,...,y_n)^T$ a resposta e $X = (x_1 \mid ... \mid x_p)$ o modelo matricial, em que $x_j = (x_{1j},...,x_{nj})^T$, $j = 1,...,p$, são os preditores. Após uma localização e uma transformação escalar, assume-se que a resposta é centralizada e os preditores são padronizados, ou seja,

Figura 2.6 – *Lasso*.



$$\sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} x_{ij} = 0 \ \ e \ \ \sum_{i=1}^{n} x_{ij}^2 = 1 \ \text{ para } j = 1,2,...,p. \tag{2.22}$$

Para quaisquer λ_1 e λ_2 fixos e não-negativos, define-se o critério *Naive Elastic net* por

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1, \tag{2.23}$$

em que

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2,$$

$$|\beta|_1 = \sum_{j=1}^{p} |\beta_j|.$$

O estimador *Naive Elastic net* $\hat{\beta}$ é o minimizador da equação (2.23)

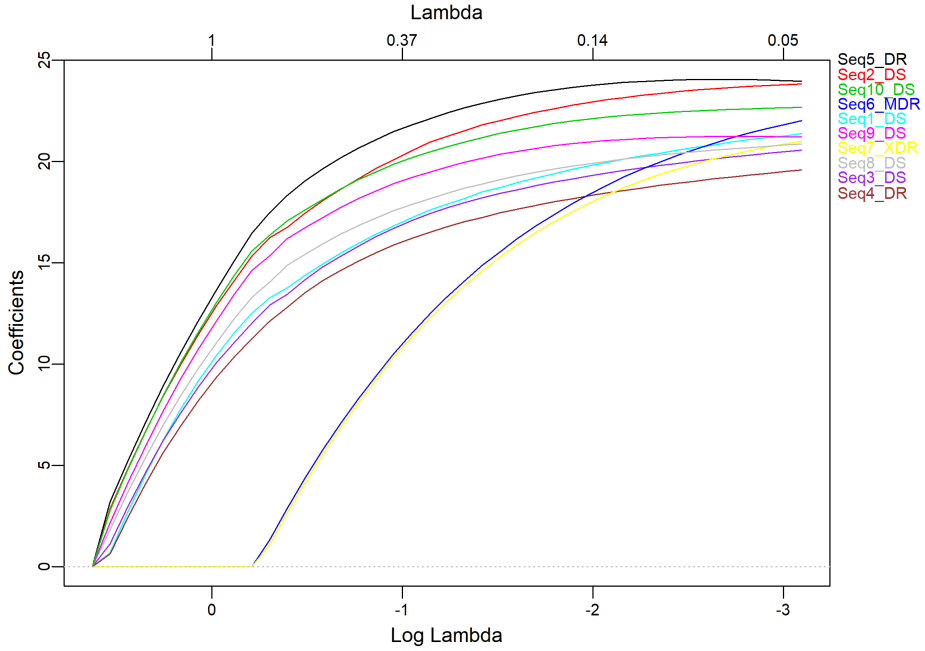
$$\hat{\beta} = \arg \min_{\beta} \{L(\lambda_1, \lambda_2, \beta)\}. \tag{2.24}$$

Esse procedimento pode ser visto como um método de mínimos quadrados penalizado. Seja $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, então resolvendo $\hat{\beta}$ na equação (2.23) é equivalente ao problema otimizado

$$\hat{\beta} = \arg\min_{\beta} |y - X\beta|^2, \text{ sujeito a } (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \leq t \text{ para algum } t. \tag{2.25}$$

Na Figura 2.8 temos o exemplo da aplicação do método *Elastic net* nas dez sequências do genoma da *mycobacterium tuberculosis*. Nesse gráfico conseguimos ver a formação dos grupos de sequências similares. Verificando o gráfico no eixo x, da esquerda para a direita, o primeiro grupo contém as sequências (Seq5_DR, Seq2_DS, Seq10_DS, Seq1_DS, Seq9_DS, Seq8_DS, Seq3_DS e Seq4_DR) e o segundo grupo contém as sequências (Seq6_MDR e Seq7_XDR).

Figura 2.7 – *Elastic net*.



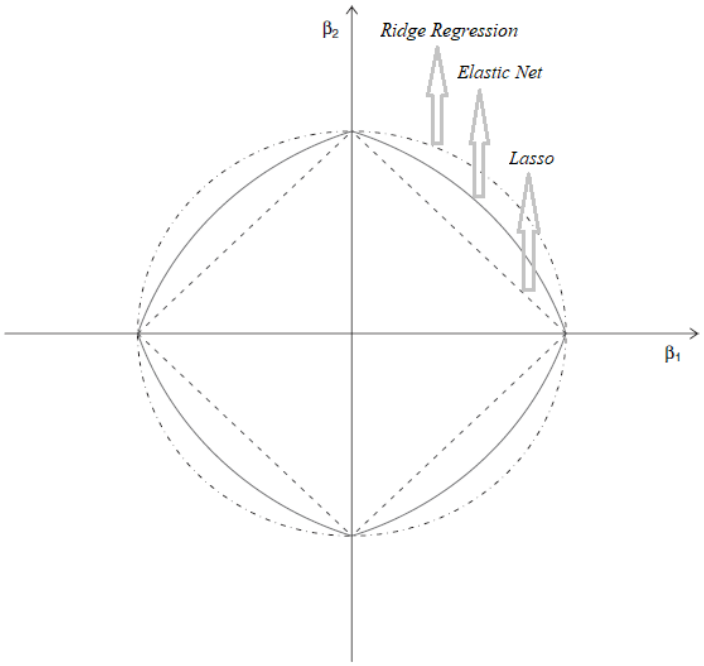
A função $(1-\alpha)|\beta|_1 + \alpha|\beta|^2$ é chamada de penalidade *Elastic Net* e é uma combinação convexa entre as penalidades que definem a estimação *Lasso* e *Ridge*, respectivamente. Quando $\alpha = 1$ o *Elastic Net* torna-se uma regressão *Ridge* simples. Quando $\alpha = 0$ temos a penalidade *Lasso* que é convexa, mas não estritamente convexa. Quando $\alpha = 0,5$ temos a penalidade *Elastic net*. Esses argumentos podem ser vistos na Figura 2.7.

Ridge regression (mantém todos os preditores):

$$\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \sum_j \beta_j^2 \right\} \tag{2.26}$$

Lasso (mantém os preditores mais significativos e remove os outros):

Figura 2.8 – Geometria das penalidades.



Fonte: Zou e Hastie (2005)

$$\hat{\beta} = \arg\ \min \left\{ \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \mu \sum_j |\beta_j| \right\} \tag{2.27}$$

Elastic net (uma combinação do *Ridge regression* e *Lasso*):

$$\hat{\beta} = \arg\ \min \left\{ \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \left((1-\alpha) \sum_j |\beta_j| + \alpha \sum_j \beta_j^2 \right) \right\} \tag{2.28}$$

2.4 Memória longa e expoente de Hurst

De acordo com Beran et al. (2013), os processos de memória longa (ou fractais) são conhecidos por desempenharem um papel importante em muitas disciplinas científicas e em campos aplicados, tais como: Física, Geofísica, Hidrologia, Economia, Finanças, Climatologia, Ciências Ambientais, Biologia, Medicina, Telecomunicações, Engenharia de Rede, entre outros. Existem várias razões para a ocorrência onipresente (presente em todos os lugares ao mesmo tempo) de processos no domínio de memória longa. Primeiro de tudo, a escala hiperbólica ocorre naturalmente (modificações por funções que variam lentamente) nos teoremas de limites para somas parciais, uma vez que os processos limitantes são necessariamente auto-similares. O fato é que no mundo dos processos estocásticos, processos auto-similares desem-

penham o mesmo papel fundamental de distribuições estáveis (incluindo a normal) no âmbito das distribuições dimensionais finitas. Fenômenos de escala hiperbólica também são um ingrediente essencial na Física e na Estatística (cuja relação é chamada de grupo de renormalização). Isto está parcialmente ligado ao papel de processos auto-similares em teoremas de limite. Outra razão para a ocorrência de fenômenos de memória longa é a agregação, que juntamente com a heterogeneidade é uma explicação frequente da dependência de longo alcance em um contexto econômico. Em redes de telecomunicações e computadores, as propriedades de distribuição dos tempos de espera podem levar a resultados semelhantes. Finalmente, há também uma conexão com os fractais (embora nem sempre seja direta, dependendo de suposições distributivas mais específicas).

Mesmo que a noção de memória longa e tópicos relacionados possam ser rastreados no início do século 20 ou até o final do século 19, é justo dizer que o assunto chamou a atenção de uma audiência matemática mais ampla (e, em particular, probabilistas e estatísticos) pelo trabalho pioneiro de Mandelbrot e colaboradores. Outro papel pioneiro semelhante pode ser atribuído a Granger em Economia, para Dobrushin (e antes, para Kolmogorov) em Física e, até mais cedo, para Hurst em Hidrologia. Essas contribuições iniciais motivaram vários probabilistas eminentes a desenvolver uma teoria de processos estocásticos no reino da auto-similaridade estocástica, leis de escala e teoremas de limites não padronizados (BERAN et al., 2013).

Iremos destacar aqui o expoente de Hurst, que é usado como uma medida de memória longa. Estudos relacionados com o expoente de Hurst foram originalmente desenvolvidos em Hidrologia para questão prática de determinar o dimensionamento ótimo da barragem para as condições de chuva volátil (grande volume de chuva em curto prazo) e seca do rio Nilo que foram observados durante um longo período. O nome “expoente de Hurst” ou “coeficiente de Hurst” deriva de Horold Edwin Hurst (1880-1978), que foi o principal pesquisador nesses estudos, onde o uso da notação padrão H para o coeficiente relaciona-se também com o seu nome.

De acordo com Esposti, Ferrario e Signorini (2008), no contexto dos processos estocásticos, a auto-similaridade é definida em termos da distribuição do processo. Seja $Y(t)$ um processo estocástico (t é um parâmetro de tempo contínuo), $Y(t)$ é chamado de auto-similar com o parâmetro de auto-similaridade H (expoente de Hurst), se para qualquer fator de alongamento positivo c, o processo reescalonado com escala de tempo ct é igual em distribuição ao processo original $Y(t)$, ou seja,

$$Y(t) \stackrel{d}{=} c^{-H} Y(ct), \tag{2.29}$$

em que $\stackrel{d}{=}$ significa igual em termos de distribuição (BERAN, 1994).

É fácil verificar se um processo auto-similar como definido na Equação (2.29), não é estacionário, isto é, diverge na distribuição $\left[Y(t) \stackrel{d}{\to} \infty\right]$ para $t \to \infty$, a menos que $H = 0$. Note que, para esses processos, a não estacionaridade, na verdade deriva da auto-similaridade, Equação (2.29), como a variância diverge com o tempo.

Com a consciência de que a auto-similaridade surge de uma maneira natural a partir de teoremas de limites para somas de variáveis aleatórias, o interesse aumentou em direção aos processos auto-similares estocásticos.

Na verdade, os processos auto-similares mais estudados e interessantes são aqueles com incrementos estacionários. Se considerarmos a sequência de incrementos de um processo auto-similar $X_i = Y_i - Y_{i-1}$, encontramos que a função de correlação é

$$p(k) = \frac{1}{2}\left[(k+1)^{2H} - 2k^{2H} + (k+1)^{2H}\right] \tag{2.30}$$

para $k \geq 0$ e $p(k) = p(-k)$ para $k < 0$.

O comportamento assintótico de $p(k)$ segue uma expansão de Taylor: primeiro nota-se que $p(k) = \frac{1}{2}k^{2H} g\left(k^{-1}\right)$ onde $g(x) = (1+x)^{2H} - 2 + (1-x)^{2H}$. Se $0 < H < 1$ e $H \neq 1/2$, então o primeiro termo não-zero na expansão de Taylor de $g(x)$, expandido na origem, é igual a $2H(2H-1)x^2$. Portanto, com k tendendo para infinito, $p(k)$ é equivalente a $H(2H-1)k^{2H-2}$, ou seja,

$$p(k) / \left[H(2H-1)k^{2H-2}\right] \to 1 \tag{2.31}$$

com $k \to \infty$

Podemos ver que para $1/2 < H < 1$ a correlação decai para zero tão lentamente que $\left[\sum_{k=-\infty}^{\infty} p(k) = \infty\right]$. O lento declínio implica que não existe um tempo de correlação correto e, assim, o valor atual da série é afetado não apenas por seus valores mais recentes, mas também por seu histórico de longo prazo. Por esse motivo, o processo X_i é muitas vezes referida como "memória longa" ou "dependência de longo alcance".

Para $H = 1/2$ as observações não são correlacionadas, e para $0 < H < 1/2$ o processo tem dependência de curto alcance, onde as correlações somam zero.

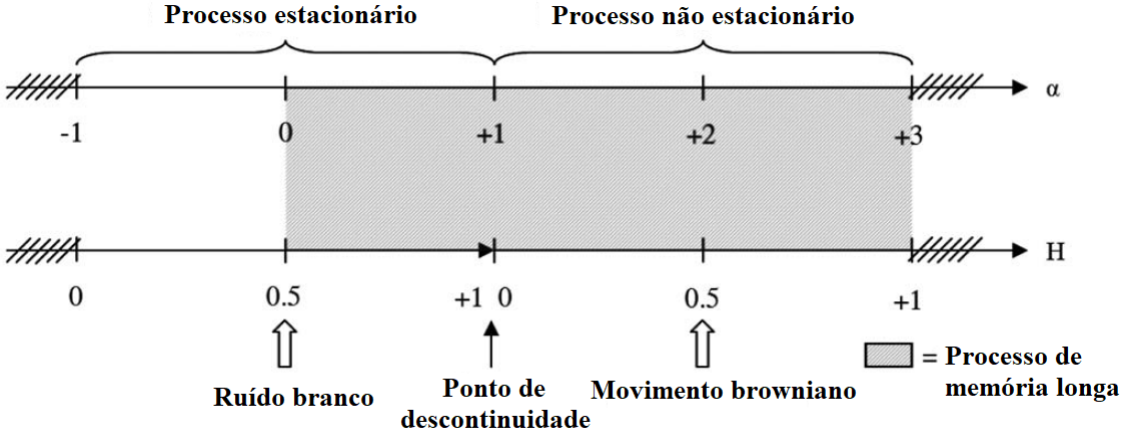
A definição de autocorrelação $[-1 \leq p(k) \leq 1]$ dada na Equação (2.31) força H para o intervalo $0 < H < 1$. Por outro lado, para $H > 1$, a Equação (2.31) deve divergir para infinito, o que contradiz a definição. Finalmente, para o processo de incremento X_i, pode-se demonstrar que o comportamento da densidade espectral perto da origem é

$$f(\lambda) \approx c_f |\lambda|^{1-2H} \quad \text{para} \quad \lambda \to 0. \tag{2.32}$$

De acordo com a definição de um processo estacionário com memória longa ou dependência de longo alcance, ou com correlações lentamente decadentes, encontramos novamente a mesma condição $1/2 < H < 1$. Além disso, o comportamento de um processo estacionário, X_i, próximo à origem, é dimensionado em um gráfico log-log como $1/f^{\alpha}$ com $\alpha = 2H - 1$.

Como o processo auto-similar, Y_i, deriva da integração dos incrementos, o processo Y_i próximo à origem é dimensionado em um gráfico log-log como $1/f^{\alpha}$ com $\alpha = 2H + 1$ de acordo com a Figura 2.9.

Figura 2.9 – Relação entre α e H. Em particular $\alpha = 2H - 1$ para um processo estacionário e $\alpha = 2H + 1$ para um processo não estacionário.



Fonte: Esposti, Ferrario e Signarini (2008)

A literatura estatística reserva os termos de "memória longa" e "intervalo de dependência longa" para processos estacionários $(0 < \alpha < 1)$ e para processos não estacionários cujos incrementos estacionários possuem uma memória longa $(2 < \alpha < 3)$. No entanto, processos não estacionários com $(1 < \alpha < 2)$ têm memória praticamente mais forte do que aqueles com

$(0 < \alpha < 1)$. Assim, na prática, eles também são referidos e tratados como processos de "memória longa".

A seguir são apresentados alguns métodos para estimação do parâmetro H.

2.4.1 Método de variância agregada

De acordo com Beran (1989), uma propriedade marcante dos processos de memória longa, é que a variância da média da amostra converge para zero mais lento que a taxa N^{-1}, onde N é o tamanho da amostra. Assim,

$$Var\left(\bar{X}_N\right) \sim cN^{2H-2}, \tag{2.33}$$

para N grande, onde $c > 0$ e $\bar{X}_N$ é a média amostral. Isto sugere o seguinte método para estimar H. Divida a série em N/m blocos de tamanho m, calcule a média amostral é dada por

$$\bar{X}_m\left(k\right) = \frac{1}{m}\sum_{t=(k-1)m+1}^{km} X\left(i\right), \;\; k = 1,2,...,N/m \tag{2.34}$$

em cada bloco e a variância amostral

$$s^2\left(m\right) = \left(N/m-1\right)^{-1}\sum_{k=1}^{N/m}\left(\bar{X}_m\left(k\right)-\bar{X}_N\right)^2, \tag{2.35}$$

em que $\overline{X}_N$ denota a média geral. Representando em um gráfico $\log s^2\left(m\right)$ versus $\log\left(m\right)$ deve render pontos espalhados ao longo de uma linha reta com inclinação igual a $2H-2$.

2.4.2 Método da variância agregada diferenciada

Teverovsky e Taqqu (2001) propuseram um método para detectar a dependência de longo alcance, mesmo na presença de não estacionariedade. É um tipo de estimador da variância obtido pegando o logaritmo da primeira diferença da Equação (2.34),

$$\log\Delta Var\left[\bar{X}_m\left(k\right)\right] \sim \log\frac{d}{dm}Var\bar{X}_m\left(k\right) + \log\Delta m. \tag{2.36}$$

em que,

$$\frac{d}{dm}Var\left[\bar{X}_m\left(k\right)\right] \sim \left(2H-2\right)Cm^{2H-3}, \tag{2.37}$$

em que, uma vez que os valores m são logaritmicamente espaçados, temos

$$\Delta \log(m) = \text{constante; isto é,}$$
$$\log \Delta m = \log m + C_1.$$

Assim,

$$\begin{aligned} \Delta Var[\bar{X}m(k)] &= (2H-3)\log m + \\ &\quad + \log(2H-2)C + \log m + C_1 \\ &= (2H-2)\log m + C_2 \end{aligned} \tag{2.38}$$

Assim, em um gráfico de log-log, esperamos ver uma linha reta com um declive igual a $2H-2$.

2.4.3 Valor absoluto agregado *(AM)*

Considere a série da média definida na Equação (2.34) e calcule seu n-ésimo momento absoluto

$$AM_n^{(m)} = \frac{1}{(N/m)} \sum_{k=1}^{(N/m)} \left| \bar{X}_k^{(m)} - \bar{X} \right|^n, \tag{2.39}$$

em que $AM_n^{(m)}$ é assintoticamente proporcional a $m^{n(H-1)}$.
Para encontrar uma estimativa para H:

- Calcule $AM_n^{(m)}$ para diferentes valores de m,
- Representar um gráfico log-log contra m e,
- O ponto deve ser espalhado ao longo de uma linha com declive $n(H-1)$.

2.4.4 Método Peng

De acordo Adler, Feldman e Taqqu (1998) esse método segue os seguintes passos:

- Calcule a soma parcial dentro de cada bloco de tamanho m, dada por

$$Y(k)^m = \sum_{t=(k-1)m+1}^{km} X_t, \;\; k = 1,2,...,(N/m) \tag{2.40}$$

- Ajustar uma regressão linear $y = a + bk$,
- Calcular a variância do resíduo $s_r^{(m)} = \frac{1}{m}\sum_{k=1}^{N/m}\left(y(k) - a - bk\right)^2$,
- Criar um gráfico de $\log s_r^{(m)}$ *vs* $\log m$,
- A inclinação deve ser $2H$.

2.4.5 Método *R/S*

O método R/S, de acordo com Beran (1994), corresponde:

Considere

$$Y_T = \sum_{t=1}^{T} X_t \tag{2.41}$$

Defina o intervalo ajustado:

$$R(t,k) = \max_{0\le i\le k}\left[Y_{t+i} - Y_t - \frac{i}{k}\left(Y_{t+1} - Y_t\right)\right] - \min_{0\le i\le k}\left[Y_{t+i} - Y_t - \frac{i}{k}\left(Y_{t+k} - Y_t\right)\right] \tag{2.42}$$

em que

$$S(t,k) = \sqrt{k^{-1}\sum_{i=t+1}^{t+k}\left(X_i - \bar{X}_{t,k}\right)^2}, \tag{2.43}$$

com $\bar{X}_{t,k} = k^{-1}\sum_{i=t+1}^{t+k} X_i$.

A razão padronizada

$$Q(t,k) = \frac{R(t,k)}{S(t,k)} \tag{2.44}$$

é conhecida como intervalo ajustado escalado ou estatística R/S.

Para os dados do rio Nilo, Hurst (1951) observou que, para um k grande,

$$\log E\left(R/S\right) \approx a + H\log k, \tag{2.45}$$

com $H > \frac{1}{2}$.

Com base nos resultados empíricos de Hurst, pode-se:

- Dividir a série em k blocos de tamanho N/k,
- Calcular a estatística $R/S\,Q(t_i,k)$, como definido na Equação (2.44), com valores iniciais $t_i = iN/k+1$ para todos os possíveis k, de tal modo que $t_i + k < N$,
- Gráfico do seu logaritmo contra o logaritmo de k,
- A inclinação estimada a partir da regressão será então estimada de H.

2.5 Ondaletas aplicada na Genética

Nos últimos anos as ondaletas vem ganhando cada vez mais espaço nas análises genéticas. Na literatura, destaca-se que uma de suas grandes vantagens está na redução do dimensionamento do conjunto de dados de microarranjos, pois um dos grandes problemas da análise genética corresponde ao elevado tamanho do conjunto de dados (FARHADIAN et al., 2014), (KIM et al., 2010; PRABAKARAN; SAHU; VERMA, 2008; LIU, 2009). A transformada de ondaleta mostrou-se também bastante eficaz para selecionar os genes discriminativos em dados de microarranjos (MEHER et al., 2012; NANNI; LUMINI, 2011). Atualmente, vem se tornando cada vez mais popular no campo da bioinformática, devido principalmente à sua capacidade em análises de multirresolução e localização espaço-frequência (LI; LIAO; KWOK, 2006). Proporciona uma avaliação rápida e sensível da importância biológica do DNA, que representa a significância do genoma (HAIMOVICH et al., 2006).

As ondaletas apresentam uma melhor capacidade para capturar componentes ocultos de dados biológicos e uma eficiente ligação entre sistemas biológicos e os objetos matemáticos usados para descrevê-las (LIÒ, 2003).

A utilização da ondaleta Haar aplicada no sinal do DNA, consiste num método que divide o banco de dados em agrupamentos com as mesmas tendências. Esses agrupamentos podem ser classificados como grupos que permitem ao pesquisador a identificação de sequências, sendo essa uma das principais vantagens da metodologia proposta, a qual é considerada uma técnica de mineração do DNA (EL-ZANATY et al., 2011). Outra vantagem da aplicação da análise de ondaletas, está na transformação de dados obtidos sob diferentes condições de crescimento. Com isso permite-se a comparação de padrões de expressões adquiridos em experiências que tem deslocamentos no tempo (SONG et al., 2007).

A aplicação de transformadas discretas decimadas e não-decimadas de ondaletas vem auxiliando grandemente a verificação de similaridades das sequências genômicas, mostrando-

se uma ferramenta bastante útil no processamento dessas análises, tendo um retorno rápido dos resultados, sendo que esses processos duravam de 3 a 4 semanas com métodos laboratoriais convencionais (SAINI; DEWAN, 2016; VANNUCCI; LIÒ, 2001).

2.6 Dados genômicos

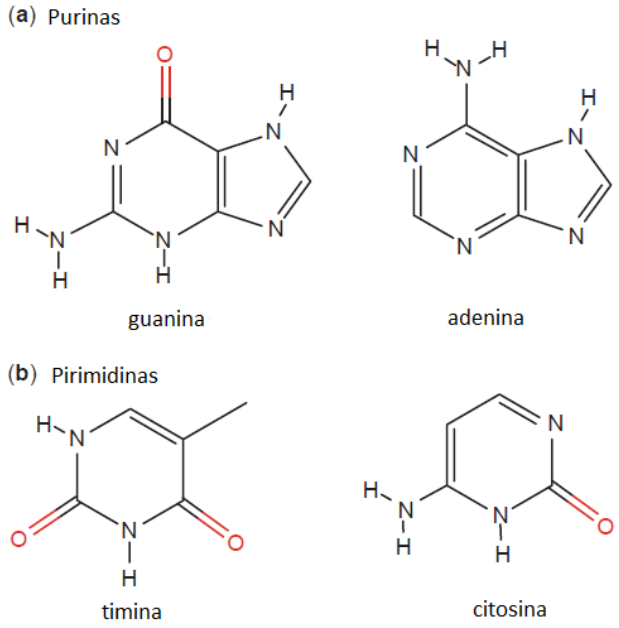
Um dos marcos históricos da genética correspondem aos experimentos do austríaco Gregor Mendel, pois seus estudos levaram a identificação de fatores responsáveis pelos traços hereditários dos organismos vivos, os quais foram chamados genes. Futuramente, descobriu-se que haviam estruturas dentro da célula chamadas cromossomos, que duplicavam-se durante a divisão da célula. Associando então esses dois fatos lançou-se a hipótese, confirmada posteriormente, de que a herança genética é carregada pelos genes arranjados nos cromossomos. Assim, os cromossomos correspondem a uma longa sequência de DNA, que contém vários genes, e outras sequências de nucleótidos, com funções específicas nas células dos seres vivos. Sob a ótica molecular, os genes podem ser considerados como segmentos contíguos e discretos de uma molécula de DNA, onde estão armazenadas as informações genéticas, que correspondem ao genoma (OKURA, 2002).

O DNA é composto por quatro bases nitrogenadas (A, C, G, T), que correspondem a Adenina, Citosina, Guanina e Timina. Essas bases são classificadas em dois grupos: purinas e pirimidinas (PEVSNER, 2009).

As purinas, compostas por Adenina (A) e Guanina (G) são maiores e contém mais de um único anel. Já as pirimidinas, Citosina (C) e Timina (T) são menores e compostas por um único anel, conforme ilustrado na Figura 2.10.

Uma sequência de DNA pode ser de uma cadeia simples ou de uma cadeia dupla. Naturalmente uma cadeia dupla de DNA ocorre pelo emparelhamento de bases. Quando temos uma cadeia dupla, as duas cadeias têm sentido oposto e são complementares uma para com a outra. Essa complementaridade significa que, para cada A, C, G, T (que são nucleotídeos do DNA), em uma cadeia, existe um T, G, C, A, respectivamente, na outra cadeia. Logo, como os cromossomos são de cadeia dupla, temos então a “dupla hélice”. As informações acerca de um gene podem estar contidas em qualquer uma das cadeias. A importância desse emparelhamento está na introdução de uma redundância completa na codificação, que permite a célula reconstituir o genoma completo a partir de apenas um filamento, o que por sua vez possibilita a replicação fiel (CRISTIANINI; HAHN, 2006).

Figura 2.10 – Bases nitrogenadas.



Fonte: Pevsner (2009)

Os dados sobre as características estruturais e funcionais do genoma de vários organismos estão sendo acumulados e analisados em laboratórios de todo o mundo. O volume de dados genômicos se expandiram a uma enorme e contínua taxa de crescimento, enquanto suas propriedades e relacionamentos fundamentais ainda não são totalmente conhecidos e estão sujeitos a revisão contínua. Um sistema em todo o mundo para reunir essas informações genômicas se concentram no Centro Nacional de Informação da Biotecnologia (NCBI) e em várias outras bases de dados genômicos com grande integração entre elas (DOUGHERTY et al., 2005).

2.6.1 Bactérias e a Genética

O estudo de bactérias e bacteriófagos são essenciais para o conhecimento em muitas áreas da Genética. Por exemplo, muito do que se sabe sobre a expressão e regulação da informação genética foi inicialmente derivada de trabalhos experimentais com elas. Além disso, o amplo conhecimento sobre bactérias resistentes tem servido como base para o uso de clonagens do DNA e procedimentos de recombinações (KLUG et al., 2012).

A importância das bactérias na pesquisa genética em organismos, baseia-se em duas características principais. Em primeiro lugar, elas têm um ciclo reprodutivo extremamente curto. Literalmente centenas de gerações, no montante de bilhões de bactérias geneticamente idênticas, podem ser reproduzidas em curtos períodos de tempo. Em segundo lugar, elas também

podem ser estudadas em uma cultura pura, ou seja, uma única espécie ou uma cepa mutante de bactérias com facilidade de serem isoladas e investigadas independentemente dos outros organismos semelhantes. Como resultado, elas têm sido indispensáveis para os progressos realizados em genética ao longo do último meio século (KLUG et al., 2012).

2.6.2 Organização do genoma bacteriano

As bactérias possuem material genético, o qual é transmitido aos descendentes no momento da divisão celular. Esse material genético não está contido dentro de um núcleo, portanto o genoma desses microrganismos está disperso no citoplasma (MACÊDO, 2016).

O genoma bacteriano é condensado e organizado em uma estrutura denominada nucleóide. Esse nucleóide ou cromossomo bacteriano é constituído por uma única molécula de DNA de fita dupla, circular, não delimitada por membrana nuclear, e é capaz de autoduplicação. Como características tem-se que seus genes contêm todas as informações necessárias à sobrevivência da célula; possuem apenas uma cópia de seu cromossomo, sendo portanto haplóides; apresentam a propriedade de replicação e transmissão das moléculas hereditárias durante a divisão celular (MACÊDO, 2016).

2.6.3 GC *content*

A GC *content* é um importante parâmetro de genomas bacterianos que vem sendo usado para escanear a composição básica do genoma, bem como para compreender a evolução da sequência codificada. Historicamente essa taxa é representada num intervalo de 25% a 75% em genomas bacterianos (MANN; CHEN, 2010).

A maioria dos artigos relatam as frequências agregadas para G e C (chamada de GC *content*), versus as frequências agregadas para A e T (chamada de AT *content*). Como essas duas quantidades são necessárias para sempre somarem 1, apenas a GC *content* é tipicamente relatada. A motivação para sempre relatar simplesmente a GC *content* é devido a uma série de reações químicas, onde a GC *content* em um genoma é frequentemente muito semelhante a AT *content*. Logo, somente um valor terá de ser comunicado, ao invés de quatro (CRISTIANINI; HAHN, 2006).

Um genoma mostra variações significativas em seu GC *content* ao longo da região de sua sequência, em contraste ao conhecimento de todo o genoma. Verificando a análise de sequências de DNA a partir de vários bancos de dados genômicos estratificados de acordo com a GC

content, observa-se que a codificação de sequências mais longas em vertebrados e genes de procariotas apresentam uma GC-*rich* , enquanto que em sequências mais curtas tem-se CG-*poor* (OLIVER; MARÍN, 1996). As regiões GC-*rich* incluem muitos genes codificadores de proteínas, e portanto a determinação da proporção de GC ajuda na identificação de regiões ricas em genes do genoma. GC *content* para toda a sequência é calculada como a razão da soma de bases G, C, sob a soma das bases A, G, C, T, ou seja,

$$GCcontent = \frac{nG + nC}{nA + nG + nC + nT}, \tag{2.46}$$

em que nA, nG, nC, nT representam o número de bases de nucleóide A, G, C, T, respectivamente, em uma sequência. A GC *content* também pode ser calculada para uma parte da sequência usando a técnica de janela, em que a GC *content* é calculada para um comprimento fixo de uma específica janela da sequência (SAINI; DEWAN, 2016).

Um dos métodos que vem sendo usado para detectar padrões de G + C no genoma de bactérias é o de ondaletas. As ondaletas de encolhimento são aplicadas para eliminar pequenas variações na GC *content*. Com isso, possíveis picos no escalograma da ondaleta representada são usados para localizar regiões do genoma com grande variação na GC *content* (VANNUCCI; LIÒ, 2001).

REFERÊNCIAS

ABU-JAMOUS, B.; FA, R.; NANDI, A. K. **Integrative cluster analysis in bioinformatics**. India: John Wiley & Sons, Ltd, 2015.

ADLER, R. J.; FELDMAN, R. E.; TAQQU, M. S. **A pratical guide to heavy tails: statistical techniques and applications**. USA: Birkhäuser, 1998.

ALDENDERFER, M. S.; BLASHFIELD, R. K. **Cluster analysis: quantitative applications in the social sciences**. Beverly Hills: Sage Publication, 1984.

ANDERBERG, M. R. **Cluster analysis for applications**. London: Academic Press, 1973.

BARBOSA, A. C. B.; BLITZKOW, D. **Ondaletas: Histórico e aplicação**. São Paulo, 2008.

BERAN, J. A test of locationf or data with slowlyd ecayings erialc orrelations. **Biometrika**, v. 76, p. 261–269, 1989.

BERAN, J. **Statistics for Long-Memory Processes**. New York: Chapman & Hall, 1994.

BERAN, J. et al. **Long-memory processes: probabilistic properties and statistical methods**. London: Springer, 2013.

CRISTIANINI, N.; HAHN, M. W. **Introduction to Computacional Genomics**. New York: Cambridge University Press, 2006.

DAUBECHIES, I. **Ten Lectures on Wavelets**. Philadelphia: Society for Industrial and Applied Mathematics, 1992.

DOUGHERTY, E. R. et al. **Genomic Signal Processing and Statistics**. New York: Hindawi Publishing Corporation, 2005.

EL-ZANATY, M. et al. Haar wavelet transform of the signal representation of DNA sequences. **International Journal of Computer Science and Communication Security**, v. 1, p. 56–62, 2011.

ESPOSTI, F.; FERRARIO, M.; SIGNORINI, M. G. A blind method for the estimation of the hurst exponent in time series: theory and application. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, v. 18, n. 3, p. 033126, 2008.

EVERITT, B. S. et al. **Cluster analysis**. Índia: John Wiley & Sons, 2011.

FARHADIAN, M. et al. Supervised wavelet method to predict patient survival from gene expression data. **The Scientific World Journal**, p. 10, 2014.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. **Journal of statistical software**, v. 33, n. 1, p. 1–20, 2010.

GENÇAY, R.; SELÇUK, F.; WHITCHER, B. **An Introduction to Wavelets and Other Filtering Methods in Finance and Economics**. San Diego, California: Academic Press, 2002.

HAIMOVICH, A. D. et al. Wavelet analysis of DNA walks. **Journal of Computational Biology**, v. 13, n. 7, p. 1289–1298, 2006.

HENNIG, C. et al. **Handbook of cluster analysis**. New York: CRC Press, 2016.

HERNANDEZ, E.; WEISS, G. **A First Course on Wavelets**. USA: CRC Press LLC, 1996.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. **Technometrics**, v. 12, n. 1, p. 55–67, 1970.

HURST, H. E. Long-term storage capacity of reservoirs. **Transactions of the American Society of Civil Engineers**, v. 116, n. 1, p. 770–799, 1951.

KAISER, G. **A Friendly Guide to Wavelets**. Cambridge: Birkhäuser, 1994.

KIM, B.-R. et al. Wavelet-Based functional clustering for patterns of high-dimensional dynamic gene expression. **Journal of Computational Biology**, v. 17, n. 8, p. 1067–1080, 2010.

KLUG, W. S. et al. **Concepts of Genetics**. 10. ed. San Francisco, Califórnia: Pearson Benjamin Cummings, 2012.

LI, S.; LIAO, C.; KWOK, J. T. **Wavelet-based feature extraction for microarray data classification**. Vancouver, Canada, 2006.

LIMA, P. C. Wavelets: uma introdução. **Matemática Universitária**, n. 33, p. 13–44, dez 2002.

LIÒ, P. Wavelets in bioinformatics and computational biology: state of art and perspectives. **Bioinformatics Review**, v. 19, p. 2–9, 2003.

LIÒ, P.; VANNUCCI, M. Finding pathogenicity islands and gene transfer events in genoma data. **Bioinformatics**, v. 16, n. 10, p. 932–940, 2000.

LIU, Y. Wavelet feature extraction for high-dimensional microarray data. **Neurocomputing**, v. 72, p. 985–990, 2009.

MACÊDO, M. M. da S. **Genética Bacteriana**. Universidade Federal de Campina Grande, 2016. Disponível em: <http://pt.slideshare.net/kaiorochars/gentica-bacteriana-14836336>.

MALLAT, S. Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. **Transactions of the American Mathematical Society**, v. 315, p. 69–87, 1989.

MALLAT, S. **A Wavelet Tour of Signal Processing**. 2. ed. San Diego, California: Academic Press, 1999.

MANLY, B. F. J.; ALBERTO, J. A. N. **Multivariate statistical methods: a primer**. Boca Raton: Chapman and Hall/CRC, 2016.

MANN, S.; CHEN, Y.-P. P. Bacterial genomic G+C composition-eliciting environmental adaptation. **Genomics**, v. 95, p. 7–15, 2010.

MATTHEW, P. K.; YAHAYA, A. Performance analysis on least absolute shrinkage selection operator, elastic net and correlation adjusted elastic net regression methods. **International Journal of Advanced Statistics and Probability**, v. 3, n. 1, p. 93–99, 2015.

MCLACHLAN, G. J. Mahalanobis distance. **Resonance**, v. 4, n. 6, p. 20–26, 1999.

MEHER, J. et al. Cascaded factor analysis and wavelet transform method for tumor classification using gene expression data. **I.J. Information Technology and Computer Science**, v. 9, p. 73–79, 2012.

MORETTIN, P. A. **Ondas e Ondaletas: da Análise de Fourier à Análise de Ondaletas**. São Paulo: Editora da Universidade de São Paulo-EDUSP, 1999. 269 p.

MORETTIN, P. A.; PINHEIRO, A.; VIDAKOVIC, B. **Wavelets in functional data analysis**. Switzerland: Springer International Publishing, 2017.

NANNI, L.; LUMINI, A. Wavelet selection for disease classification by dna microarray data. **Expert Systems with Applications**, v. 38, p. 990–995, 2011.

NASON, G. P. **Wavelet Methods in Statistics with R**. New York: Springer, 2008.

OGDEN, R. T. **Essential wavelets for statistical applications and data analysis**. Boston: Birkhäuser, 1997.

OGUTU, J. O.; SCHULZ-STREECK, T.; PIEPHO, H.-P. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. **BMC Proceedings**, v. 6, n. 2, p. 1–6, 2012.

OKURA, V. K. **Bioinformática de Projetos Genoma de Bactérias**. Dissertação (Mestrado) — Universidade Estadual de Campinas, Campinas, 2002.

OLIVEIRA, K. F. de. **Análise da Transformada Wavelet Direcional Adaptativa na Codificação de Imagens**. Dissertação (Mestrado) — Universidade de Brasília, Brasília, 2009.

OLIVER, J. L.; MARÍN, A. A relationship between GC content and coding-sequences length. **Molecular Evolution**, v. 43, p. 216–223, 1996.

PERCIVAL, D. B.; WALDEN, A. T. **Wavelets methods for time series analysis**. Cambridge: Cambridge University Press, 2000.

PEVSNER, J. **Bioinformatics and Functional Genomics**. 2. ed. Hoboken, New Jersey: John Wiley & Sons, 2009.

PRABAKARAN, S.; SAHU, R.; VERMA, S. A wavelet approach for classification of microarray data. **International Journal of Wavelets, Multiresolution and Information Processing**, v. 6, n. 3, p. 375–389, 2008.

SAINI, S.; DEWAN, L. Application of discrete wavelet transform for analysis of genomic sequences of mycobacterium tuberculosis. **SpringerPlus**, v. 5, n. 1, 2016.

SONG, J. Z. et al. The wavelet-based cluster analysis for temporal gene expression data. **Journal on Bioinformatics and Systems Biology**, p. 7, 2007.

TEVEROVSKY, V.; TAQQU, M. Testing for long-range dependence in the presence of shifting means or a slowly declining trend using a variance type estimator. **Journal of Time Series Analysis**, v. 18, p. 279–304, 2001.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society**, v. 58, n. 1, p. 267–288, 1996.

UZINSKI, J. C. **Momentos Nulos e Regularidade Wavelet na Deteccção de Falhas em Sinais**. Dissertação (Mestrado) — Universidade Estadual Paulista Júlio Mesquita Filho, Ilha Solteira - SP, 2013.

VANNUCCI, M.; LIÒ, P. Non-decimated wavelet analysis of biological sequences: applications to protein structure and genomics. **Sankhyā: The Indian Journal of Statistics**, v. 63, p. 218–233, 2001.

VIDAKOVIC, B. **Statistical Modeling by Wavelets**. USA: John Wiley & Sons, 1999.

WOJTASZCZYK, P. **A Mathematical Introduction to Wavelets**. New York: Cambridge University Press, 1997.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 67, p. 301–320, 2005.

SEGUNDA PARTE

ARTIGO 1

Evaluation of genome similarities using the non-decimated wavelet transform

Redigido conforme as normas da revista Genetics and Molecular Research (versão publicada)

GMR

Evaluation of genome similarities using the non-decimated wavelet transform

L.M. Ferreira, T. Sáfadi and R.R. Lima

Departamento de Estatística, Universidade Federal de Lavras, Lavras, MG, Brasil

Corresponding author: T. Sáfadi
E-mail: safadi@des.ufla.br

Genet. Mol. Res. 16 (3): gmr16039758
Received June 23, 2017
Accepted August 24, 2017
Published September 21, 2017
DOI http://dx.doi.org/10.4238/gmr16039758

ABSTRACT. The wavelets have become increasingly popular in the field of bioinformatics due to their capacity in multiresolution analysis and space-frequency localization; the latter particularity is acquired due to a moving window that runs through the analyzed space. As a feature, they have a better ability to capture hidden components of biological data and an efficient link between biological systems and the mathematical objects used to describe them. The decomposition of signals/sequences at different levels of resolution allows obtaining distinct characteristics in each level. The energy (variance) obtained at each level provides a new set of information that can be used to search similarities between sequences. We show that the behavior of GC-content sequence can be succinctly described regarding the non-decimated wavelet transform, and we indicate how this characterization can be used to improve clustering of the similar strains of the genome of the *Mycobacterium tuberculosis*, having a very efficient level of detail. The clustering analysis using the energy obtained at each level of the analyzed sequences was essential to verify the dissimilarity of the sequences.

Key words: Non-decimated wavelet transform; Cluster analysis; Genome

INTRODUCTION

In the last decades, the analysis using technique of wavelets has been growing increasingly. One of the great advantages associated with this method corresponds to the computational gain, that is, the analyses are processed almost in real time. The applicability is in several areas of science, like Physics, Mathematics, Engineering, Genetics, among others.

The wavelet transform is a technique of seeing and represents a signal. Mathematically, it is represented by a function oscillating in time or space. As a characteristic, it has sliding windows that expand or compress to capture low- and high-frequency signals, respectively (Percival and Walden, 2000). Its origin occurred in the field of seismic study to describe the disturbances arising from a seismic impulse (Morlet et al., 1982).

Among the wavelet techniques, we have the discrete non-decimated wavelet transform (NDWT), whose main characteristic is that it can work with any size of signals/sequences. In this technique, the coefficients are translation invariants, that is, the choice of origin is irrelevant since all the observations are used in the analysis, a situation that does not occur in the discrete decimated wavelet transform (DWT).

Discrete wavelet transforms were used to identify gene locations in genomic sequences (Ning et al., 2003), identifying long-range correlations, locating periodicities in DNA sequences (Vannucci and Liò, 2001), and in the analysis of G+C patterns (Dodin et al., 2000).

The clustering analysis is often adopted to deal with DNA sequences efficiently. A wavelet-based feature vector model was proposed by Bao and Yuan (2015) for clustering of DNA sequences.

Human tuberculosis (TB) is caused by an intracellular pathogen, *Mycobacterium tuberculosis* (MTB) and it replicates rapidly in the lungs with high oxygen concentration. Global TB control measures are affected by the emergence of drug resistant (DR), multidrug resistant (MDR), and extensively drug resistant (XDR) strains. Resistance in these MTB strains to anti-TB drugs occurs due to chromosomal mutations (Saini and Dewan, 2016). Global control of tuberculosis is hampered by slow, insensitive diagnostic methods, particularly for the detection of DR forms and in patients with human immunodeficiency virus infection. Early detection is essential to reduce the death rate and interrupt transmission. Boehme et al. (2010) concerned with this situation and developed a more efficient method for the detection of DR and MDR strains. Perdigão et al. (2010) worked to characterize the genetic changes associated with the high number of XDR that threatens the global control of TB worldwide.

Apart from molecular methods based on whole genome sequences of MTB, signal processing of complete genomic sequences can help display and explore structural patterns capable of being interpreted and compared. Graphical representations obtained from signal processing methods can provide insight into the evolution, structure, and function of genomes (Anastassiou, 2000).

The genome of MTB is approximately 4.4 million base pairs long and is one of the largest known bacterial genomes. This bacterium is the cause of the TB disease that has killed thousands of people around the world. The GC-content is an important parameter of bacterial genomes used to scan the basic composition of the genome as well as to understand the evolution of the coded sequence.

Saini and Dewan (2016) highlights the potential of discrete wavelet transforms in the analysis and comparison of genomic sequences of MTB with different resistance characteristics. Based on the calculation of the energy of wavelet decomposition coefficients

of complete genomic sequences, they showed that the genomic sequences could be grouped into two broad categories wherein the DR and drug susceptible (DS) sequences formed one group while the MDR and XDR sequences formed the other group.

The main advantage of the NDWT method concerning DWT is the possibility to work with any size of genome sequences, that is, there is no loss of genome information. In the DWT method, the sequence must have a power of two, where loss of information of the genome inevitably occurs.

The NDWT method can be used in any genome type, increasing the speed of the analysis, because the analyses with this method are processed almost in real time.

In this study, the NDWT was applied to the GC-content sequences of the MTB genome strains and the energies obtained to the detail level coefficients were used to study their similarities. Stacked plots of the detail levels provide an effective means of exploring the relationships between genomic strains at different scales. The proposed methodology is applied to MTB sequences, being 4 DR, 4 DS, 1 MDR, and 1 XDR.

MATERIAL AND METHODS

The sequences analyzed correspond to the strains of the MTB genome. Ten sequences were analyzed, obtained from the National Center for Biotechnology Information (NCBI, 2017).

The GC-content of all the sequences was evaluated using a sliding window of 10,000 bases. The sequences were decomposed using a discrete NDWT. We considered the Daubechies wavelet (4 null moments) with 5 levels of decomposition. Statistical measures of energy for each of the decomposed sequences were evaluated. We used the measures of energy to know if the sequences were similar or not, and the clustering analysis was performed using the Mahalanobis distance in a hierarchical method with the average linkage.

The free software R (R Core Team, 2017) was used.

Wavelet

A wavelet function is the interpretation of a short-wave with rapid growth and decay. The theory is based on the representation of functions in different scales and resolutions (time-scale) that is considered one of its main characteristics (Daubechies, 1992).

In the analysis of wavelet, the window is oscillating and is called the mother wavelet. There are arbitrary translations and dilations. In this way, the mother wavelet generates other wavelets (Hernandez and Weiss, 1996).

By definition: a wavelet is a function $\psi(t) \in L^2(\mathbb{R})$, such that their family of functions is given by Equation 1:

$$\psi_{j,k} = 2^{-j/2}\psi\left(2^{-j}t - k\right) \quad \text{(Equation 1)}$$

where j and k are arbitrary integers on an orthonormal basis in Hilbert space $L^2(\mathbb{R})$ (Wojtaszczyk, 1997).

Pyramidal algorithm

The pyramidal algorithm, developed by Mallat (1989), uses the low-pass filter l_k

obtained using the scale function or father wavelet (ϕ) and the high-pass filter h_k obtained using the mother wavelet (ψ) with coefficients given by Equation 2:

$$l_k = \sqrt{2}\int_{-\infty}^{+\infty}\phi(t)\phi(2t-k)\,dt$$

$$h_k = \sqrt{2}\int_{-\infty}^{+\infty}\psi(t)\phi(2t-k)\,dt \quad \text{(Equation 2)}$$

The approximation coefficients (coarse scales) are obtained by a low-pass filter, and the coefficients of details (finer scales) are obtained by a high-pass filter.

Discrete NDWTs

The characteristic of the discrete NDWT is to keep the same amount of data in even and odd decimations on each scale and continue to do the same on each subsequent scale, being D_0 the dyadic decimation, D_1 the odd decimation, H the high-pass filter, and L the low-pass filter. Consider, for example, an input vector (y_1,..., y_n). Then, apply and keep both D_0H_y and D_1H_y, even and odd indexed of the observation-filtered wavelets. Each of these sequences is length $n/2$. Thus, in total, the number of wavelet coefficients in both decimals on the finer scale is 2 x n/2 = n (Nason, 2008).

Figure 1 shows that in the finer scale the coefficients of wavelets are d_0 and d_1. The next finer scales are d_{00}, d_{01}, d_{10}, d_{11}.

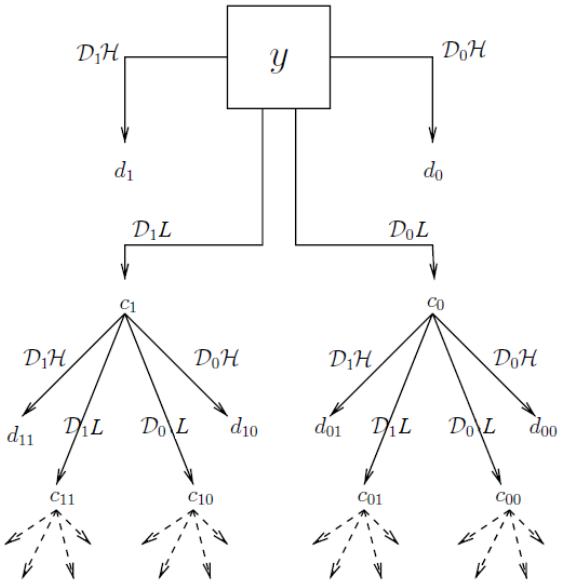


Figure 1. Diagram of the discrete non-decimated wavelet transform (NDWT). Source: Nason (2008).

In the discrete NDWT, the coefficients are translational invariants, that is, it means that the circular displacement of the data was reflected in the same direction of the coefficients. Another feature is the ability to handle data of an arbitrary size that does not require the sample size n to have a power of two, which is what occurs in the discrete DWT. The main advantage of this method is associated with zero pass filters, which means that it operates circularly to the data allowing functionalities at different scales to be aligned with the sequence of the original data (Vannucci and Liò, 2001).

Scalogram

The scalogram is a very effective tool for interpreting the sign of the wavelet. It is defined as a graph of the sum of squares of the wavelet coefficients at the different levels.

The energy *E* (j) for the coefficients, d_{jk}, of the wavelet on each level *j*, is given by:

$$E(j) = \sum_{k=0}^{n} d_{j,k}^2 \qquad j = 1, \ldots, J. \qquad \text{(Equation 3)}$$

Equation 3 corresponds to the calculation of the energy in the discrete NDWT on level *j* (Gençay et al., 2002).

Daubechies wavelets

The Daubechies wavelet is a family of orthogonal wavelets that define a discrete wavelet transformation and are characterized by a maximum number of null moments (degree of smoothing) for some given support. With each wavelet type of this class, there is a scaling function (called father wavelet), which generates an orthogonal multiresolution analysis.

According to Daubechies (1992), for each integer *r*, the orthonormal basis for $L^2(\mathbb{R})$ is defined as in Equation 4:

$$\phi_{r,j,k} = 2^{-j/2} \phi_r \left(2^{-j} x - k\right), \qquad j, k \in \mathbb{Z} \qquad \text{(Equation 4)}$$

where the function $\phi_r(x)$ in $L^2(\mathbb{R})$ has the property that $\phi_r(x-k) \mid k \in \mathbb{Z}$ is an orthonormal sequential basis in $L^2(\mathbb{R})$. Here *j* is the scale index, *k* is the translation index, and *r* is the filtering index.

Null moments

The waveforms have some null moments, that is, a function $\psi \in L^2(\mathbb{R})$ has *m* null moments if it satisfies Equation 5:

$$\int x^l \psi(x)\, dx = 0 \qquad \text{(Equation 5)}$$

where $l = 0, \ldots, m - 1$

If the wavelets have *m* null moments, then all the coefficients of the wavelet of any polynomial of degree *m* or less can be exactly zero (Nason, 2008).

Null moments and smoothness are mathematically related. The greater the number of null moments of a wavelet, the smoother it is. Moreover, the greater the smoothness of the wavelet, the greater is the probability of perfect reconstruction of the signal decomposed by the wavelet transform.

Clustering analysis

When two samples are close, they should also have similar values for the measured variables; therefore, the greater proximity between the measures related to the samples, the greater the similarity between them. The dendrogram (which has a tree structure) hierarchized this similarity so that one can have a two-dimensional view of the similarity of the whole set of samples used in the study, that is, the dendrogram organizes certain factors and variables (Abonyi and Feil, 2007).

The average linkage clustering was developed as an antidote to the extremes of both single and complete linkage. Although there are some variants of the method, each essentially computes an average of the similarity of a case under consideration with all cases in the existing cluster and, subsequently, joins the case to that cluster if a given level of similarity is achieved using this average value. The most commonly used variant of average linkage computes the arithmetic average of similarities among the cases. Other variants of average linkage are designed to calculate the similarity between the centroids of two clusters that might be merged. The average linkage has been used extensively in the biological sciences but has only recently begun to see much use in the social sciences (Aldenderfer and Blashfield, 1984).

The Mahalanobis distance increases with increasing distances between the two groups and with decreasing within-group variation. By also employing within-group correlations, the Mahalanobis distance takes account of the (possibly nonspherical) shape of the groups (Everitt et al., 2011).

The use of Mahalanobis D^2 according to Equation 6 implies that the investigator is willing to assume that the covariance matrices are at least approximately the same in the two groups. When this is not so, D^2 is an inappropriate inter-group measure, and for such cases, several alternatives have been proposed.

$$D^2 = \left(\overline{x}_A - \overline{x}_B\right)' W^{-1} \left(\overline{x}_A - \overline{x}_B\right) \quad \text{(Equation 6)}$$

where $\overline{x}_A'$ means vector of the group A, $\overline{x}_B'$ means vector of the group B, and W is the pooled within-group covariance matrix for the two groups.

GC-content

The GC-content is an important parameter of bacterial genomes used to scan the basic composition of the genome as well as to understand the evolution of the coded sequence. Historically, this rate is represented in a range of 25 to 75% in bacterial genomes (Mann and Chen, 2010). In the mammalian genome, approximately 50% of all genes are controlled by promoters with high GC-contents. Chang et al. (2015) examined a method for stable quantification of such GC-rich DNA sequences.

For each genome sequence, the GC-content is calculated as the ratio of the sum of bases G, C, under the sum of the bases A, G, C, and T, according to Equation 7:

$$GCcontent = \frac{nG + nC}{nA + nG + nC + nT} \quad \text{(Equation 7)}$$

where nA, nG, nC, and nT represent the number of nucleotide bases A, G, C, and T, respectively, in a sequence. The GC-content can also be calculated for a part of the sequence using the window technique, wherein the GC-content is calculated for a fixed length of a specific window of the sequence. The determination of GC-content ratio helps in identifying gene-rich regions of the genome (Saini and Dewan, 2016). These gene-rich regions bring significant biological information about the genome. Cheng et al. (2016) and Wei et al. (2016) worked with high GC-content, aiming the development of new molecular markers, highlighting the importance of working with gene-rich regions.

RESULTS AND DISCUSSION

Table 1 shows the information for each sequence obtained at the site of the NCBI. In the second column, the identification of each accession number is shown. The third column shows the characteristic of each strain according to Ilina et al. (2013), i.e., DS, DR, MDR, and XDR. In the fourth column, we have the total rate of GC-content, whose values are very close. The last column shows the total size of each sequence; note that all sequences have the length around 4 million.

Table 1. Description of 10 sequences of the *Mycobacterium tuberculosis* genome.

Sequence number	NCBI accession number	Resistance type	Total rate of GC-content	Total sequence length
Seq1	CP002992.1	DS	0.6560	4398525
Seq2	CP000717.1	DS	0.6562	4424435
Seq3	CP001641.1	DS	0.6561	4398812
Seq4	CP001642.1	DR	0.6559	4405981
Seq5	CP001664.1	DR	0.6563	4408224
Seq6	CP001658.1	MDR	0.6561	4398250
Seq7	CP001976.1	XDR	0.6561	4399120
Seq8	CP002884.1	DS	0.6561	4414325
Seq9	AL123456.3	DS	0.6561	4411532
Seq10	CP000611.1	DS	0.6561	4419977

DS = drug susceptible; DR = drug resistant; MDR = multidrug resistant; XDR = extensively drug resistant.

Initially, we show the dissimilarity between the rate of GC-content sequences considering the mean of all sequences and how much each sequence differs from the general average. The result is shown in Figure 2. The sequences that presented averages smaller than the general mean were: 1 (DS), 4 (DR), 9 (DS), and 10 (DS). Moreover, those with averages higher than the general average were: 2 (DS), 3 (DS), 5 (DR), 6 (MDR), 7 (XDR).

Figure 3 shows the decomposition on 5 levels of details of the sequence 1, using the NDWT. The wavelet used is Daubechies with 4 null moments.

The signal was decomposed up to the fifth level, because from the sixth level onwards the energy is quite low, as can be seen in Figure 4, that is, there is a decay of the energy along of the levels. It is also noted that energy concentrates more on the first two levels.

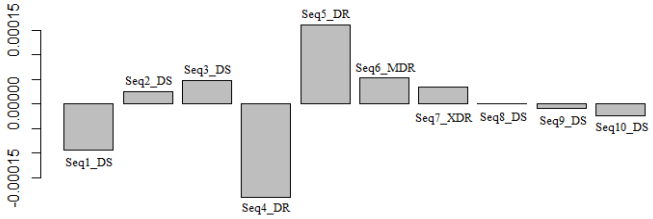


Figure 2. Dissimilarity of the sequence signals. DS = drug susceptible; DR = drug resistant; MDR = multidrug resistant; XDR = extensively drug resistant.

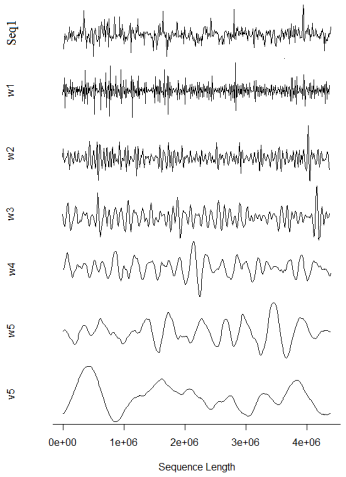


Figure 3. Decomposition of the sequence 1 in 5 levels using the discrete non-decimated wavelet transforms (NDWT). The first signal corresponds to the original signal, and in the sequence we have, w1 is the 1st level of decomposition, w2 the second level of decomposition, and so on. The last signal v5 represents the approximation coefficients of the smoother level.

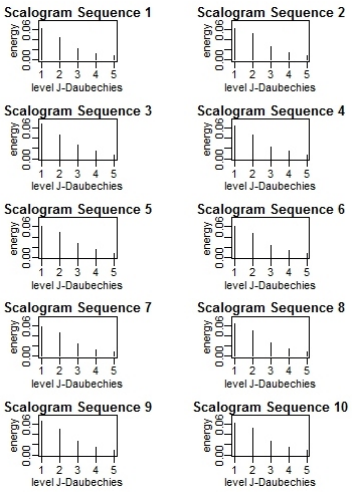


Figure 4. Scalogram for each GC-content sequence.

The approximation coefficients (v5) for all sequences are plotted (Figure 5), and it is possible to note that the regions of higher and lower peaks are coincident. For the windows of the 2 million nucleotides, the sequences 6 (MDR) and 7 (XDR) presented peaks in distinct regions from the other sequences. The highest peaks are regions rich in genes.

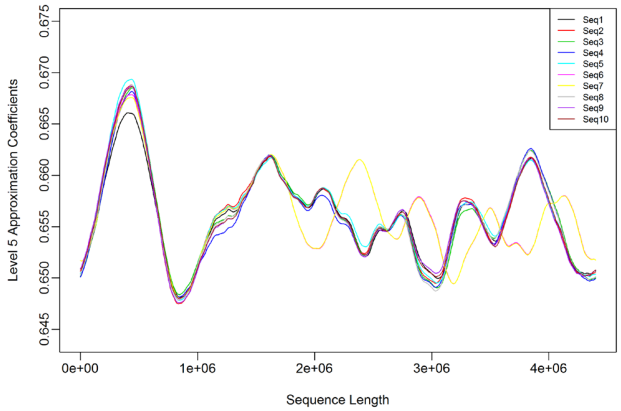


Figure 5. Approximation coefficients (v5) for all sequences.

The energies obtained in each level were considered to the clustering analysis. We use the Mahalanobis distance in a hierarchical method using the average linkage. The formation of 3 groups was verified (Figure 6).

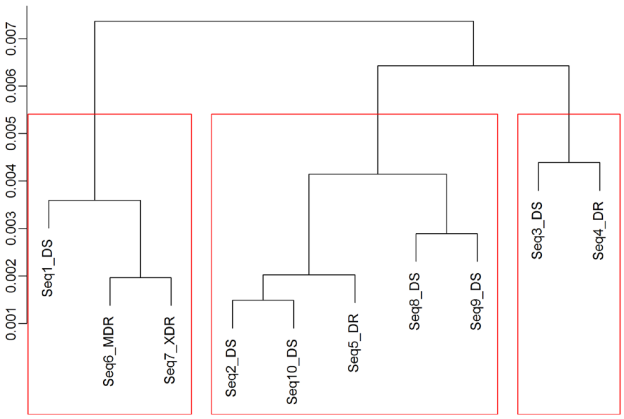


Figure 6. Clustering of sequences based on the energies of each level. DS - drug susceptible, DR - drug resistant, MDR - multidrug resistant, XDR - extensively drug resistant.

The first group (shown on the left side of Figure 6) has the sequences: 1 (DS), 6 (MDR), and 7 (XDR). The second group has the sequences: 2 (DS), 10 (DS), 5 (DR), 8 (DS), and 9 (DS), and the third group has the sequences: 3 (DS) and 4 (DR).

Figures 7 to 9 show each group with their features. The first group (Figure 7) show that the sequences 6 (MDR) and 7 (XDR) almost overlap completely. These strains correspond to a single patient in KwaZulu-Natal, South Africa. However, the sequence 1 (DS) presents different peaks. As a characteristic, this strain was isolated in Russia belonging to the AI family (according to RFLP genotyping), and it is sensitive to all common drugs used in the treatment of tuberculosis.

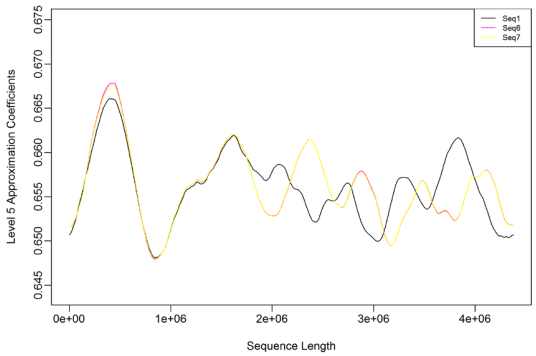


Figure 7. Clustering of sequences 1 (DS), 6 (MDR), and 7 (XDR).

Figure 8 shows the second group, with the sequences 2 (DS), 5 (DR), 8 (DS), 9 (DS), and 10 (DS) that present very similar behavior. The sequence 2 (DS) is a susceptible strain representing the largest portion of patients' tuberculosis isolates recovered during an epidemic in the Western Cape of South Africa. The sequence 5 (DR) corresponds to a drug-resistant strain, having an accelerated rate of transmission between humans under agglomeration conditions. The sequence 8 (DS) is a susceptible strain used for comparative genomic studies. The sequence 9 is a susceptible strain derived from the original human lung H37, isolated in 1934. It has been widely used all over the world in biomedical research. Unlike some clinical isolates, it retains total virulence in animals with tuberculosis and is susceptible to drugs and receptive to genetic manipulation. Moreover, the sequence 10 (DS) is an avirulent susceptible strain derived from its virulent parent strain H37 (isolated from a 19-year-old male patient with chronic pulmonary tuberculosis named Edward R. Baldwin in 1905). This strain was obtained through an aging and dissociation process of an *in vitro* culture in the year 1935.

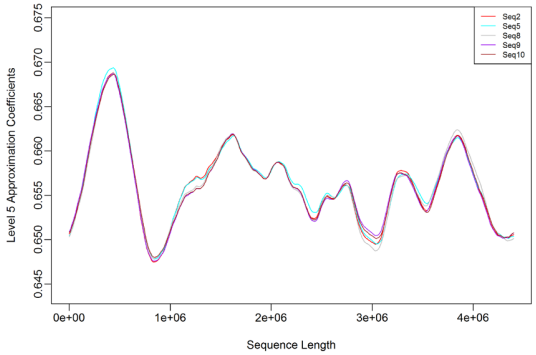


Figure 8. Clustering of sequences 2 (DS), 5 (DR), 8 (DS), 9 (DS), and 10 (DS).

Figure 9 shows the third group with the sequences 3 (DS) and 4 (DR). The sequence 3 (DS) is a susceptible strain belonging to the Beijing family, sequenced for comparative genomic studies. The sequence 4 (DR) is a resistant strain isolated in 2004, referring to a patient with secondary pulmonary tuberculosis, sequenced for comparative genomic studies.

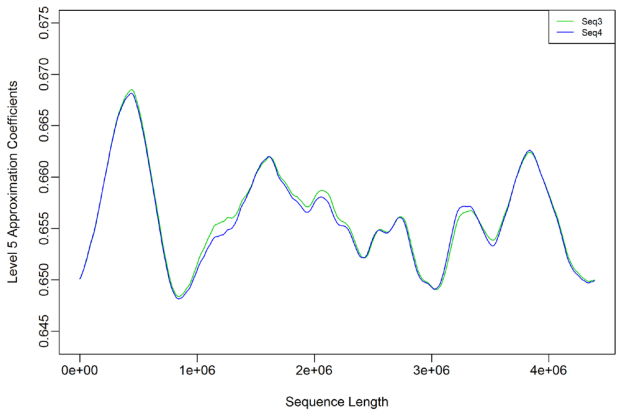


Figure 9. Clustering of sequences 3 (DS) and 4 (DR).

It is interesting to note that although the graphs 8 and 9 are very similar, the proposed methodology detected differences between the analyzed sequences.

Saini and Dewan (2016) based on the calculation of the energy of wavelet decomposition coefficients of complete genomic sequences showed that the genomic sequences of MTB could be grouped only into two groups. The first group with DS and DR sequences (lower energy) and the second group with MDR and XDR sequences (highest energy).

Our method, considering the energy at each level of detail, was able to identify more than two groups. It was possible to detect particularities of sequences 1 (DS), 3 (DS), and 4 (DR) with the proposed methodology.

CONCLUSIONS

The use of the discrete NDWT in the analysis of MTB genome strains allowed us to consider the entire genome sequence without the need to have a power of 2.

Similarities between genome sequences are best detected if one considers the energy at each level of detail of the wavelet decomposition.

ACKNOWLEDGMENTS

L.M. Ferreira thanks CAPES (Aperfeiçoamento de Pessoal de Nível Superior) for the financial support.

REFERENCES

Abonyi J and Feil B (2007). Cluster analysis for data mining and system identification. Birkhäuser, Basel, Boston, Berlin.

Anastassiou D (2000). Frequency-domain analysis of biomolecular sequences. *Bioinformatics* 16: 1073-1081. https://doi.org/10.1093/bioinformatics/16.12.1073

Aldenderfer MS and Blashfield RK (1984). Cluster analysis sage university papers series. Quantitative applications in the social sciences. Sage Publications, Inc., Beverly Hills.

Bao JP and Yuan RY (2015). A wavelet-based feature vector model for DNA clustering. *Genet. Mol. Res.* 14: 19163-19172. https://doi.org/10.4238/2015.December.29.26

Boehme CC, Nabeta P, Hillemann D, Nicol MP, et al. (2010). Rapid molecular detection of tuberculosis and rifampin resistance. *N. Engl. J. Med.* 363: 1005-1015. https://doi.org/10.1056/NEJMoa0907847

Chang GJ, Seyfert HM and Shen XZ (2015). Adaption of SYBR Green-based reagent kit for real-time PCR quantitation of GC-rich DNA. *Genet. Mol. Res.* 14: 8509-8515. https://doi.org/10.4238/2015.July.28.20

Cheng JL, Li J, Qiu YM, Wei CL, et al. (2016). Development of novel SCAR markers for genetic characterization of *Lonicera japonica* from high GC-RAMP-PCR and DNA cloning. *Genet. Mol. Res.* 15: gmr7737. https://doi.org/10.4238/gmr.15027737

Everitt BS, Landau S, Leese M and Stahs D (2011). Cluster analysis. John Wiley & Sons, King's College London, UK.

Daubechies I (1992). Ten Lectures on Wavelets. Society for industrial and applied mathematics, Philadelphia.

Dodin G, Vandergheynst P, Levoir P, Cordier C, et al. (2000). Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. *J. Theor. Biol.* 206: 323-326. https://doi.org/10.1006/jtbi.2000.2127

Gençay R, Selçuk F and Whicher B (2002). An introduction to wavelets and other filtering methods in finance and economics. Academic Press, San Diego, California.

Hernandez E and Weiss G (1996). A first course on wavelets. CRC Press LLC, USA.

Ilina EN, Shitikov EA, Ikryannikova LN, Alekseev DG, et al. (2013). Comparative genomic analysis of *Mycobacterium tuberculosis* drug resistant strains from Russia. *PLoS One* 8: e56577. https://doi.org/10.1371/journal.pone.0056577

Mallat SG (1989). Multiresolution approximations and wavelet orthonormal bases of L2(R). *Trans. Am. Math. Soc.* 315: 69-87.

Mann S and Chen YPP (2010). Bacterial genomic G+C composition-eliciting environmental adaptation. *Genomics* 95: 7-15. https://doi.org/10.1016/j.ygeno.2009.09.002

Morlet J, Arens G, Fourgeau E and Giard D (1982). Wave propagation and sampling theory- Part I: Complex signal and scattering in multilayered media. *Geophysics* 47: 203-221. https://doi.org/10.1190/1.1441328

Nason GP (2008). Wavelet methods in statistics with R. Springer, New York.

National Center for Biotechnology Information (2017). *Mycobacterium tuberculosis*. Genoma. Available at [https://www.ncbi.nlm.nih.gov/assembly/GCF_000224435.1/]. Accessed April 2, 2017.

Ning J, Moore CN and Nelson JC (2003). Preliminary wavelet analysis of genomic sequences. In: Proceedings of the IEEE computer society conference on bioinformatics CSB '03, Stanford, California, 509-510.

Percival DB and Walden AT (2000). Wavelet methods for time series analysis. Cambridge University Press.

Perdigão J, Macedo R, Malaquias A, Ferreira A, et al. (2010). Genetic analysis of extensively drug-resistant *Mycobacterium tuberculosis* strains in Lisbon, Portugal. *J. Antimicrob. Chemother.* 65: 224-227. https://doi.org/10.1093/jac/dkp452

R Core Team (2017). A Language and environment for statistical computing. Vienna, Austria. Available at [https://www.R-project.org/].

Saini S and Dewan L (2016). Application of discrete wavelet transform for analysis of genomic sequences of *Mycobacterium tuberculosis*. *Springerplus* 5: 64. https://doi.org/10.1186/s40064-016-1668-9

Vannucci M and Liò P (2001). Non-decimated wavelet analysis of biological sequences: applications to protein structure and genomics. *The Indian J. Statistics* 63: 218-233.

Wei CL, Cheng JL, Khan MA, Yang LQ, et al. (2016). An improved DNA marker technique for genetic characterization using RAMP-PCR with high-GC primers. *Genet. Mol. Res.* 15: gmr8721. https://doi.org/10.4238/gmr.15038721

Wojtaszczyk P (1997). A mathematical introduction to wavelets. Cambridge University Press, New York.

ARTIGO 2

Wavelet-domain Elastic net for clustering on genomes strains

Redigido conforme as normas da revista Genetics and Molecular Biology (versão publicada)

Genetics and Molecular Biology, 41, 4, 884-892 (2018)

DOI: http://dx.doi.org/10.1590/1678-4685-GMB-2018-0035

Research Article

Wavelet-domain elastic net for clustering on genomes strains

Leila Maria Ferreira[1], Thelma Sáfadi[2] and Juliano Lino Ferreira[3]

[1]*Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, Departamento de Estatística, Universidade Federal de Lavras (UFLA), Lavras, MG, Brazil.*
[2]*Departamento de Estatística, Universidade Federal de Lavras (UFLA), Lavras, MG, Brazil.*
[3]*Empresa Brasileira de Pesquisa Agropecuária (Embrapa) Pecuária Sul. Bagé, RS, Brazil.*

Abstract

We propose to evaluate genome similarity by combining discrete non-decimated wavelet transform (NDWT) and elastic net. The wavelets represent a signal with levels of detail, that is, hidden components are detected by means of the decomposition of this signal, where each level provides a different characteristic. The main feature of the elastic net is the grouping of correlated variables where the number of predictors is greater than the number of observations. The combination of these two methodologies applied in the clustering analysis of the *Mycobacterium tuberculosis* genome strains proved very effective, being able to identify clusters at each level of decomposition.

Keywords: Elastic net, genome, GC-content, cluster analysis, wavelet transform.

Received: February 01, 2018; Accepted: March 11, 2018.

Introduction

Mycobacterium tuberculosis (MTB), also called Koch bacillus, is a species of the pathogenic bacterium of the genus *Mycobacterium* and the causative agent of most cases of tuberculosis (TB) (Taylor *et al.*, 2003). TB is the ninth leading cause of death worldwide and the leading cause of death from a single infectious agent, ranking above HIV/AIDS. In 2016, there were an estimated 1.3 million TB deaths among HIV-negative people (down from 1.7 million in 2000) and an additional 374,000 deaths among HIV-positive people. An estimated 10.4 million people fell ill with TB in 2016: 90% were adults, 65% were male, 10% were people living with HIV (74% in Africa), and 56% were in five countries: India, Indonesia, China, the Philippines, and Pakistan. Drug-resistant TB is a continuing threat. In 2016, there were 600,000 new cases with TB resistance to rifampicin, the most effective first-line drug, of which 490,000 had multidrug-resistant TB (MDR). Almost half (47%) of these cases were in India, China, and the Russian Federation (WHO, 2017).

By the end of 2016, extensively drug-resistant TB (XDR) had been reported by 123 WHO member states. Of these, 91 countries and five territories reported representative data from continuous surveillance or surveys regarding the proportion of MDR cases that had XDR. Combining their data, the average proportion of MDR cases with XDR was 6.2% (95% CI: 3.6–9.5%) (WHO, 2017)

Several studies are underway regarding the study of MTB strains in order to detect the mutations that the strains suffer when becoming drug resistant. Boehme *et al.* (2010) studied an automated molecular test for MTB resistance to rifampin (RIF) in patients with suspected drug-sensitive or multidrug-resistant pulmonary tuberculosis. Perdigão *et al.* (2010) performed a study to characterize the genetic changes associated with the high number of XDR TB that threatens the control of TB worldwide. Zhou *et al.* (2017) investigated the association between genotype and drug resistance profiles of MTB strains circulating in China in a national drug resistance survey. Müller *et al.* (2013) studied programmatically selected multidrug-resistant strains drive and the emergence of extensively drug-resistant TB in South Africa. Smith *et al.* (2014) investigated the reduced virulence of an XDR outbreak strain of MTB in a murine model. Other related studies were done by Iwamoto *et al.* (2012), Sandegren *et al.* (2011), Buu *et al.* (2012), and Treviño *et al.* (2015).

In order to extract information regarding the genome of MTB, one of the techniques used is the GC-content. This is an important parameter of bacterial genomes used to scan the basic composition of the genome, as well as to understand the evolution of the coded sequence. Hildebrand *et al.* (2010) showed that the GC-content is highly correlated to genomic GC-content, that is, there is selection on genomic base composition in many bacteria.

Send correspondence to Leila Maria Ferreira. Departamento de Estatística, Universidade Federal de Lavras (UFLA), P.O. Box 3037, 37200-000 Lavras, MG, Brazil. E-mail: leilamaria2003@yahoo.com.br.

The use of wavelet analysis in genomic data has been growing a growing field. One of the features of this analysis is the extraction of characteristics that are hidden, thus increasing the precision of the results. Conceptually, the wavelet transform is a technique for seeing and representing a signal. This signal is decomposed in resolution levels, where each level adds details. Mathematically, it is represented by a function oscillating in time or space. The method has sliding windows that expand or compress to capture low and high frequency signals, respectively (Percival and Walden, 2000). Its origin occurred in the field of seismic study to describe the disturbances arising from a seismic impulse (Morlet *et al.*, 1982).

Among the wavelet techniques, we used the discrete non-decimated wavelet transform (NDWT), which has as its main characteristic that it can work with any size of signals/sequences. In this technique, the coefficients are translation invariants, that is, the choice of origin is irrelevant, since all the observations are used in the analysis, a situation that does not occur in the discrete decimated wavelet transform (DWT). Discrete wavelet transforms have been used to identify gene locations in genomic sequences (Ning *et al.*, 2003), identifying long-range correlations, locating periodicities in DNA sequences (Vannucci and Liò, 2001), and for analysis of G+C patterns (Dodin *et al.*, 2000).

The NDWT method can be used in any genome type, increasing the speed of the analysis, which is processed almost in real time. Bao and Yuan (2015) created a wavelet-based feature vector (WFV) model that outperformed the other models in terms of both the clustering results and the running time, confirming that wavelets are an efficient method in the analysis of DNA sequences.

The clustering analysis that has been worked with genomic data is the elastic net, which is the regular regression method that linearly combines the L_1 and L_2 penalties of the LASSO and Ridge regression methods. The main feature of this method is the grouping of correlated variables where the number of predictors is greater than the number of observations. Elastic net employs a grouping effect, in which strongly correlated predictors tend to be in or out of the model. Sáfadi (2017) showed that the wavelet-domain elastic net methodology was effective for clustering of time series data, that is, the interaction of wavelets with elastic net is an efficient method of grouping. Another characteristic of the method is the speed with which the analyses are processed. Mol *et al.* (2009) proved that there exists a particular "elastic net representation" of the regression function such that, if the number of data increases, the elastic net estimator is consistent not only for prediction but also for variable/feature selection, demonstrating the adaptive capacity of the elastic net. Cho *et al.* (2009) proposed a simple stepwise procedure that identifies disease-causing SNPs simultaneously by employing elastic net regularization, a variable selection method that allows addressing multicollinearity in the study of rheumatoid arthritis, showing the efficiency of genetic data interaction with elastic net. The studies of Waldmann *et al.* (2013), Hughey and Butte (2015), Ayers and Cordell (2010), Ogutu *et al.* (2012), and Furqan and Siyal (2016) also show the significant relationship between genetic data interaction and elastic net.

In this work, the discrete non-decimated wavelet transform was applied to GC-content sequences; the detailed level coefficients are used to study similarities of MTB genome strains through elastic net methodology. The visualization of the graphs obtained with the elastic net allowed identifying the groupings of similar strains. The proposed methodology was applied to ten MTB sequences, with two being 2 drug-resistant, 6 six drug-susceptible, one multi drug-resistant and one extensively drug-resistant.

Material and Methods

In the analyses, the free software R (R Core Team, 2017) was used.

Table 1 shows the description of each strain of the MTB genome, obtained from the National Center for Biotechnology Information (NCBI, 2017). The methodology used was as follows:

1. The GC-content of all the sequences was evaluated using a sliding window of 10,000 base pairs (bp).

The GC-content is an important parameter of bacterial genomes used to scan the basic composition of the genome as well as to understand the evolution of the coded sequence. Generally, the CG-content ranges from 25 to 75% in bacterial genomes (Mann and Chen, 2010). In the mammalian genome, approximately 50% of all genes are controlled by promoters with high GC-content. Chang *et al.* (2015) examined a method for stable quantification of such GC-rich DNA sequences.

For each genome sequence, the GC-content is calculated as the ratio of the sum of G and C bases divided by the sum of the A, G, C and T bases (Equation 1):

$$GCcontent = \frac{nG + nC}{nA + nG + nC + nT} \quad (1)$$

where nA, nG, nC, and nT are the number of A, G, C and T nucleotide bases, respectively, in a sequence. The GC-content can also be calculated for a part of the sequence using the window technique, wherein the GC-content is calculated for a fixed length of a specific window of the sequence. The determination of GC-content ratio helps in identifying gene-rich regions of the genome (Saini and Dewan, 2016). Theses gene-rich regions provide significant biological information about the genome. Cheng *et al.* (2016) and Wei *et al.* (2016) worked with high GC-content aiming to develop new molecular markers, highlighting the importance of working with gene-rich regions.

2. The sequences were decomposed using a discrete non-decimated wavelet transform. We used the Daubechies wavelet (4 null moments) with 5 levels of decomposition.

Table 1 - Descriptions of the *Mycobacterium tuberculosis* strains.

Sequences	Descriptions of the strains
Seq1_DS	Strain was isolated in Russia belonging to the AI family (according to RFLP genotyping) and it is sensitive to all common drugs used in the treatment of tuberculosis.
Seq2_DS	Susceptible strain representing the largest portion of tuberculosis isolates recovered during an epidemic in the Western Cape of South Africa.
Seq3_DS	Susceptible strain belonging to the Beijing family, sequenced for comparative genomic studies.
Seq4_DR	Resistant strain isolated in 2004, referring to a patient with secondary pulmonary tuberculosis, sequenced for comparative genomic studies.
Seq5_DR	Drug-resistant strain, having an accelerated rate of transmission between humans under agglomeration conditions.
Seq6_MDR	Strain from a single patient in KwaZulu-Natal, South Africa.
Seq7_XDR	Strain from a single patient in KwaZulu-Natal, South Africa.
Seq8_DS	Susceptible strain used for comparative genomic studies.
Seq9_DS	Susceptible strain derived from the original human lung H37, isolated in 1934. It has been widely used all over the world in biomedical research. Unlike some clinical isolates, it retains total virulence in animals with tuberculosis and is susceptible to drugs and receptive to genetic manipulation.
Seq10_DS	A virulent susceptible strain derived from its virulent parent strain H37 (isolated from a 19-year-old male patient with chronic pulmonary tuberculosis, named Edward R. Baldwin in 1905). This strain was obtained through an aging and dissociation process of an *in vitro* culture in 1935.

A wavelet function is the interpretation of a short wave with rapid increase and decrease. The theory is based on the representation of functions in different scales and resolutions (time-scale), being considered one of its main characteristics (Daubechies, 1992).

In the analysis of wavelets, the oscillating window is called the mother wavelet. There are arbitrary translations and dilations, and thus the mother wavelet generates other wavelets (Hernadez and Weiss, 1996).

By definition: a wavelet is a function $\psi(x) \in L^2(\mathbb{R})$, such that the function family is given by Equation 2:

$$\psi_{j,k} = 2^{\frac{-j}{2}} \psi(2^{-j}x - k), \quad (2)$$

where j and k are arbitrary integers on an orthonormal basis in Hilbert space $L^2(\mathbb{R})$ (Wojtaszczyk, 1997).

The characteristic of the discrete non-decimated wavelet transform (NDWT) is to keep the same amount of data in the even and odd decimations on each scale and continue to do the so on each subsequent scale. The coefficients are translational invariants, that is, the circular displacement of the data is reflected in the same direction of the coefficients. Another feature is the ability to handle data of arbitrary size that does not require the sample size to be a power of two, which is what occurs in the discrete decimated wavelet transform (Nason, 2008). The main advantage of this method is associated with zero-pass filters, which means that it operates circularly to the data allowing functionalities at different scales to be aligned with the sequence of the original data (Vannucci and Liò, 2001).

Percival and Walden (2000) highlight that the NDWT method can also be used to form a multiresolution analysis (MRA). The approximation coefficients and coefficients of details of MRA are such that circularly shifting the time-series by any amount will circularly shit each approximation coefficients and coefficients of details by a corresponding amount. The NDWT method is computed using $O(N \log_2 N)$ multiplications.

The Daubechies wavelet is a family of orthogonal wavelets that define a discrete wavelet transformation, characterized by a maximum number of null moments (degree of smoothing) for some given support. With each wavelet type of this class, there is a scaling function (called father wavelet), which generates an orthogonal multiresolution analysis.

According to Daubechies (1992), for each integer r, the orthonormal basis for $L^2(\mathbb{R})$ is defined by Equation 3:

$$\psi_{r,j,k} = 2^{\frac{-j}{2}} \psi_r(2^{-j}x - k), \quad j,k \in \mathbb{Z} \quad (3)$$

in which the function $\psi_r(x)$ in $L^2(\mathbb{R})$ has the property that $\psi_r(x-k)|k \in \mathbb{Z}$ is an orthonormal sequential basis in $L^2(\mathbb{R})$. Here, j is the scale index, k is the translation index, and r is the filtering index.

3. The elastic net methodology was used at each level of decomposition, aiming at the identification of similar sequences.

According to Zou and Hastie (2005), given a set of data with n observations and p predictors, considering $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, the response $\mathbf{X} = (\mathbf{x_1}|\ldots|\mathbf{x_p})$ the matrix model, where in $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, p$, are the predictors, and considering that the response is centralized and the predictors are standardized, that is, correlated,

$$\sum_{i=1}^{n} y_i = 0 \sum_{i=1}^{n} x_{ij} = 0 \text{ and } \sum_{i=1}^{n} x_{ij}^2 = 1 \quad \text{for } j = 1, \ldots, \text{p.} \quad (4)$$

For any λ_1 e λ_2 fixed and non-negative, the elastic net criterion is defined as:

$$L(\lambda_1, \lambda_2, \beta) = |\mathbf{y} - \mathbf{X}\beta|^2 \lambda_2 |\beta|^2 + \lambda_1 |\beta|^2, \quad (5)$$

wherein

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2$$

$$|\beta|_1 = \sum_{j=1}^{p} |\beta_j|.$$

The elastic net estimator $\hat{\beta}$ is the minimizer of Equation 5

$$\hat{\beta} = \arg\min_{\beta} \{L(\lambda_1, \lambda_2, \beta\}. \quad (6)$$

This procedure can be seen as a penalized least squares method. Let $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$, so solving $\hat{\beta}$ in equation 5 is equivalent to the optimized problem

$$\hat{\beta} = \arg\min_{\beta} |\mathbf{y} - \mathbf{X}\beta|^2, \text{subject to } (1-\alpha)|\beta|_1 + \alpha|\beta|^2 \leq t \text{ for some } t. \quad (7)$$

The function $(1-\alpha)|\beta|_1 + \alpha|\beta|^2$ is called elastic net penalty and is a convex combination between the penalties that define the LASSO and Rigde estimation, respectively. When $\alpha = 1$ the elastic net becomes a simple Ridge regression. When $\alpha = 0$, we have the LASSO penalty, which is convex but not strictly convex. When $\alpha = 0.5$ we have elastic net penalty. These arguments can be seen in Figure 1.

The Ridge regression estimator (keeping all predictors) is:

$$\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{n} (Y_i - X\beta)^2 + \lambda \sum_j \beta_j^2 \right\}. \quad (8)$$

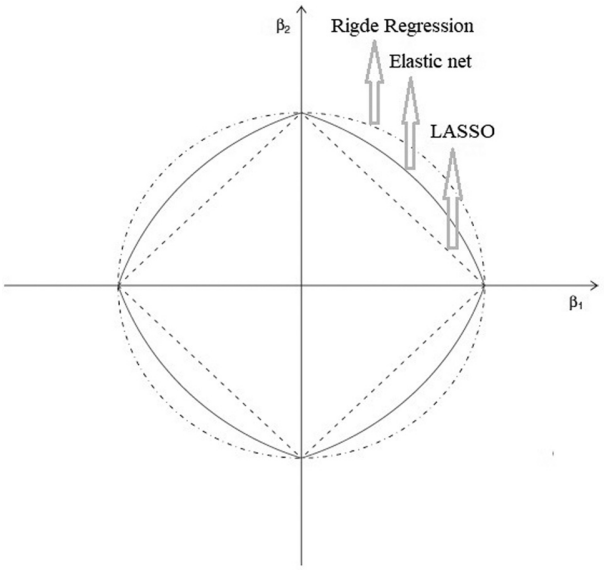


Figure 1 - Geometry of the penalties. Source: Zou and Hastie (2005).

The LASSO estimator (keeping the most significant predictors and removing the others) is:

$$\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{n} (Y_i - X\beta)^2 + \mu \sum_j |\beta_j| \right\}. \quad (9)$$

The elastic net is a combination of Ridge regression and LASSO; its estimator is given by

$$\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{n} (Y_i - X\beta)^2 + \lambda \left((1-\alpha) \sum_j |\beta_j| + \alpha \sum_j \beta_j^2 \right) \right\}. \quad (10)$$

Results and Discussion

Table 2 contains the information for each sequence of strains of the MTB genome obtained from NCBI. Note that the GC-content total rate values are very close, indicating that there are no differences between the sequences.

Figure 2 shows the size and signal behavior visualization of each GC-content sequence. Note that the sequences show practically the same behavior. The *x*-axis shows the amount of nucleotides of each sequence.

We applied the proposed methodology to the GC-content sequences. After comparing locality and smoothness of the decomposing wavelet, the Daubechies with 4 vanishing moments, db4, to the non-decimated wavelet decomposition was selected. The coefficients at each multiresolution level are denoted by d1, d2, d3, d4, d5, and s5 with d1 being the level of the finest detail and s5 the smoothest level.

First, the elastic net was applied on the GC-content sequence and on the smooth level coefficients (s5, Figure 3), which revealed three groups (Figure 3a). The first group was composed of three members, the second of five, whereas the third one contained only Seq2_DS. Considering the

Table 2 - Description of the *Mycobacterium tuberculosis* genome.

Sequence number	NCBI Access number	Resistance type	Total Rate of GC-content	Infraspecific name
Seq1	CP002992.1	DS	0.6560	CTRI-2
Seq2	CP000717.1	DS	0.6562	F11
Seq3	CP001641.1	DS	0.6561	CCDC5079
Seq4	CP001642.1	DR	0.6559	CCDC5180
Seq5	CP001664.1	DR	0.6563	str. Haarlem
Seq6	CP001658.1	MDR	0.6561	KZN 1435
Seq7	CP001976.1	XDR	0.6561	KZN 605
Seq8	CP002884.1	DS	0.6561	CCDC5079
Seq9	AL123456.3	DS	0.6561	H37Rv
Seq10	CP000611.1	DS	0.6561	H37Ra

DS = drug susceptible; DR = drug resistant; MDR = multidrug resistant; XDR = extensively drug resistant.

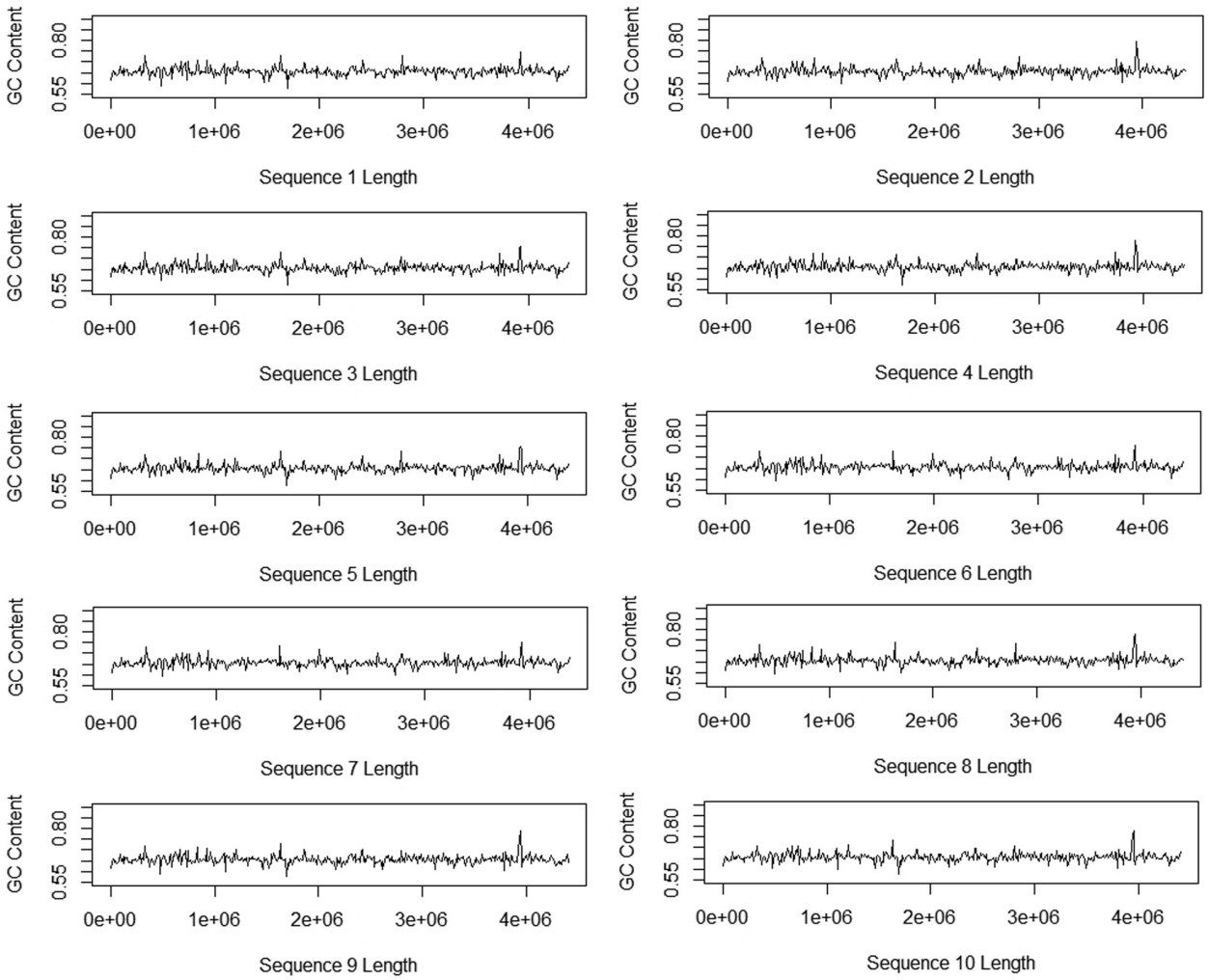


Figure 2 - GC-content sequence sign (10,000 bp window) of MTB strains.

smooth level coefficients, (Figure 3b) the first group is made up by the sequences Seq6_MDR and Seq7_XDR and the others are in the second group.

Comparing the formation of the groups without NDWT with the groups formed with NDWT, referring to the smooth level of decomposition, we found that in the second group formed without NDWT the elastic net failed to distinguish XDR and MDR strains from DS strains. This is a contradictory situation, since belong completely different strains. In the NDWT, referring to the smooth level of decomposition, this separation occurs very clearly, showing that the XDR and MDR strains are different in relation the other strains analyzed and between them are similar.

Considering each level of detail, the elastic net was applied to d1 to d5 coefficients. Figure 4 shows the elastic net plots on each level. We summarized the clustering observed in Table 3.

Table 3 shows that the 6-MDR and 7-XDR sequences were pooled at all levels of detail. These strains correspond to a single patient in KwaZulu-Natal, South Africa. At level 1, the highlight is for the 1-DS sequence that alone forms a group; this strain was isolated in Russia from the AI family (according to RFLP genotyping), and was sensitive to all common drugs used in the treatment of TB. For levels 2 and 3, the sequence 2-DS formed a group; this is a susceptible strain representing the largest portion of TB isolates from patients recovered during an epidemic in the Western Cape region of South Africa. Level 2 also highlights the 4-DR sequence, which is a resistant strain isolated in 2004, referring to a patient with secondary pulmonary TB, sequenced for comparative genomic studies.

The 5-DR sequence corresponds to a drug-resistant strain, with an accelerated rate of transmission between humans under agglomeration conditions. The 8-DS sequence is a susceptible strain used for comparative genomic studies. The 9-DS sequence is a susceptible strain derived from the original human lung H37, isolated in 1934. It has been widely used all over the world in biomedical research. Unlike some clinical isolates, it retains total virulence in animals with TB and is susceptible to drugs and receptive to

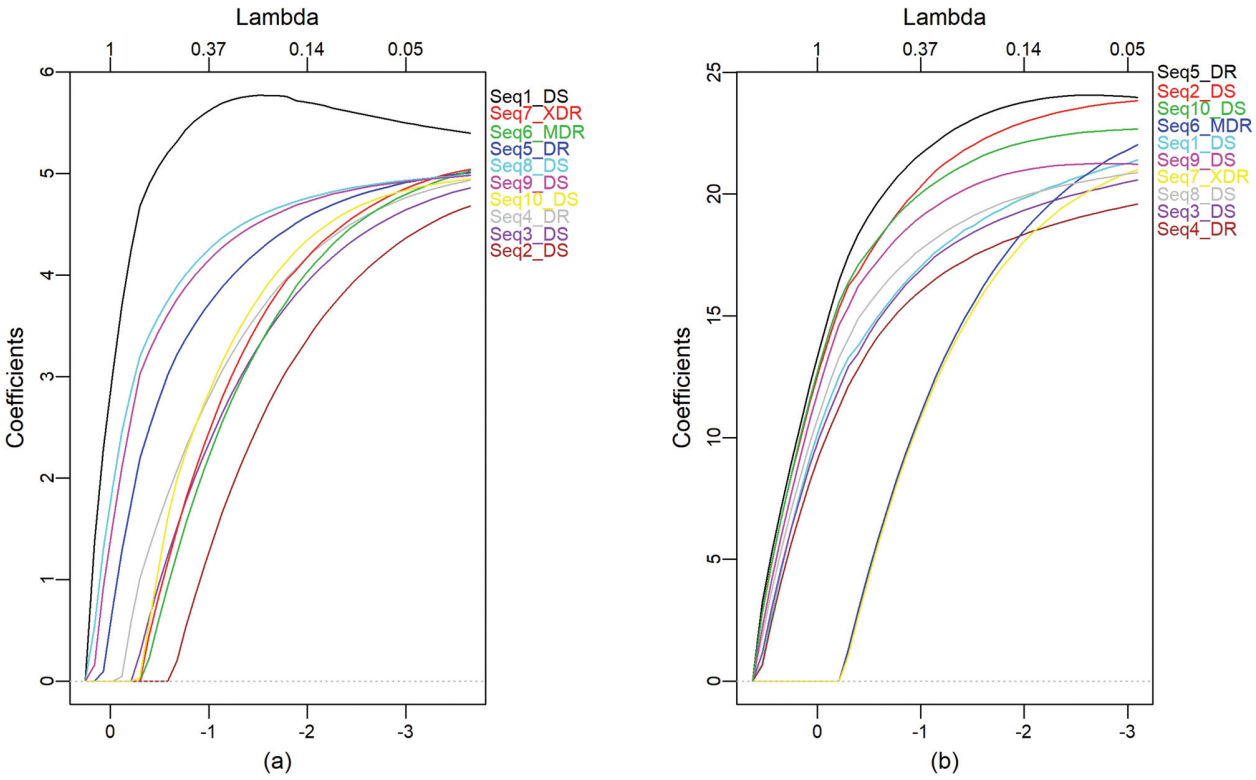


Figure 3 - Elastic net for: (a) signals of the GC-content sequences, (b) s5 coefficients.

Table 3 - Formation of the groups at each level of decomposition.

Levels	Groups				
	1	2	3	4	5
1	DS{1}	DS{2, 3} DR{4, 5}	DS{8, 9, 10}{6-MDR, 7-XDR}		
2	DS{2}	DS{3, 10}	DR{4}	{6-MDR, 7-XDR}	DS{1, 8, 9} DR{5}
3	DS{2}	DS{3, 10} {6-MDR, 7-XDR}	DS{1, 8, 9} DR{4, 5}		
4	{6-MDR, 7-XDR}	DS{1, 2, 3, 8, 9, 10} DR{4, 5}			
5	{6-MDR, 7-XDR}	DS{1, 2, 3, 8, 9, 10} DR{4, 5}			

genetic manipulation. These sequences appear grouped at all levels, except for the first detail level.

In addition, the DS (3 and 10) sequences appear grouped at all levels, except for the first level of detail. The sequence 3-DS is a susceptible strain belonging to a Beijing family, sequenced for comparative genomic studies, and the 10-DS sequence is an avirulent susceptible strain derived from its virulent parent strain H37 (isolated in 1905 from a 19-year-old male patient named Edward R. Baldwin who had chronic pulmonary TB). This strain was obtained in 1935 through an aging and dissociation process of an *in vitro* culture.

Concerning group formation, at levels 4 and 5 these groups were the same, forming two groups. At level 2, the largest number of groups were formed, totaling five. At this level, a larger specification of the groups occurs, with two strains isolated.

Saini and Dewan (2016), based on the calculation of the energy of wavelet decomposition coefficients of complete genomic sequences, showed that the genomic sequences of MTB could be grouped only into two groups. The first group with DS and DR sequences (lower energy) and the second group with MDR and XDR sequences (highest energy). Ferreira *el al*. (2017), considering the energy at each level of detail, were able to identify more than two groups, as particularities of 1 (DS), 3 (DS), and 4 (DR) sequences were detected with the proposed methodology.

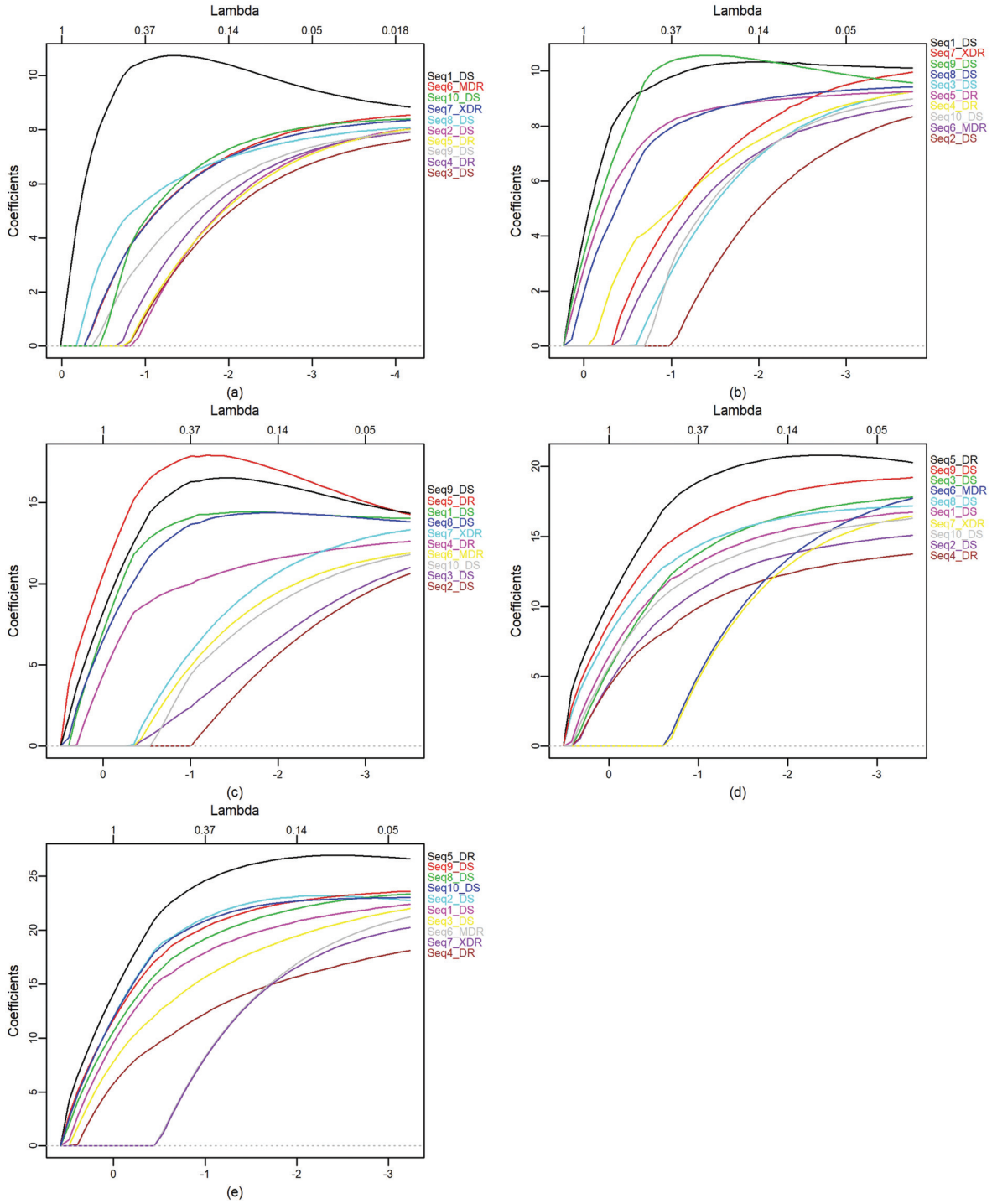


Figure 4 - Elastic net for: (a) d1, (b) d2, (c) d3, (d) d4, and (e) d5 coefficients.

Conclusions

The combination of the NDWT and elastic net methodologies, applied in the analysis of clustering of the *Mycobacterium tuberculosis* genome strains, proved very effective. Through this analysis, it was possible to see group formation at each level of decomposition.

Acknowledgments

The authors would like to thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for the financial support.

References

Ayers KL and Cordell HJ (2010) SNP Selection in genome-wide and candidate gene studies via penalized logistic regression. Genet Epidemiol 34:879-891.

Bao JP and Yuan RY (2015) A wavelet-based feature vector model for DNA clustering. Genet Mol Res 14:19163-19172.

Boehme CC, Nabeta P, Hillemann D, Nicol MP, Shenai S, Krapp F, Allen J, Tahirli R, Blakemore R, Rustomjee R *et al.* (2010) Rapid molecular detection of tuberculosis and rifampin resistance. N Engl J Med 363:1005-1015.

Buu TN, van Soolingen D, Huyen MNT, Lan NTN, Quy HT, Tiemersma EW, Kremer K, Borgdorff MW and Cobelens FGJ (2012) Increased transmission of*Mycobacterium tuberculosis*Beijing genotype strains associated with resistance to streptomycin: a population-based study. PLoS One 7:e42323.

Chang GJ, Seyferty HM and Sen XZ (2015) Adaption of SYBR green-based reagent kit for real-time PCR quantitation of GC-rich DNA. Genet Mol Res 14:8509-8515.

Cheng JL, Qiu YM, Wei CL, Yang LQ and Fu JJ (2016) Development of novel SCAR markers for genetic characterization of *Lonicera japonica* from high GC-RAMP-PCR and DNA cloning. Genet Mol Res 15:gmr7737.

Cho S, Kim H, Oh S, Kim K and Park T (2009) Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis.BMC Proc 3 Suppl 7:S25.

Daubechies I (1992) Ten Lectures onWavelets. Society for Industrial and Applied Mathematics, Philadelphia, 378 p.

Dodin G, Vandergheynst P, Levoir P, Cordier C and Marcourt L (2000) Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. J Theor Biol 206:323-326.

Ferreira LM, Sáfadi T and Lima RR (2017) Evaluation of genome similarities using the non-decimated wavelet transform. Genet Mol Res 16:gmr16039758.

Furqan MS and Siyal MY (2016) Elastic-net copula Granger causality for inference of biological networks. PLoS One 11:e0165612.

Hernandez E and Weiss G (1996) A first course on wavelets. CRC Press, Boca Raton, 489 p.

Hildebrand F, Meyer A and Walker AE (2010) Evidence of selection upon genomic GC-content in bacteria. PLoS Genetics 6:e1001107.

Hughey JJ and Butte AJ (2015) Robust meta-analysis of gene expression using the elastic net. Nucleic Acids Research 43:1-11.

Iwamoto T, Grandjean L, Arikawa K, Nakanishi N, Caviedes L, Coronel J, Sheen P, Wada T, Taype CA, Shaw MA *et al.* (2012) Genetic diversity and transmission characteristics of Beijing family strains of *Mycobacterium tuberculosis* in Peru. PLoS One 7:e49651.

Mann S and Chen YPP (2010) Bacterial genomic G+C composition-eliciting environmental adaptation. Genomics 95:7-15.

Mol C, Vito E and Rosasco L (2009) Elastic-net regularization in learning theory. J Complex 25:201-230.

Morlet J, Arens G, Fourgeau E and Giard D (1982) Wave propagation and sampling theory- Part I: Complex signal and scattering in multilayered media. Geophysics 47:203-221.

Müller B, Chihota VN, Pillay M, Klopper M, Streicher EM, Coetzee G, Trollip A, Hayes C, Bosman ME, Pittius NCGv *et al.* (2013) Programmatically selected multidrug-resistant strains drive the emergence of extensively drug-resistant tuberculosis in South Africa. PLoS One 8:e70919.

Nason GP (2008) Wavelet methods in statistics with R. Springer, New York, 268 p.

Ning J, Moore CN and Nelson JC (2003) Preliminary wavelet analysis of genomic sequences. In: Proceedings of the IEEE computer society conference on bioinformatics. IEEE Computer Society, Stanford, pp 509-510.

Ogutu JO, Schulz-Streeck T and Piepho HP (2012) Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions.BMC Proc 6 Suppl 2:S10.

Percival DB and Walden AT (2000) Wavelet methods for time series analysis. Cambridge University Press, Cambridge, 611 p.

Perdigão J, Macedo R, Malaquias A, Ferreira A, Brum L and Portugal I (2010) Genetic analysis of extensively drug-resistant *Mycobacterium tuberculosis* strains in Lisbon, Portugal. J Antimicrob Chemother 65:224-227.

Sáfadi T (2017) Wavelet-domain elastic net for clustering of volatilities. Int J Stat Econ 18:73-80.

Saini S and Dewan L (2016) Application of discrete wavelet transform for analysis of genomic sequences of *Mycobacterium tuberculosis*. SpringerPlus 5:64.

Sandegren L, Groenheit R, Koivula T, Ghebremichael S, Advani A, Castro E, Pennhag A, Hoffner S, Mazurek J, Pawlowski A *et al.* (2011) Genomic stability over 9 years of an isoniazid resistant *Mycobacterium tuberculosis* outbreak strain in Sweden. PLoS One 6:e16647.

Smith KLJ, Saini D, Bardarov S, Larsen M, Frothingham R, Gandhi NR, Jacobs Jr WR, Sturm AW and Lee S (2014) Reduced virulence of an extensively drug-resistant outbreak strain of *Mycobacterium tuberculosis* in a murine model. PLoS One 9:e94953.

Taylor GM, Stewart GR, Cooke M, Chaplin S, Ladva S, Kirkup J, Palmer S and Young DB (2003) Koch's Bacillus – a look at the first isolate of *Mycobacterium tuberculosis* from a modern perspective. Microbiology 149:3213–3220.

Treviño SF, Otero RM, Noriega ER, Díaz EG, Gómez HRP, García VB, Cabrera LV and González EG (2015) Genetic diversity of *Mycobacterium tuberculosis* from Guadalajara, Mexico and identification of a rare multidrug resistant Beijing Genotype. PLoS One 10:e0118095.

Vannucci M and Liò P (2001) Non-decimated wavelet analysis of biological sequences: Applications to protein structure and genomics. Sankhya: Indian J Stat 63:218-233.

Waldmann P, Mészáros G, Gredler B, Fuerst C and Sölkner J (2013) Evaluation of the lasso and the elastic net in genome-wide association studies.Front Genet 4:270.

Wei CL, Cheng JL, Khan MA, Yang LQ, Imani S, Chen HC and Fu JJ (2016) An improved DNA marker technique for genetic characterization using RAMP-PCR with high-GC primers. Genet Mol Res 15:gmr8721.

Wojtaszczyk P (1997) A Mathematical Introduction to Wavelets. Cambridge University Press, New York, 274 p.

Zhou Y, Hof Svd, Wang S, Pang Y, Zhao B, Xia H, Anthony R, Ou X, Li Q, Zheng Y *et al.* (2017) Association between genotype and drug resistance profiles of *Mycobacterium tuberculosis* strains circulating in China in a national drug resistance survey. PLoS One 12:e0174197.

Zou H and Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B 67:301-320.

Internet Resources

NCBI (2017) *Mycobacterium tuberculosis*. Genome, https://www.ncbi.nlm.nih.gov/assembly/GCF_000224435.1/ (accessed 2 July 2017).

R Core Team (2017) R - a language and environment for statistical computing, https://www.R-project.org/ (accessed 10 June 2017).

WHO (2017) Global tuberculosis report 2017, http://www.who.int/tb/publications/global_report/en/ (accessed 5 June 2017).

Associate Editor: Ana Tereza R. Vasconcelos

ARTIGO 3

Evaluation of genome similarities: a wavelet-domain approach

Redigido conforme as normas da revista Genetics and Molecular Biology (versão sob revisão)

Evaluation of genome similarities: a wavelet-domain approach

Leila Maria Ferreira[1], Thelma Sáfadi[2] and Juliano Lino Ferreira[3]

[1]Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, Departamento de Estatística, Universidade Federal de Lavras (UFLA), Lavras, MG, Brazil.
[2]Departamento de Estatística, Universidade Federal de Lavras (UFLA), Lavras, MG, Brazil.
[3]Empresa Brasileira de Pesquisa Agropecuária (Embrapa) Pecuária Sul. Bagé, RS, Brazil.

Short title: Evaluation of genome similarities

Send correspondence to Thelma Sáfadi. Departamento de Estatística, Universidade Federal de Lavras (UFLA), P.O. Box 3037, 37200-000 Lavras, MG, Brazil. E-mail: safadi@ufla.br

Abstract

Several statistical methods have been used in the analysis of genomic similarities. We propose a new methodology that associates the decomposition of the genome in several levels of details (wavelet transform) and the Hurst exponent. The wavelets are intrinsically related to the multiresolution analysis, allowing the study to be done at various levels of resolution, having a variation of the thicker scales to the finer scales. As strong point, the wavelets are adaptive to the varying lengths of coding and non-coding regions in genomic sequences and does not require any training. The Hurst exponent is used as a measure of long memory of processes. The advantage of this method is the ability to quantify the degree of correlation between observations. The values of the Hurst exponent at each level of detail are used to obtain clusters thus allowing visualization of the similarity between genomes. The analysis the Hurst exponent is obtained throw five different methods and the methodology is applied to ten sequences of the genomes of *Mycobacterium tuberculosis*. The results were efficient, where the Aggregated Variance method presented the best result with respect to the formation of the groups of similar strains.

Keywords: *Mycobacterium tuberculosis*; GC content; Non-decimated wavelet transform; Hurst exponent.

Introduction

Mycobacterium is a genus comprising a number of Gram-positive, acid-fast, rod-shaped aerobic bacteria and is the only member of the family *Mycobacteriaceae* within the order *Actinomycetales*. Like other closely related *Actinomycetales*, such as *Nocardia* and *Corynebacterium*, *mycobacteria* have unusually high genomic DNA GC content and are capable of producing mycolic acids as major components of their cell wall. *Mycobacterium tuberculosis* (MTB) is the causative agent of tuberculosis (TB), a chronic infectious disease with a growing incidence worldwide. This species is responsible for more morbidity in humans than any other bacterial disease. It infects 1.7 billion people a year (≈33% of the entire world population) and causes over 3 million deaths/year. This bacterium does not form a polysaccharide capsule, and is an extremely slow growing obligate aerobe. The sluggish growth rate is a result of the tough cell wall that resists the passage of nutrients into the cell and inhibits waste products to be excreted out of the cell. The specialized cell envelope of this organism resembles a modified Gram positive cell wall (NCBI, 2018).

Concerned about the growing numbers of people dying with tuberculosis, studies are being targeted on combating drug resistant strains. Since the launch of the Global Project on Anti-tuberculosis Drug Resistance Surveillance in 1994, data on drug resistance have been systematically collected and analyzed from 160 countries worldwide (82% of the 194 WHO Member States), which collectively have more than

97% of the world's population and TB cases. This includes 90 countries that have continuous surveillance systems based on routine diagnostic drug susceptibility testing (DST) of all TB patients and 70 countries that rely on epidemiological surveys of representative samples of patients. Surveys conducted about every 5 years represent the most common approach to investigating the burden of drug resistance in resource-limited settings. Among the drug resistant strains the most worrying are Multidrug Resistant (MDR) and Extensively Drug Resistant (XDR) (WHO, 2018).

Recently, the use of wavelets has increasingly been used in the analysis of bacterial genomes, for example: Linehan (2016) studied about wavelet packet analysis of amino acid chain sequences in the proteins of *mesophile* and *thermophile* bacteria. Song et al. (2004) worked in the comparative genomics via wavelet analysis for closely related bacteria. König et al. (2006) studied about discovery functional gene expression patterns in the metabolic network of *Escherichia coli* with wavelets transforms. Sun et al. (2011) used the wavelet analysis to rapidly determine the characteristic morphology of the spore coat of bacteria. Cattani (2012) worked on the existence of wavelet symmetries in *Archaea* DNA.

Considering sequences of the genome of the bacterium *Mycobacterium tuberculosis*, Ferreira et al. (2017) showed that the clustering analysis using the energy (variance) obtained at each decomposition level by means of the discrete non-decimated wavelet transform (NDWT) was essential to verify the similarity of the sequences. Ferreira et al. (2018) worked with the combination of the two methodologies NDWT and Elastic net, applied in the analysis of clustering of the same strains of the *Mycobacterium tuberculosis* genome. In this proposal, through visualization of the graphs obtained with the Elastic net at each decomposition level, it was possible to identify the groups of the similar strains.

One of the most widely used bacterial genome analysis techniques is GC content (Pevsner, 2009). As the genome is composed of nitrogenous bases of a molecule of DNA or RNA, the GC content transforms these bases in percentage that will represent the signal to be analyzed by means of a determined statistic.

Conceptually the wavelet transform is a technique of seeing and represents a signal. This signal is decomposed in resolution levels, where each level brings a detailing, which corresponds with the multiresolution analysis (Wojtaszczyk, 1997). Mathematically, it is represented by a function oscillating in time or space. As characteristic, it has sliding windows that expand or compress to capture low and high frequency signals, respectively (Percival and Walden, 2000). We considered the discrete non-decimated wavelet transform (NDWT), whose main characteristic is that it can work with any size of signals/sequences (Vannucci and Liò, 2001; Nason, 2008).

Studies involving the Hurst exponent were originally developed in hydrology for the practical matter of determining optimum dam sizing for the Nile river's volatile rain and drought conditions that had been observed over a long period of time. The name "Hurst exponent", or "Hurst coefficient", derives from Harold Edwin Hurst (1880–1978), who was the lead researcher in these studies; the use of the standard notation *H*

for the coefficient relates to his name also (Palma, 2007; Beran, 2013). Its applicability in genome analysis of bacteria can be portrayed by the following works: (Liu et al., 2015; Audit and Ouzounis, 2003; Peng et al., 2017; Zhou and Yu, 2014; Liu et al., 2012).

The proposal of this paper is to verify the similar genomes, whose strains belong to the genome of *Mycobacterium tuberculosis*, through of the interaction of two techniques (wavelet transform and Hurst exponent), wherein five methods were tested for the estimation of the Hurst exponent at each level of signal decomposition.

Materials and Methods

The specification of each sequence considered is presented in Table 1.

Table 1 - Descriptions of the *Mycobacterium tuberculosis* strains.

Sequences	Descriptions of the strains
Seq1_DS	Strain was isolated in Russia belonging to the AI family (according to RFLP genotyping) and it is sensitive to all common drugs used in the treatment of tuberculosis.
Seq2_DS	Susceptible strain representing the largest portion of patients' tuberculosis isolates recovered during an epidemic in the Western Cape of South Africa.
Seq3_DS	Susceptible strain belonging to the Beijing family, sequenced for comparative genomic studies.
Seq4_DR	Resistant strain isolated in 2004, referring to a patient with secondary pulmonary tuberculosis, sequenced for comparative genomic studies.
Seq5_DR	Drug resistant strain, having an accelerated rate of transmission between humans under agglomeration conditions.
Seq6_MDR	Strain correspond to a single patient in KwaZulu-Natal, South Africa.
Seq7_XDR	Strain correspond to a single patient in KwaZulu-Natal, South Africa.
Seq8_DS	Susceptible strain used for comparative genomic studies.
Seq9_DS	Susceptible strain derived from the original human lung H37, isolated in 1934. It has been widely used all over the world in biomedical research. Unlike some clinical isolates, it retains total virulence in animals with tuberculosis and is susceptible to drugs and receptive to genetic manipulation.
Seq10_DS	A virulent susceptible strain derived from its virulent parent strain H37 (isolated from a 19 year old male patient with chronic pulmonary tuberculosis, named Edward R. Baldwin in 1905). This strain was obtained through an aging and dissociation process of an in vitro culture in the year 1935.

At the first moment of the analysis it is necessary to obtain the signal referring to the strains of the genome of *Mycobacterium tuberculosis*. The technique used is GC content, using a sliding window of 10,000 base pairs (bp) (Saini and Dewan, 2016).

The GC content is calculated as the ratio of the sum of bases G, C, under the sum of the bases A, G, C and T, according to Equation 1:

$$GCcontent = \frac{nG + nC}{nA + nG + nC + nT} \quad (1)$$

where nA, nG, nC and nT represent the number of nucleotide bases A, G, C and T, respectively, in a sequence.

In the Table 2 we have the description of each one of the 10 analyzed sequences obtained from the National Center for Biotechnology Information (NCBI, 2018), as well as the total rate of GC content.

Table 2 - Description of the *Mycobacterium tuberculosis* genome.

Sequence number	NCBI Access number	Resistance type	Total Rate of GC content	Infraspecific name
Seq1	CP002992.1	DS	0.6560	CTRI-2
Seq2	CP000717.1	DS	0.6562	F11
Seq3	CP001641.1	DS	0.6561	CCDC5079
Seq4	CP001642.1	DR	0.6559	CCDC5180
Seq5	CP001664.1	DR	0.6563	str. Haarlem
Seq6	CP001658.1	MDR	0.6561	KZN 1435
Seq7	CP001976.1	XDR	0.6561	KZN 605
Seq8	CP002884.1	DS	0.6561	CCDC5079
Seq9	AL123456.3	DS	0.6561	H37Rv
Seq10	CP000611.1	DS	0.6561	H37Ra

DS = drug susceptible; DR = drug resistant; MDR = multidrug resistant; XDR = extensively drug resistant.

Having the signals of each of the sequences, we enter into the phase of decomposition of these signals through of the discrete non-decimated wavelet transform (NDWT), whose description follows just below (Kang and Vidakovic, 2016).

Considering that ϕ and ψ are scaling and wavelet functions respectively, we represent a data vector $y = (y_0, y_1, \ldots, y_{m-1})$ of size m as a function f in terms of shifts of the scaling function at some multiresolution level J such that $J - 1 < \log_2 m \leq J$, as

$$f(x) = \sum_{k=0}^{m-1} y_k \phi_{J,k}(x),$$

where $\phi_{J,k}(x) = 2^{J/2}\phi(2^J(x-k))$. The data interpolating function f can be re-expressed of according to Equation 2:

$$f(x) = \sum_{k=0}^{m-1} c_{J_0,k}\phi_{J_0,k}(x) + \sum_{j=J_0}^{J-1}\sum_{k=0}^{2^n-1} d_{jk}2^{j/2}\psi\left(2^j(x-k)\right), \tag{2}$$

where $\phi_{J_0,k}(x) = 2^{J_0/2}\phi\left(2^{J_0}(x-k)\right)$,

$\psi_{jk}(x) = 2^{j/2}\psi\left(2^j(x-k)\right)$,

$j = J_0, \dots, J-1; k = 0, 1, \dots, m-1.$

The coefficients $c_{J_0,k}$ and $d_{jk}, j = J_0, \dots, J-1; k = 0, \dots, m-1$, represent the NDWT of vector *y*.

We considered the Daubechies wavelet with 4 null moments and 5 levels of details, the coefficients of each level are represented by (d1, d2, d3, d4, d5), where d1 corresponds at the level with less details and d5 at the level with more details.

The Hurst exponent belongs to the range $(0,1)$, wherein for $0.5 < H < 1$ it is said that process has long-range dependence, for $H = 0.5$ it is uncorrelated, while for $0 < H < 0.5$ the process has short-range dependence. In simple terms, this means that high values of *H* correspond to processes which highly depend on previous values and have memory (persistent) while low values express the opposite, being anti-trended (anti-persistent) (Feng and Vidakovic, 2017; Šiljak and Şeker, 2014).

For the estimation of the Hurst exponent five methods were used:

Aggregated Variance Method

According Beran (1989) one striking property of long memory processes is that the variance of the sample mean converges to zero slower than the rate N^{-1}, where N is the sample size.

$$Var(\bar{X}_N) \sim cN^{2H-2} \tag{3}$$

for large N, where $c > 0$ and $\bar{X}_N$ is the sample mean. This suggests the following method for estimating H. Divide the series into N/m blocks of size m, compute the sample mean

$$\bar{X}_m(k) = \frac{1}{m}\sum_{t=(k-1)m+1}^{km} X(i), \quad k = 1, 2, \dots, N/m \tag{4}$$

in each block and the sample variance

$$s^2(m) = (N/m-1)^{-1}\sum_{k=1}^{N/m}(\bar{X}_m(k) - \bar{X}_N)^2, \tag{5}$$

where $\bar{X}_N$ denotes the overall mean. Plotting $\log s^2(m)$ versus $\log(m)$ should yield points scattered along a straight line with slope equal to $2H-2$.

This method was also used in the works: (Montanari et al., 2000), (Montanari et al., 1997) and (Caccia et al., 1997).

Differenced Aggregated Variance Method

Teverovsky and Taqqu (2001) proposed a method for detecting long-range dependence even in the presence of nonstationarity. It is a variance-type estimator obtained by taking the logarithm of the first-order difference of Equation (4),

$$\log \Delta Var[\bar{X}_m(k)] \sim \log \frac{d}{dm} Var\bar{X}_m(k) + \log \Delta m \tag{6}$$

On one hand,

$$\frac{d}{dm} Var[\bar{X}_m(k)] \sim (2H-2)Cm^{2H-3} \tag{7}$$

On the other hand, since the m values are logarithmically spaced, we have

$$\Delta \log(m) = const; \text{ that is, } \log \Delta m = \log m + C_1$$

Hence

$$\begin{aligned} \Delta Var[\bar{X}_m(k)] &= (2H-3)\log m + \log(2H-2)\, C + \log m + C_1 \\ &= (2H-2)\log m + C_2 \end{aligned}$$

Thus, in a log-log plot we would expect to see a straight line with a slope equal to $2H-2$.

We can cite the following works that also used this estimation method of H: (Lahmiri, 2016), (Rea et al., 2009), (Abaev et al., 2013), (Jaiswal et al., 2015) and (Ranganai and Kubheka, 2016).

Aggregated Absolute Value Method

Consider the series of the average define in Equation (4), and compute its *n-th* absolute moment (Bărbulescu et al., 2010)

$$AM_n^{(m)} = \frac{1}{(N/m)} \sum_{k=1}^{(N/m)} \left| \bar{X}_k^{(m)} - \bar{X} \right|^n \tag{9}$$

$AM_n^{(m)}$ is asymptotically proportional to $m^{n(H-1)}$.

To find an estimate for H: compute $AM_n^{(m)}$ for different values of m; plot it in a log-log plot against m; the point should be scattered along a line with slope $n(H-1)$.

Other authors also used this method: (Danladi et al., 2017), (Jonkers, 2003), (Owczarczuk, 2012), (Montanari et al., 1999) and (Rehman, 2009).

Peng Method

According to Adler et al. (1998) the method follows the following steps: compute the partial sum within each block of size m

$$Y(k)^m = \sum_{t=(k-1)m+1}^{km} X_t\,, \;\; k = 1, 2, \ldots, (N/m); \tag{10}$$

fit a regression line $y = a + bk$; compute the variance of the residual $s_r^{(m)} = \frac{1}{m}\sum_{k=1}^{N/m}(y(k) - a - bk)^2$; plot $\log s_r^{(m)}$ vs $\log m$; the slope should be $2H$.

Foo (2011) and Pacheco and Román (2006) also worked with this method.

R/S Method

The method R/S Beran (1994) corresponds: consider

$$Y_T = \sum_{t=1}^{T} X_t$$

Define the adjusted range

$$R(t,k) = \max_{0\le i\le k}\left[Y_{t+i} - Y_t - \frac{i}{k}(Y_{t+1} - Y_t)\right] - \min_{0\le i\le k}\left[Y_{t+i} - Y_t - \frac{i}{k}(Y_{t+k} - Y_t)\right] \quad (11)$$

Consider

$$S(t,k) = \sqrt{k^{-1}\sum_{i=t+1}^{t+k}\left(X_i - \bar{X}_{t,k}\right)^2} \quad (12)$$

where $\bar{X}_{t,k} = k^{-1}\sum_{i=t+1}^{t+k} X_i$.

The standardized ratio

$$Q(t,k) = \frac{R(t,k)}{S(t,k)} \quad (13)$$

is known as rescaled adjusted range or R/S - statistic.

For the River Nile data, Hurst (1951) observed that, for large k,

$$\log E(R/S) \approx a + H\log k \quad (14)$$

with $H > \frac{1}{2}$.

Based on Hurst's empirical findings, one can do: divide the series into k block of size N/k; compute the R/S statistics $Q(t_i,k)$, as defined in Equation (13), with starting values $t_i = iN/k + 1$ for all possible k such that $t_i + k < N$; plot its logarithm against the logarithm of k; the estimated slope from the regression will be then the estimate of H.

We can cite the following works that also used this method: (Perez et al., 2009), (Zhao et al., 2015), (Nikolopoulos et al., 2014), (Liu et al., 2015), (Danilenko, 2009) and (Nikolopoulos et al., 2016).

The values of the Hurst exponent obtained at each level of detail are considered in the cluster analysis. The clustering analysis was done in each method using distance of Mahalanobis in a hierarchical method with the average linkage.

All analyzes and figures were performed in the free software R, version 3.4.0 (R Core Team 2018). The packages used were waveslim, fArma and cluster (Whitcher, 2015), (Wuertz, 2013) and (Maechler et al., 2017). The choice of the number of groups formed in each method was made using the package NbClust (Charrad et al., 2014).

Results and discussion

In what follows we will present in detail the analysis for the Aggregate Variance Method.

In the Table 3 we have the values calculated for the Hurst exponent at each decomposition level. Note that in all levels of decomposition the value of the Hurst exponent is less than 0.5 showing short-range dependence.

Figure 1 corresponds to the formation of three groups referring to the Aggregated Variance Method. The first group is formed only by the sequence Seq1_DS, which strain was isolated in Russia belonging to the AI family (according to

RFLP genotyping), and it is sensitive to all common drugs used in the treatment of tuberculosis. The second group shows the similarity of the sequences Seq6_MDR and Seq7_XDR, these sequences correspond to a single patient in KwaZulu-Natal, South Africa.

Table 3 - Hurst Exponent obtained by Aggregated Variance Method.

Sequences	Levels				
	Level 1	Level 2	Level 3	Level 4	Level 5
Seq1_DS	-0.1764	0.0128	-0.1432	-0.0692	0.0723
Seq2_DS	-0.0707	0.0251	-0.2434	0.0858	0.0513
Seq3_DS	-0.1070	0.0170	-0.2113	0.0471	0.0705
Seq4_DR	-0.1303	-0.0515	-0.1438	0.0648	0.0499
Seq5_DR	-0.0362	-0.0401	-0.2597	0.0257	0.0771
Seq6_MDR	-0.0167	0.0233	-0.1412	0.0400	0.1977
Seq7_XDR	-0.0490	0.0176	-0.1443	0.0347	0.1979
Seq8_DS	-0.1711	0.0331	-0.2831	0.0765	0.0595
Seq9_DS	-0.1537	0.0068	-0.4009	0.0759	0.0604
Seq10_DS	-0.2241	-0.0554	-0.1840	0.0714	0.0506

Analyzing the formation of the largest group, we noticed that the similar sequences coincide with the results obtained by Ferreira et al. (2018), because in each level of decomposition it presents a group formation similar to the methodology we are proposing. The sequences that appear in this group are: Seq2_DS and Seq3_DS (whose characteristics are respectively: susceptible strain representing the largest portion of patients' tuberculosis isolates recovered during an epidemic in the Western Cape of South Africa and susceptible strain belonging to the Beijing family, sequenced for comparative genomic studies); Seq5_DR (drug resistant strain, having an accelerated rate of transmission between humans under agglomeration conditions); Seq4_DR and Seq10_DS (whose characteristics are respectively: resistant strain isolated in 2004, referring to a patient with secondary pulmonary tuberculosis, sequenced for comparative genomic studies and a virulent susceptible strain derived from its virulent parent strain H37 which was isolated from a 19 year old male patient with chronic pulmonary tuberculosis, named Edward R. Baldwin in 1905. This strain was obtained through an aging and dissociation process of an in vitro culture in the year 1935) and Seq8_DS and Seq9_DS (whose characteristics are respectively: susceptible strain used for comparative genomic studies and susceptible strain derived from the original human lung H37, isolated in 1934. It has been widely used all over the world in biomedical research. Unlike some clinical isolates, it retains total virulence in animals with tuberculosis and is susceptible to drugs and receptive to genetic manipulation).

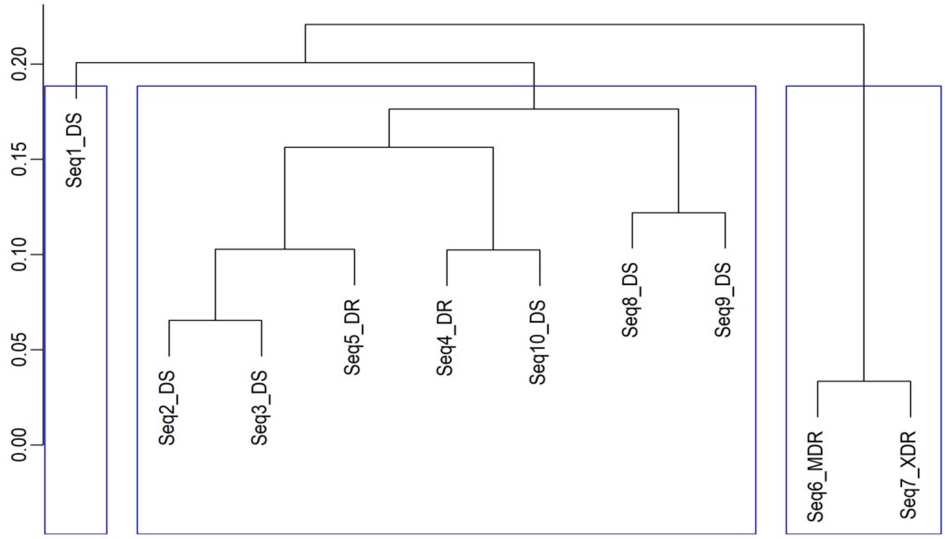


Figure 1 - Clustering of the sequences by means of the Aggregated Variance Method.

Ferreira et al. (2017) found the sequence Seq1_DS in the same group of the sequences Seq6_MDR and Seq7_XDR, however on the plot of the last decomposition level (not showed here) the sequence Seq1_DS presents a different behavior of the sequences Seq6_MDR and Seq7_XDR. This difference was detected here.

The results obtained for the Hurst exponent at each level analyzed in the methods: Differenced Aggregated Variance, Aggregated Absolute Value, Peng and R/S are respectively represented in the Tables 4, 5, 6 and 7.

The figures referring to the formation of groups in the methods: Differenced Aggregated Variance, Aggregated Absolute Value, Peng and R/S are respectively represented in the Figures 2a, 2b, 2c and 2d.

Note that for Aggregated Variance and Aggregated Absolute Value/Moment methods all decomposition levels had H less than 0.5 while R/S, Peng and Differenced Aggregated Variance methods presented H less than 0.5 for the first three levels and the greater than 0.5 for the last two levels showing long-range dependence. The negative values of the Hurst exponent in some methods were found due to the estimated of H to be empirical, being able render H to be negative (Jeon et al., 2014). The values found for the Hurst exponent above 1 in the Peng or Variance of Residuals Method was also found in the works (Jaiswall et al., 2015) and (Rea et al., 2009).

The Aggregated Variance, Differenced Aggregated Variance and Aggregated Absolute Value methods presented the formation of three groups and the Peng and R/S methods presented the formation of two groups. The formation of the groups in each method analyzed is summarized in the Table 8.

The Aggregated Absolute Value method was the one that most approached the formation of groups of the Aggregated Variance method, however the formation of similar sequences in their larger group does not match the results obtained by Ferreira et al. (2017) and Ferreira et al. (2018).

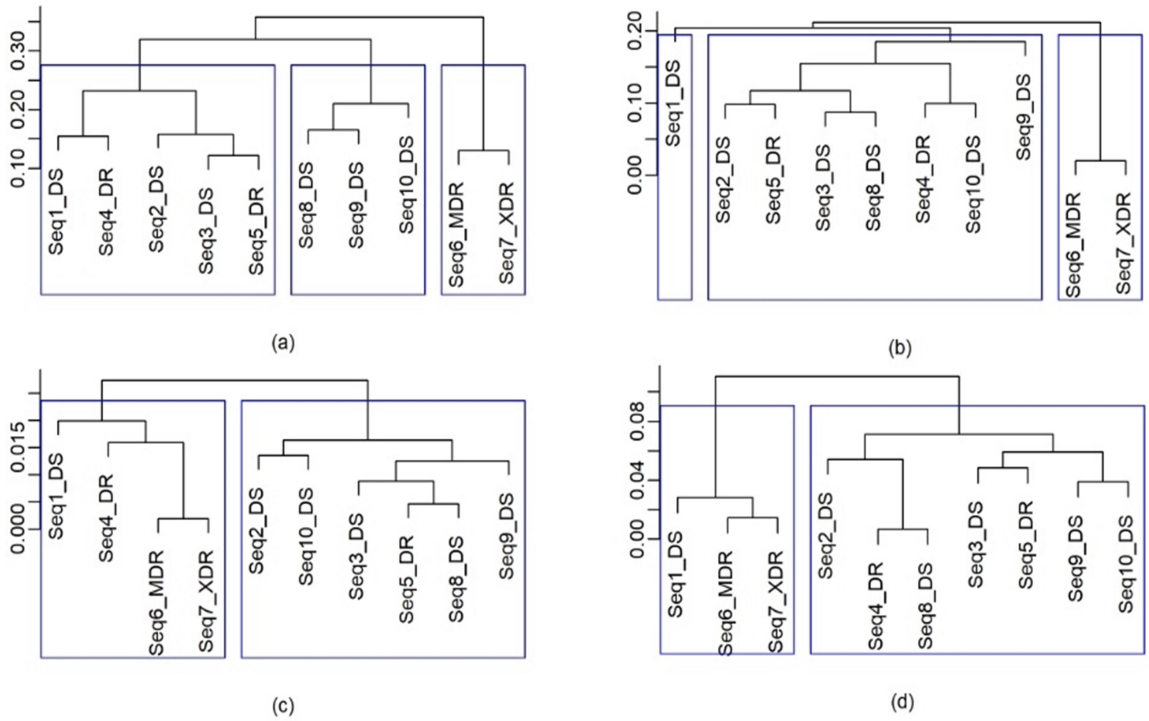


Figure 2a - Clustering of the sequences by means of the Differenced Aggregated Variance Method; **2b** Clustering of the sequences by means of the Aggregated Absolute Value Method; **2c** Clustering of the sequences by means of the Peng Method; **2d** Clustering of the sequences by means of the R/S Method.

The Differenced Aggregated Variance, Peng and R/S methods also do not present coherence in the formation of groups of similar sequences according to Ferreira et al. (2017) and Ferreira et al. (2018).

Table 4 - Hurst Exponents obtained by Differenced Aggregated Variance Method.

Sequences	Levels				
	Level 1	Level 2	Level 3	Level 4	Level 5
Seq1_DS	0.0856	0.2538	0.1420	0.4270	0.8014
Seq2_DS	0.1829	0.3861	0.0152	0.5178	0.7536
Seq3_DS	0.2079	0.2289	0.0406	0.5227	0.7516
Seq4_DR	0.1106	0.1875	0.2269	0.5156	0.7400
Seq5_DR	0.2563	0.2904	0.0615	0.4320	0.7598
Seq6_MDR	0.1837	-0.0118	0.0230	0.4316	0.7683
Seq7_XDR	0.2818	-0.0551	0.0289	0.4454	0.8397
Seq8_DS	0.0153	0.0968	0.0447	0.5061	0.7775
Seq9_DS	-0.0902	0.1226	-0.0757	0.5057	0.8092
Seq10_DS	-0.1168	0.2491	0.0808	0.5249	0.8382

Table 5 - Hurst Exponents obtained by Aggregated Absolute Value Method.

Sequences	Levels				
	Level 1	Level 2	Level 3	Level 4	Level 5
Seq1_DS	-0.0812	0.1258	-0.0419	0.0219	0.1731
Seq2_DS	0.0584	0.1269	-0.1355	0.1846	0.1430
Seq3_DS	-0.0061	0.1291	-0.0976	0.1442	0.1668
Seq4_DR	-0.0064	0.0583	-0.0196	0.1615	0.1403
Seq5_DR	0.0804	0.0738	-0.1317	0.1101	0.1702
Seq6_MDR	0.0781	0.1282	-0.0354	0.1448	0.2949
Seq7_XDR	0.0588	0.1265	-0.0382	0.1398	0.2947
Seq8_DS	-0.0432	0.1368	-0.1687	0.1731	0.1505
Seq9_DS	-0.0592	0.0999	-0.2610	0.1796	0.1560
Seq10_DS	-0.0900	0.0432	-0.0708	0.1713	0.1409

Table 6 - Hurst Exponents obtained by Peng Method.

Sequences	Levels				
	Level 1	Level 2	Level 3	Level 4	Level 5
Seq1_DS	-0.0090	0.0088	0.2287	0.6945	1.1861
Seq2_DS	-0.0164	0.0114	0.2075	0.6817	1.1802
Seq3_DS	-0.0144	-0.0045	0.2109	0.6856	1.1813
Seq4_DR	-0.0124	0.0010	0.2165	0.6849	1.1950
Seq5_DR	-0.0149	0.0006	0.2049	0.6841	1.1862
Seq6_MDR	-0.0124	0.0027	0.2189	0.6986	1.2021
Seq7_XDR	-0.0111	0.0026	0.2202	0.6990	1.2016
Seq8_DS	-0.0121	0.0015	0.2081	0.6839	1.1846
Seq9_DS	-0.0126	-0.0059	0.1983	0.6811	1.1868
Seq10_DS	-0.0128	0.0078	0.1951	0.6813	1.1786

Table 7 - Hurst Exponents obtained by R/S Method.

Sequences	Levels				
	Level 1	Level 2	Level 3	Level 4	Level 5
Seq1_DS	0.2208	0.2773	0.3641	0.6241	0.8847
Seq2_DS	0.1870	0.2087	0.3635	0.6892	0.8398
Seq3_DS	0.1756	0.2646	0.3427	0.6955	0.8417
Seq4_DR	0.1638	0.1951	0.3211	0.7044	0.8407
Seq5_DR	0.2018	0.2610	0.3736	0.6772	0.8609
Seq6_MDR	0.2329	0.2768	0.3811	0.6395	0.8985
Seq7_XDR	0.2234	0.2669	0.3785	0.6362	0.9009
Seq8_DS	0.1620	0.1987	0.3167	0.7042	0.8440
Seq9_DS	0.2055	0.2214	0.3207	0.6935	0.8420
Seq10_DS	0.2258	0.2530	0.3315	0.6953	0.8404

Table 8 - Clustering the sequences for each method analyzed.

Methods	Groups		
	1	2	3
Aggregated Variance	Seq1_DS	Seq2_DS, Seq3_DS, Seq5_DR, Seq4_DR, Seq10_DS, Seq8_DS, Seq9_DS	Seq6_MDR, Seq7_XDR
Differenced Aggregated Variance	Seq1_DS, Seq4_DR, Seq2_DS, Seq3_DS, Seq5_DR	Seq8_DS, Seq9_DS, Seq10_DS	Seq6_MDR, Seq7_XDR
Aggregated Absolute Value	Seq1_DS	Seq2_DS, Seq5_DR, Seq3_DS, Seq8_DS, Seq4_DR, Seq10_DS, Seq9_DS	Seq6_MDR, Seq7_XDR
Peng	Seq1_DS, Seq4_DR, Seq6_MDR, Seq7_XDR	Seq2_DS, Seq10_DS, Seq3_DS, Seq5_DR, Seq8_DS, Seq9_DS	
R/S	Seq1_DS, Seq6_MDR, Seq7_XDR	Seq2_DS, Seq4_DR, Seq8_DS, Seq3_DS, Seq5_DR, Seq9_DS, Seq10_DS	

Among the five methods analyzed for the estimation of the Hurst exponent, the formation of groups of the similar strains of the genome of *Mycobacterium tuberculosis* which more closely approached the results obtained by Ferreira et al. (2017) and Ferreira et al. (2018) was the Aggregated Variance Method. Even though each method presented different group formations, in all methods the sequences Seq6_MDR and Seq7_XDR remained in the same group.

Conclusions

The proposed methodology applied in the analysis of clustering of the strains of genome of the *Mycobacterium tuberculosis* showed relevant results. This methodology can be applied to any type of genome. The use of the discrete non-decimated wavelet transform allows the use of the entire genome sequence without the length condition being the power of two.

The Aggregated Variance method presented the best result with respect to the formation of the groups of similar strains.

Acknowledgments

Leila Maria Ferreira thanks CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for the financial support.

References

Abaev P, Razumchik R and Uglov I (2013) Statistical analysis and modeling of SIP traffic for parameter estimation of server hysteretic overload control. Journal of Telecommucations and Information Technology:22–31.

Adler RJ, Feldman RE and Taqqu MS (1998) A pratical guide to heavy tails: statistical techniques and applications. Birkhäuser, Boston, 534 p.

Audit B and Ouzounis CA (2003) From genes to genomes: universal scale-invariant properties of microbial chromosome organization. Journal of Molecular Biology 332:617–633.

Bărbulescu A, Serban C and Maftei C (2010) Evaluation of Hurst exponent for precipitation time series. Latest Trends on Computers 2:590–595.

Beran J (1989) A test of locationf or data with slowlyd ecayings erialc orrelations. Biometrika 76:261–269.

Beran, J. 1994. Statistics for Long-Memory Processes. Chapman & Hall, 315 p.

Beran J, Feng Y, Ghosh S and Kulik R (2013) Long-memory processes probabilistic properties and statistical methods. Springer-Verlag Berlin Heidelberg, 884 p.

Caccia DC, Percival D, Cannon MJ, Raymond G and Bassingthwaighte JB (1997) Analyzing exact fractal time series: evaluating dispersional analysis and rescaled range methods. Physica A: Statistical Mechanics and its Applications 246:609–632.

Cattani C (2012) On the Existence of Wavelet Symmetries in Archaea DNA. Computational and Mathematical Methods in Medicine 2012:21 p.

Charrad M, Ghazzali N, Boiteau V and Niknafs A (2014) NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software 61:1–36. http://www.jstatsoft.org/v61/i06/

Danilenko S (2009) Long-term memory effect in stock prices analysis. Economics & Management 151-155.

Danladi A, Yohanna M and Silikwa WN (2017) Routing protocols source of self-similarity on a wireless network. Alexandria Engineering Journal.

Feng C and Vidakovic B (2017) Estimation of the Hurst exponent using trimean estimators on nondecimated wavelet coefficients. Journal of Latex Class Files arXiv preprint arXiv:1709.08775.

Ferreira LM, Sáfadi T and Ferreira JL (2018). Wavelet-domain Elastic net for clustering on genomes strains. Genet Mol Biol AHEAD.

Ferreira LM, Sáfadi T and Lima RR (2017) Evaluation of genome similarities using the non-decimated wavelet transform. Genet Mol Res 16:gmr16039758.

Foo DAC (2011) A method of detecting the presence of long run dependencies in time series. Journal of Accounting, Finance and Economics 1:31–47.

Hurst HE (1951) Long-term storage capacity of reservoirs. Transactions of the American Society of Civil Engineers 116:770–799.

Jaiswal R, Lokhande S, Bakre A and Gutte K (2015) Performance analysis of IPv4 and IPv6 internet traffic. ICTACT Journal on Communication Technology 6:1208–1217.

Jeon S, Nicolis O and Vidakovic B (2014) Mammogram diagnostics via 2-D complex wavelet-based self-similarity measures. The São Paulo Journal of Mathematical Sciences 8:265–284.

Jonkers ART (2003) Long-range dependence in the Cenozoic reversal record. Physics of the Earth and Planetary Interiors 135:253–266.

Kang M and Vidakovic B (2016) WavmatND: A Matlab package for non-decimated wavelet transform and its applications. arXiv preprint arXiv:1604.07098.

König R, Schramm G, Oswald M, Seitz H, Sager S, Zapatka M, Reinelt G and Eils R (2006) Discovering functional gene expression patterns in the metabolic network of Escherichia coli with wavelets transforms. BMC Bioinformatics 7:1–14.

Lahmiri S (2016) Clustering of Casablanca stock market based on hurst exponent estimates. Physica A: Statistical Mechanics and its Applications 456:310–318.

Linehan JB (2016) Wavelet Packet Analysis of Amino Acid Chain Sequences in the Proteins of Mesophile and Thermophile Bacteria. DePaul Discoveries 5:1–8.

Liu X, Wang B and Xu L (2015) Statistical analysis of Hurst exponents of essential/nonessential genes in 33 bacterial genomes. PLoS One 10:1–9.

Liu X, Wang YS and Wang J (2012) A statistical feature of Hurst exponents of essential genes in bacterial genomes. Integrative Biology 4:93–98.

Maechler M, Rousseeuw P, Struyf A, Hubert M and Hornik K (2017) cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6.

Montanari A, Taqqu MS and Teverovsky V (1999) Estimating long-range dependence in the presence of periodicity: an empirical study. Mathematical and Computer Modelling 29:217–228.

Montanari A, Rosso R and Taqqu MS (2000) A seasonal fractional ARIMA model applied to the Nile River monthly flows at Aswan. Water Resources Research 36:1249–1259.

Montanari A, Rosso R and Taqqu MS (1997) Fractionally dilferenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation. Water Resources Research 33:1035–1044.

Nason GP (2008) Wavelet methods in statistics with R. Springer, New York, 268 p.

NCBI-National Center for Biotechnology Information. *Mycobacterium tuberculosis*. Genoma, https://www.ncbi.nlm.nih.gov/genome/166 (March 1, 2018)

Nikolopoulos D, Petraki E, Temenos N, Kottou S, Koulougliotis D and Yannakopoulos PH (2014) Hurst exponent analysis of indoor radon profiles of greek apartment dwellings. Physical Chemistry & Biophysics 4:1–8.

Nikolopoulos D, Cantzos D, Petraki E, Yannakopoulos PH and Nomicos C (2016) Traces of long-memory in pre-seismic MHz electromagnetic time series-part 1: investigation through the R/S analysis and time-evolving spectral fractals. Journal of Earth Science & Climatic Change 7:1–17.

Ogden RT (1997) Essential wavelets for statistical applications and data analysis. Boston: Birkhäuser, 206 p.

Owczarczuk M (2012) Long memory in patterns of mobile phone usage. Physica A: Statistical Mechanics and its Applications 391:1428–1433.

Pacheco JCR and Román DT (2006) Performance analysis of time-domain algorithms for self-similar traffic. International Conference on Electronics, Communications and Computers.

Palma W (2007) Long-memory time series theory and methods. John Wiley & Sons, Inc., Canada, 304 p.

Peng C, Lin Y, Luo H and Gao F (2017) A comprehensive overview of online resources to identify and predict bacterial essential genes. Frontiers in Microbiology 8:2331.

Perez CS, Dominguez LR and Pacheco R (2009) What is the required number of users for the generation of aggregated H-ss traffic? International Journal of Computers and Communications 3:1–8.

Percival DB and Walden AT (2000) Wavelet methods for time series analysis. Cambridge University Press, 594 p.

Pevsner J (2009) Bioinformatics and Functional Genomics. Second Edition. John Wiley & Sons, Hoboken, New Jersey, 951 p.

R Core Team. A Language and environment for statistical computing. Vienna, Austria. https://www.R-project.org/ (March 2, 2018)

Rehman S (2009) Study of Saudi Arabian climatic conditions using Hurst exponent and climatic predictability index. Chaos, Solitons & Fractals 39:499–509.

Ranganai E and Kubheka SB (2016) Long memory mean and volatility models of platinum and palladium price return series under heavy tailed distributions. SpringerPlus 5:1–20.

Rea W, Oxley L, Reale M and Brow J (2009) Estimators for long range dependence: an empirical study. Electronic Journal of Statistics 0:1–16.

Šiljak H and Şeker S (2014) Hurst analysis of induction motor vibrations from aging process. Balkan Journal of Electrical & Computer Engineering 2:16–19.

Song J, Ware T, Liu SL and Surette M (2004) Comparative Genomics via Wavelet Analysis for Closely Related Bacteria. EURASIP Journal on Applied Signal Processing 2004:5–12.

Sun W, Romagnoli JA, Palazoglu A and Stroeve P (2011) Characterization of Surface Coats of Bacterial Spores with Atomic Force Microscopy and Wavelets. Industrial & Engineering Chemistry Research 50:2876–2882.

Teverovsky V and Taqqu MS (2001) Testing for long-range dependence in the presence of shifting means or a slowly declining trend using a variance type estimator. Journal of time series analysis 18:279–304.

Vannucci M and Liò P (2001) Non-decimated wavelet analysis of biological sequences: applications to protein structure and genomics. Sankhyā: The Indian Journal of Statistics 63:218–233.

Whitcher B (2015) waveslim: Basic wavelet routines for one-, two- and three-dimensional signal processing. R package version 1.7.5. https://CRAN.R-project.org/package=waveslim

WHO-World Health Organization. Global tuberculosis report 2017. http://www.who.int/tb/publications/global_report/en/ (March 5, 2018)

Wojtaszczyk P (1997) A mathematical introduction to wavelets. Cambridge University Press, New York, 276 p.

Wuertz D (2013) fArma: ARMA Time Series Modelling. R package version 3010.79. https://CRAN.R-project.org/package=fArma

Zhao G, Zhou G and Wang J (2015) Application of R/S method for dynamic analysis of additional strain and fracture warning in shaft lining. Hindawi Publishing Corporation Journal of Sensors 2015:1–7.

Zhou Q and Yu YM (2014) Comparative analysis of bacterial essential and nonessential genes with Hurst exponent based on chaos game representation. Chaos, Solitons & Fractals 69:209–216.

CONSIDERAÇÕES GERAIS

As três técnicas propostas para a análise de agrupamentos de genomas apresentaram resultados muito significativos.

A primeira técnica utilizando a energia em cada nível de decomposição conseguiu captar mais detalhes de informações das sequências do genoma da *Mycobacterium tuberculosis* do que Saini e Dewan (2016), que utilizaram a energia total.

A segunda técnica, constituída pela interação de ondaletas e *Elastic net*, apresenta como uma das suas grandes vantagens a visualização da formação dos grupos em cada nível de decomposição, onde os níveis mais suaves obtiveram formações de grupos similares com a primeira técnica e ao trabalho de Saini e Dewan (2016).

A terceira técnica utilizando a combinação de ondaletas e o expoente de Hurst, foi possível verificar a formação de grupos similares com a primeira e a segunda técnica. Nessa terceira técnica o método que se destacou foi o de variância agregada.

Essas três técnicas representam uma nova forma de análise de agrupamento de genomas, podendo ser aplicadas em diferentes genomas, independente do tamanho desses genomas. As análises são processadas muito rapidamente. Detalhes que poderiam passar despercebidos são detectados, devido ao grau de refinamento que os dados são processados.

Como trabalhos futuros seria interessante um estudo sobre o tamanho ideal da janela a ser utilizada para o cálculo da taxa GC *content*. Testar outros tipos de ondaletas no método não decimado. E comparar os resultados obtidos do *Elastic Net* com o *Lasso*.