DULCÍDIA CARLOS GUEZIMANE ERNESTO

TRANSFORMADA DE WAVELET DISCRETA NÃO DECIMADA PARA O AGRUPAMENTO DE GENOMAS DE VÍRUS DAS FAMÍLIAS CORONAVIRIDAE E PARAMYXOVIRIDAE

LAVRAS – MG

2024

DULCÍDIA CARLOS GUEZIMANE ERNESTO

TRANSFORMADA DE WAVELET DISCRETA NÃO DECIMADA PARA O AGRUPAMENTO DE GENOMAS DE VÍRUS DAS FAMÍLIAS CORONAVIRIDAE E PARAMYXOVIRIDAE

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

Profa. Dra. Thelma Sáfadi
Orientadora

Profa. Dra. Leila Maria Ferreira
Coorientadora

LAVRAS-MG

2024

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

Ernesto, Dulcídia Carlos Guezimane.
Transformada de wavelet discreta não decimada para o agrupamento de genomas de vírus das famílias Coronaviridae e Paramyxoviridae / Dulcídia Carlos Guezimane Ernesto. - 2023.
111 p.

Orientador(a): Profa. Dra Thelma Sáfadi.
Coorientador(a): Profa. Dra. Leila Maria Ferreira.
Tese (doutorado) - Universidade Federal de Lavras, 2023.
Bibliografia.

1. Transformada de wavelet. 2. Regressão penalizada. 3. Expoente de Hurst. I. Sáfadi, Profa. Dra Thelma. II. Ferreira, Profa. Dra. Leila Maria. III. Título.

DULCÍDIA CARLOS GUEZIMANE ERNESTO

TRANSFORMADA DE WAVELET DISCRETA NÃO DECIMADA PARA O AGRUPAMENTO DE GENOMAS DE VÍRUS DAS FAMÍLIAS CORONAVIRIDAE E PARAMYXOVIRIDAE

UNDECIMATED DISCRETE WAVELET TRANSFORM FOR THE CLUSTERING OF VIRUS GENOMES OF THE CORONAVIRIDAE AND PARAMYXOVIRIDAE FAMILIES

APROVADA em 18 de dezembro de 2023.

Profa. Dra. Thelma Sáfadi	UFLA
Prof. Dr. Paulo Henrique Sales Guimarães	UFLA
Profa. Dra. Airlane Pereira Alencar	USP
Profa. Dra. Ana Paula Festucci de Herval	UFLA
Profa. Dra. Leila Maria Ferreira	UFLA

Profa. Dra. Thelma Sáfadi
Orientadora

Profa. Dra. Leila Maria Ferreira
Coorientadora

LAVRAS-MG

2024

Ás minhas filhas Kyanda de Ernesto e Shantel Ernesto:

Filhas são como flores que embelezam o nosso mundo com bênçãos.

Por isso eu dedico este trabalho a elas

Pois foram o meu combustível quando tudo parecia impossível

Foram a minha inspiração quando tudo parecia perdido

E minhas companheiras fieis durante todos 4 anos longe de casa

Todo o meu amor a elas eu dedico

E que Deus seja sempre benevolente e piedoso para elas

AGRADECIMENTOS

Agradeço a Deus pelo dom da vida, por ter sido meu protetor e meu porto seguro durante esta caminhada. O apoio das minhas filhas Kyanda e Shantel, que estiveram ao meu lado todos os dias.

A minha mãe Maria Ilda, a minha avó Leonor da Conceição, por me colocarem nas vossas orações.

As minhas amigas Handina Langa, Bibi da Conceição, Mercês, e Santa Helena, agradeço pela amizade e pelos momentos delicados que me apoiaram e me deram forças durante esta etapa. Ao meu amigo Manuel João Castigo, pelo suporte.

À minha orientadora, Thelma Sáfadi, pela atenção, confiança, e principalmente pelos ensinamentos durante este período tão importante. Aprendi muito, amadureci muito, e sou eternamente grata a si Professora. Que Deus possa abençoa-la poderosamente.

À minha coorientadora, Leila Maria Ferreira, agradeço muito pela dedicação e tempo gasto nos nossos encontros de estudo semanais. Agradeço pela infinita paciência durante este percurso. Que Deus possa te abençoar poderosamente.

Ao programa de Pós-graduação em Estatísticae Experimentação Agropecuária. Aos docentes do programa de pós-graduação em Estatística e Experimentação Agropecuária da Universidade Federal de Lavras, pela dedicação e pelo conhecimento transmitido durante a pós-graduação.

À Universidade Pùngué pelo apoio, e pela dispença para continuar com os estudos.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, pelo apoio financeiro.

Muito obrigada!

Porque ainda que a figueira não floresça, nem haja fruto na vide; ainda que decepcione o produto da oliveira, e os campos não produzam mantimento; ainda que as ovelhas da malhada sejam arrebatadas, e nos currais não haja gado;

Todavia eu me alegrarei no Senhor; exaltarei no Deus da minha salvação

(Habacuque 3:17,18)

RESUMO

Este trabalho teve por objetivo a implementação de duas formas de análise de similaridades de sequenciamentos de duas famílias de vírus, sob o domínio das wavelets. As wavelets são costumeiramente utilizadas quando se trabalha com uma extensa base de dados não estacionária. A técnica da transformada de wavelet trabalha com dados em tempo real, e permite que a série temporal possa ser decomposta em níveis, permitindo deste modo que a cada nível de decomposição seja possível se ampliar o nível de detalhe da série, e assim se observar detalhes omissos, que não se podem observar na série original. Após a decomposição do conteúdo GC de cada um dos sequenciamentos em estudo, duas formas distintas de agrupamento foram implementadas de modo a verificar, os sequenciamentos com algum nível de similaridade. Fez-se a análise de agrupamento com recurso à regressão penalizada no domínio das penalizações lasso, ridge e elastic net, e por outro lado, recorreu-se também ao expoente de Hurst implementado por meio de 5 técnicas diferentes nomeadamente: método peng, análise R/S, variância agregada, variância agregada diferenciada e pelo método dos momentos absolutos. Ao fim do estudo pode-se concluir que variantes mais fracas da família Coronaviridae é que se associam as estirpes da família Paramyxoviridae. E por outro lado, o elastic net (do 1º ao 3º nível), o método dos momentos absolutos e o método de variância agregada diferenciada, tiveram melhor desempenho em relação as demais metodologias.

Palavras-chave: transformada de Wavelet; agrupamentos; regressão penalizada; expoente de Hurst.

ABSTRACT

This work aimed to implement two forms of analysis of sequence similarities of two virus families, under the wavelet domain. Wavelets are commonly used when working with an extensive non-stationary database. The wavelet transform technique works with data in real time, and allows the time series to be decomposed into levels, thus allowing at each level of decomposition to increase the level of detail in the series, and thus observe details omissions, which cannot be observed in the original series. After decomposing the GC content of each of the sequences under study, two different forms of grouping were implemented in order to verify sequences with some level of similarity. Cluster analysis was carried out using penalized regression in the domain of lasso, ridge and elastic net penalties, and on the other hand, we also used the Hurst exponent implemented through 5 different techniques, namely: peng method, R analysis /S, aggregate variance, differentiated aggregate variance and by the method of absolute moments. At the end of the study, it can be concluded that weaker variants of the Coronaviridae family are associated with strains of the Paramyxoviridae family. And on the other hand, the elastic net (from the 1st to the 3rd level), the absolute moments method and the differentiated aggregate variance method performed better in relation to the other methodologies.

Keywords: wavelet transform; similarities; penalized regression; Hurst exponente.

INDICADORES DE IMPACTO

A tese apresenta impactos sociais e tecnológicos na área de saúde pois oferece metodologias distintas para que o pesquisador na área de saúde ou genética possa ter mecanismos de encontrar sequenciamentos genéticos que possam ser semelhantes e deste modo contribuir com a descoberta da cura para uma determinada doença. Portanto, tem-se como finalidade impactar principalmente pesquisadores das áreas de saúde e bem-estar, e ao mesmo tempo impactar direta ou indiretamente a sociedade através da inclusão de metodologias que facilitem o processamento de informação em base de dados muito grandes. O trabalho apresenta uma fundamentação teórica na qual sustenta cada uma das metodologias propostas, juntamente com várias contribuições relacionadas a tecnologias de processamento de dados para uma base de dados muito grande, e as mesmas tecnologias aqui apresentadas podem ser aplicadas a outras famílias de vírus na área de saúde e bem-estar de modo a modernizar a forma de análise de dados. O trabalho visa fornecer métodos que tornam mais simples a busca por técnicas de comparação entre sequenciamentos tornando deste modo mais necessária a relação entre a Estatística e a Genética. Portanto, teve-se como objetivo a implementação de formas distintas de análise de similaridade de sequenciamentos de duas famílias de vírus, sob o domínio das wavelets. Ao fim do estudo, tem-se como resultado um trabalho que fornece metodologias claras e exequíveis para qualquer família de vírus sem restrição, e que pode ser usado por pesquisadores da UFLA e de diversas áreas com pesquisas relacionadas a genética, saúde e bem-estar.

IMPACT INDICATORS

The thesis presents social and technological impacts in the healthcare field by offering distinct methodologies for researchers in health or genetics to have mechanisms for finding genetic sequences that may be similar and thus contribute to the discovery of a cure for a specific disease. Therefore, its purpose is to primarily impact researchers in health and well-being fields and, at the same time, directly or indirectly impact society through the inclusion of methodologies that facilitate information processing in very large databases. The work provides a theoretical foundation that supports each proposed methodology, along with several contributions related to data processing technologies for large databases. The same technologies presented here can be applied to other virus families in the healthcare domain to modernize data analysis methods. The work aims to provide methods that simplify the search for sequence comparison techniques, thereby emphasizing the relationship between Statistics and Genetics. Therefore, the objective was to implement distinct forms of sequence similarity analysis for two virus families under the domain of wavelets. At the end of the study, the result is a work that provides clear and achievable methodologies for any virus family without restriction, which can be used by researchers at UFLA and various fields with research related to genetics, health, and well-being.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

A	Adenina
C	Citosina
T	Timina
G	Guanina
Beta	Human Coronarirus HKU1
Bet1	Betacoronavirus England 1
Bet2	Human coronavirus OC43
Bet4	Bovine coronavirus
Bet5	SARS coronavirus TOR2
Alfa	Coronavirus humano NL63
Alf1	Camel alpha coronavirus
Alf2	Human coronarirus 229E
MERS	Middle East respiratory syndrome-related coronavirus
Del	Deltacoronavirus
Gama1	Coronavirus duck
Gama2	Turkey coronavirus
Influ1	Human parainfluenza virus 1
Influ3	Human parainfluenza virus 3
Influ4	Human parainfluenza virus 4a viral cRNA
Influ5	Parainfluenzavirus 5 strain W3A
Hendra	Hendrahenipavirus
GC	Guanina e Citosina

SUMÁRIO

PRIMEIRA PARTE – UM PANORAMA GERAL

1 INTRODUÇÃO GERAL

Em pesquisas atuais é indispensável a implementação de métodos estatísticos no processamento de dados, de modo a assegurar a credibilidade das mesmas. Diversas metodologias têm sido utilizadas por pesquisadores para o estudo de sequências genéticas, no entanto, as wavelets têm se tornado uma alternativa eficaz e eficiente para o processamento de sequências genéticas em tempo real, para o estudo de similaridades. A transformada de wavelet discreta não decimada de Daubechies, em estudos anteriores mostrou ser uma alternativa poderosa na decomposição do sinal em tempo real.

Ademais, encontrar semelhanças no sequenciamento de algumas espécies de vírus evolutivamente semelhantes pode ser uma alternativa para os cientistas encontrarem a fórmula para uma vacina preventiva ou para o desenvolvimento de uma cura para alguma doença. Assim, a implementação de métodos estatísticos que possam tornar o processo de estudo das semelhanças mais eficiente é bem-vinda, no sentido em que quanto mais rápido e com capacidade para processar milhares de nucleotídeos numa sequência genômica, melhores estimativas serão obtidas. Além disso, se o investigador pretende identificar vírus e bactérias emergentes e trabalhar com uma base de dados genômica muito grande, é essencial implementar uma metodologia de agrupamento que seja viável para sequências genômicas em tempo real.

Neste trabalho recorreu-se a duas formas distintas de formação de clusters, nomeadamente: Regressão penalizada, nas penalizações Lasso, Ridge e elastic net, e a outra forma de agrupamento foi com recurso ao expoente de Hurst, nos métodos: Peng, análise R/S, método dos momentos absolutos, variância agregada, variância agregada diferenciada.

A regressão penalizada tem sido um recurso bastante usado na modelagem de modelos que melhor possam se adequar aos dados em diversas áreas de conhecimento, pois num cenário atual em que o pesquisador está sujeito a trabalhar com um número elevado de variáveis preditoras que possam explicar uma variável resposta existe uma certa necessidade de se escolher as melhores variáveis preditoras. O mesmo se aplica na análise de similaridades entre duas sequências genômicas.

Para Neto et al. (2019), o método Ridge é o método com um estimador de encolhimento, e a vantagem é que permite uma análise gráfica bastante útil do processo de estimação. Por outro lado, o método Lasso é um método que faz seleção de variáveis preditoras, ou seja, o lasso minimiza a soma de quadrados com uma restrição (Tibshirani, Hoefling e Tibshirani, 2011).

Vários foram os pesquisadores que usaram os métodos de regressão penalizada Lasso ou Ridge, dentre eles, Neto et al. (2019), aplicaram o método de regressão Lasso, Ridge e regressão polinomial na previsão de temperatura do fluído e ganho de energia de um sistema solar em operações com nano fluídos, ao fim do estudo a regressão Ridge não apresentou o melhor desempenho pois o Ridge considerou um modelo com todos os preditores e obteve um alto erro de teste e o lasso excluiu alguns preditores e obteve um resultado melhor que o Ridge.

O método elastic net tem sido um recurso usado na formação de grupos para a análise de sequenciamentos semelhantes, pois é uma metodologia que agrega o que tem de melhor nos métodos de regressão Lasso e Ridge. Vários foram os estudos feitos até então sobre o elastic net na área de ciências da saúde, (Cardoso-Berensztejn et al., 2014) aplicaram o método elastic net para estudar aspectos ecocardiográficos em pacientes com polineuropatia amiloidóptica familiar com mutação VAL30MET. Durante o estudo, fez-se uma análise multivariada usando o modelo de regressão elastic net para identificar as variáveis mais relevantes ao estudo.

Ferreira e Lima (2017), desenvolveram uma pesquisa na qual avaliaram a similaridade de genoma do Mycobacterium tuberculosis combinando a transformada discreta não decimada de wavelet e o elastic net, neste estudo observou-se que a formação de grupos das cepas do genoma em estudo mostrou uma performance muito precisa e com capacidade de reagrupamento a cada nível de decomposição.

Li Guigui et al. (2021) exploraram o padrão de medicação das prescrições antigas e modernas da medicina chinesa para o tratamento da úlcera péptica com base em metodologias que utilizam a análise de tendências de Mann Kendall e o expoente de Hurst foram utilizados para a análise de tendências. Por outro lado, no mesmo estudo, a análise wavelet é utilizada como recurso para a análise de periodicidade.

A justificativa para a realização desse estudo é a necessidade de poder-se contribuir com uma metodologia que seja capaz de processar grandes quantidades de base de dado no domínio de tempo e frequência de forma rápida e eficiente, para séries não estacionárias, e que fosse capaz de ser um contributo para a medicina na procura da cura para outras doenças.

O objetivo do trabalho é de encontrar similaridades nos sequenciamentos genéticos dos vírus das famílias Coronaviridae e Paramyxoviridae. E como objetivos específicos, busca-se aplicar a transformada de wavelet discreta não decimada nos sequenciamentos de modo a verificar a distribuição do conteúdo GC, e fazer uma análise de agrupamentos por meio dos métodos lasso e ridge, elastic net, e pelo cálculo do expoente de Hurst implementado em 5 técnicas distintas.

A estrutura da tese é composta por duas partes. Na primeira parte estão apresentados os conteúdos que fundamentam a base teórica de toda a pesquisa e na segunda parte estão apresentados os seguintes ensaios: 1. "*REGRESSÃO PENALIZADA NO ESTUDO DE SIMILARIDADES DE GENOMAS DE VÍRUS DAS FAMÍLIAS CORONAVIRIDAE E PARAMYXOVIRIDAE*", publicado na Revista Contemporânea; 2. "*The use of an elastic net method in the study of genome similarities of Coronaviridae and Paramyxoviridae viruses*", submetido na revista Biometrical Journal; e 3. "*ANALYSIS OF SIMILARITIES IN RESPIRATORY TRACT GENOMES: An approach using the Hurst exponent and the non-decimated discrete wavelet transform*", publicado na revista Cadernos Pedagógicos.

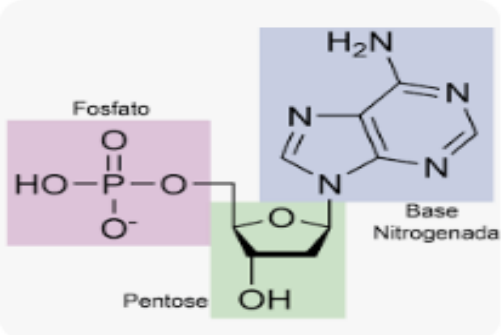
2 REFERENCIAL TEÓRICO

Nesta seção é apresentada a fundamentação teórica do trabalho, abrangendo a organização do genoma viral, o histórico e composição do vírus, o conteúdo GC, escalograma, análise de Fourier, wavelets, transformada de wavelet decimada e não decimada, wavelets de Daubechies, Algoritmo piramidal, análise de agrupamento, 5 formas distintas de estimar o expoente de Hurst, regressão penalizada Lasso, Ridge e Elastic net.

2.1 O GENOMA VIRAL: Organização do genoma viral

De acordo com Zaha, Ferreira e Passaglia (2014), os organismos vivos são compostos por células, nas quais, a propriedade de um organismo depende das suas células individuais cuja continuidade ocorre por meio do seu material genético. As células armazenam suas informações hereditárias na forma de moléculas de DNA (Ácido desoxirribonucleico) de fita dupla, e longas cadeias poliméricas pareadas e não ramificadas, formadas sempre pelos mesmos quatro tipos de monômeros. Esses monômeros são nomeados a partir de um alfabeto de quatro letras A (Adenina), T (Tinina), C (Citosina) e G (Guanina), e estão ligados um ao outro em uma longa sequência linear que codifica a informação genética, assim como as sequências de 1s e 0s que codificam as informações em um arquivo de computador (Alberts et al., 2017). A seguir na Figura 1 apresenta-se a estrutura dos nucleotídeos.

Figura 1- Estrutura dos nucleotídeos



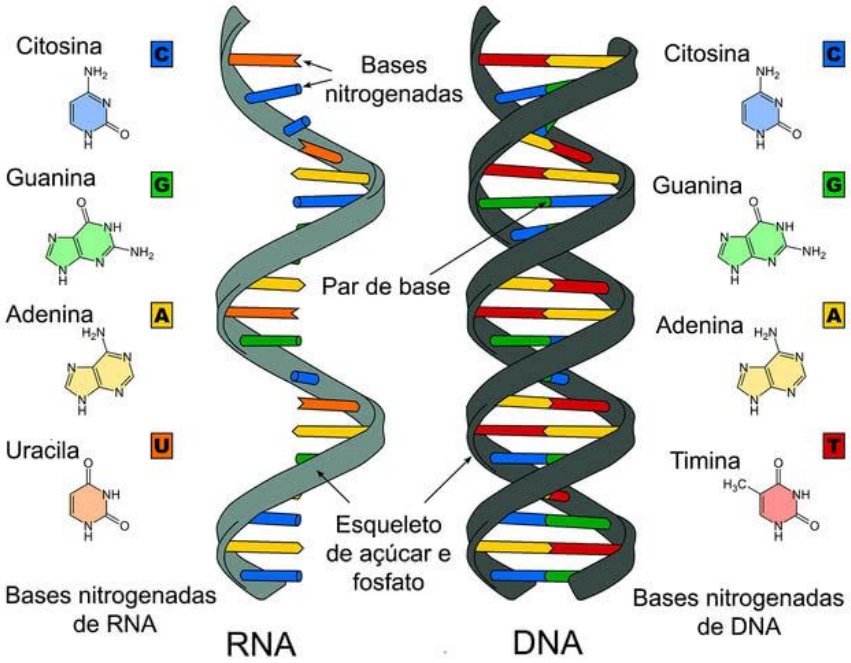
Fonte: Escola Educação (2022)

A Figura 1, corresponde à estrutura dos nucleotídeos e está relacionada à sua função. Essas moléculas são polímeros formados por cadeias de nucleotídeos, cuja composição (tipo e sequência) determina suas características químicas (Zaha, Ferreira e Passaglia, 2014).

A Figura 2, mostra como um DNA e um RNA se apresentam, e de acordo com (Alberts et al., 2017), usando métodos químicos, os cientistas podem ler o sequenciamento completo dos

nucleotídeos em qualquer molécula de DNA, e deste modo, decifrar toda a informação hereditária que cada organismo contém. No início de um processo de síntese proteica, o DNA de um gene é copiado e forma-se uma molécula linear que se chama RNA (ácido ribonucleico) e o processo de cópia é chamado de transcrição (Griffiths et al., 2006).

Figura 2- Composição do DNA e RNA

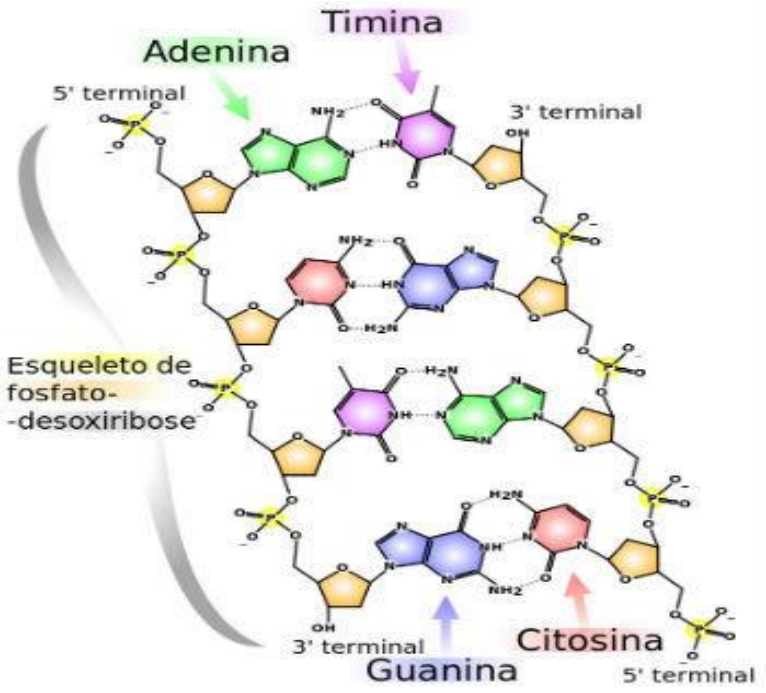


Fonte: https://www.diferenca.com/dna-e-rna/ (2023)

Como o DNA e o RNA, as proteínas carregam informações em uma forma de sequência linear de símbolos, da mesma maneira que uma mensagem humana é escrita em um código alfabético. Existem muitas moléculas diferentes de proteína em cada célula, e exceto pela água elas formam a maior parte da massa da célula (Griffiths et al., 2006).

Segundo Zaha, Ferreira e Passaglia (2014), a composição química do RNA é ilustrada na Figura 3. O processo de transcrição do RNA, define a orientação do gene que progride de 5’ para 3’.

Figura 3- Estrutura química do RNA



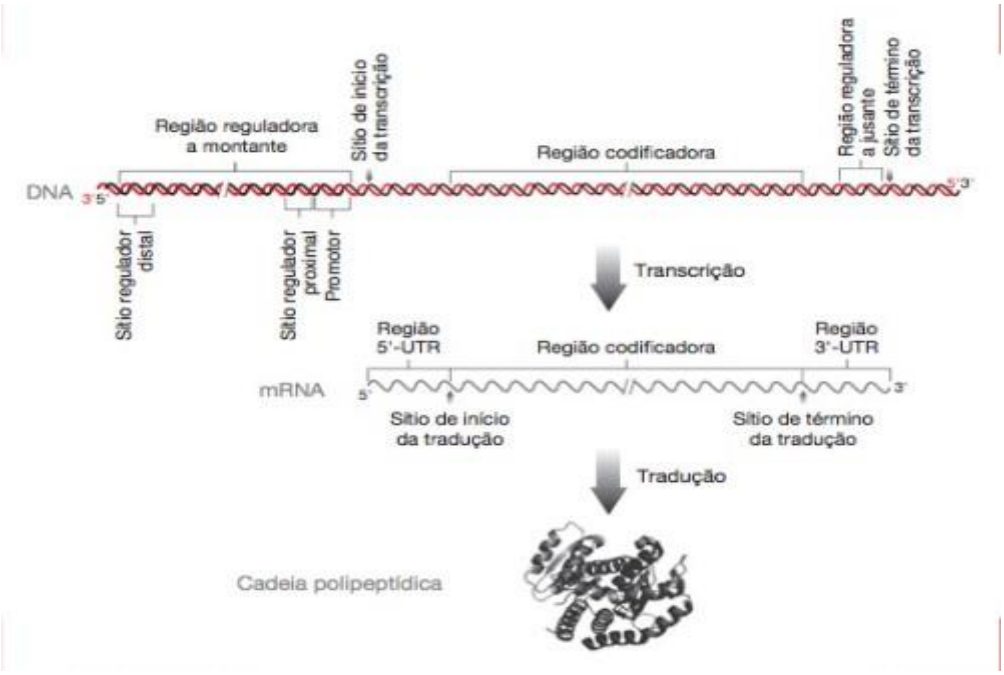
Fonte: https://www.todamateria.com.br/dna/ (2023)

Portanto, cada gene está orientado de 5' para 3' considerando a orientação da fita de DNA cuja sequência estará representada no RNA (fita codificadora). O RNA transcrito é complementar á fita-molde, a qual, tem a mesma orientação e sequência correspondente à da fita codificadora (Zaha, Ferreira e Passaglia, 2014).

A Figura 4 representa a estrutura de um gene procariótico, mas de um modo geral existem 3 principais tipos de RNA nas células, que atuam diretamente na síntese de proteínas, nomeadamente:

a) RNA mensageiro (mRNA), essa molécula é o veículo que leva a informação de um gene para a maquinaria molecular de síntese de proteínas ou seja, ele é responsável pela transferência de informação genética do DNA aos ribossomos, neste local ocorre a síntese das proteínas e representa de 1 a 5\% do RNA total da célula (Zaha, Ferreira e Passaglia, 2014);
b) RNA ribossômico, representa cerca de 75\% do RNA total da célula, e forma fitas duplas por pareamentos internos (Zaha, Ferreira e Passaglia, 2014);
c) tRNA (RNA transportador), que transporta os resíduos de aminoácidos até os ribossomos para a síntese das proteínas. Representa 10 a 15% do RNA total da célula (Zaha, Ferreira e Passaglia, 2014), ou seja, o papel principal é de levar o aminoácido para o sistema de tradução.

Figura 4- Esquema representativo da estrutura e da expressão de um gene procariótico que codifica uma proteína



Fonte: Zaha, Ferreira e Passaglia (2014)

Antes que se explore a composição genômica, é interessante conhecer um pouco do histórico do surgimento ou descoberta do vírus que causa a covid19.

2.2 Histórico e composição genética da Síndrome respiratória aguda grave

A Síndrome Respiratória Aguda Grave (SARS) é geneticamente formada por uma fita de RNA no qual o vírus, é responsável por infectar uma grande variedade de espécies de mamíferos e aves (Mittal, Ni e Seo, 2020). Ao microscópio eletrônico, ele apresenta projeções iguais a picos de glicoproteínas na sua superfície que se assemelham com o formato de uma coroa (Mittal, Ni e Seo, 2020).

Ainda na perspectiva de Mittal, Ni e Seo (2020), há pesquisadores que acreditam que alguns vírus têm origem por meio de roedores, como o HCoV-OC43 e HKU1, e por outro lado, outros acreditam que originaram dos morcegos, que consistem em HCoV-NL63, HCov-229E, SAR-CoV e MERS-CoV. Outros pesquisadores acreditam que o SARS-CoV-2 se originou diretamente do SARS-CoV, enquanto outros apontam que foi produzido no laboratório (Andersen e Godoy, 2020). No entanto, a origem do SARS-CoV-2 permanece desconhecida.

Acredita-se que os morcegos podem ser o vetor de transmissão do vírus RaTG13 pois apresenta uma alta similaridade na sequência com o SARS-CoV-2 que veio do mesmo ramo

(Ning et al., 2021). Um estudo revelou que o bat RaTG13 e o SARS-CoV-2 reconhecem o mesmo receptor ACE2 e têm a mesma capacidade de infectar células por meio desse receptor.

Os mesmos estudos também identificaram que a crista de ligação ACE2 no ponto de ligação ao receptor RaTG13 do morcego (RBM) contém quatro resíduos, que é o mesmo que SARS-CoV-2 (Ning et al., 2021).

O SARS-Cov-2, pertence ao grupo dos coronavírus, e apresenta um genoma de fita simples e envelopado (Woo et al., 2021). Os coronavírus contêm os maiores genomas de RNA conhecidos 26–32 quilobases (Platto et al., 2021), que é o genoma de vírus de RNA mais longo, conhecido até o momento (Woo et al., 2021). É altamente contagioso e causa dor de cabeça, febre, tosse, fadiga, mialgia e produz expectoração e hemoptise (Ning et al., 2021), pertencem à ordem Nidovirales, família Coronaviridae e subfamília Coronavirinae (Ning et al., 2021).

Segundo (Ning et al., 2021) o coronavírus foi isolado pela primeira vez em 1937, e em 2003 considerado patogênico para a saúde humana. A variante A é a mais próxima que foi descoberta em morcegos, que tinha 96,2\% de similaridade de sequência com o vírus humano, sendo considerado o genoma do vírus humano original (Platto et al., 2021). Ainda na perspectiva de (Platto et al., 2021) a variante B, difere da variante A por duas mutações. A variante C difere de sua variante original B por uma alteração GV e é o principal tipo europeu, inicialmente encontrado em pacientes da França, Itália, Inglaterra e Suécia (Platto et al., 2021).

Por definição, genoma é um código genético que possui toda a informação hereditária de um ser, e é codificado no DNA, ou seja, um conjunto de todos os diferentes genes que se encontram em cada núcleo de uma determinada espécie. É indispensável conhecer o genoma de uma determinada espécie, pois o genoma contém informação genética, hereditária do ser vivo. E compreender o genoma ajuda a estudar os seres a quem aqueles dados pertencem.

Ademais, no SARS-Cov-2 o genoma do RNA codifica quatro proteínas estruturais e 16 não estruturais, as proteínas estruturais são as proteínas S, M, N e E (Platto et al., 2021), e a responsabilidade das proteínas estruturais é a infecção do hospedeiro, fusão da membrana, montagem viral, morfogênese e liberação de partículas virais (Mittal et al., 2020). As proteínas E e M formam o envelope e a membrana que reveste o vírus, a proteína N se liga ao genoma do RNA e a proteína S (Spike), que forma as proteções características da superfície do vírus, interage com o receptor da membrana plasmática da célula-alvo para mediar a penetração do vírus nela.

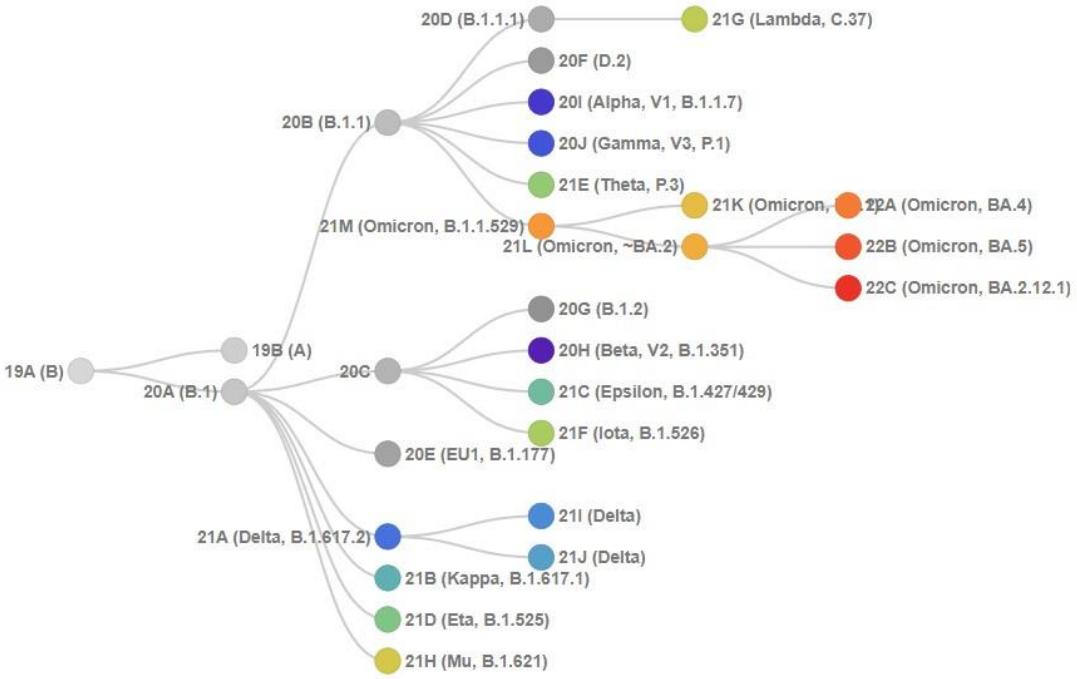
Entre essas proteínas estruturais, as proteínas S triméricas projetam-se do envelope do vírus e são a maquinaria chave que facilita a entrada do vírus na célula hospedeira (Mittal et al.,

2020), portanto as proteínas não estruturais oferecem facilidade de replicação viral e transcrição.

As variantes do vírus podem aparecer e ser mantidas em morcegos com propriedades que aumentam muito sua agressividade, pois o vírus tem a capacidade de usar o receptor ACE2 como uma via de entrada nas células hospedeiras (Platto et al., 2021).

A Figura 5, mostra como as variantes se relacionam entre si. Ademais, de acordo com o site NCBI, o genoma de referência de coronavírus relacionado à síndrome respiratória aguda grave é Viral Proj 15500.

Figura 5- Árvore filogenética do Coronaviridae



Fonte: https://covariants.org/variants(2021)

2.3 O conteúdo GC

O conteúdo GC geralmente e usado para mapear a composição do genoma e entender a evolução da sua sequência de codificação (Saini e Dewan, 2016). Para se ter o sinal referente aos genomas em estudo, iremos estimar o conteúdo GC com uma janela deslizante de tamanho $n = 100$.

Segundo Ferreira, Safadi e Ferreira (2020), o conteúdo GC é calculado como a razão da soma das bases Guanina (G) e Citosina (C) para a soma das bases Adenina (A), Guanina (G), Citosina (C) e Tinina (T).

$$GC_{content} = \frac{nG+nC}{nA+nT+nC+nG}, \quad 2.1$$

Em que: nA, nT, nC e nG representam o número de nucleotídeos das bases de A, T, C, G respectivamente. Pode-se interpretar as proporções obtidas segundo Saini e Dewan (2016), em que as regiões ricas em GC (Guanina e Citosina) podem indicar que possuem bastantes genes codificadores de proteínas, ou seja, ao determinarmos a proporção GC, podemos identificar regiões ricas em genes no genoma e ao mesmo tempo obtermos os sinais de cada sequência.

Após a obtenção do sinal de cada sequência, ira seguir a fase de decomposição do mesmo por meio da transformada não decimada de wavelet discreta.

2.4 Wavelets

As wavelet são funções capazes de decompor, descrever ou representar uma outra função, e estão descritas em função do tempo e de frequência (Daubechies,1992). Tem sido uma alternativa bastante usada para processar dados de series temporais não estacionarias em diversas áreas, inclusive na genética. Nesta seção será apresentada a fundamentação teórica relativa as wavelets, que sustenta a pesquisa desenvolvida.

2.4.1 Análise de wavelet

A análise de wavelets e uma forma matemática de analisar funções, que são bastante usadas por pesquisadores de diversas áreas nos últimos anos. As wavelet são funções que podem ser suaves ou não, simétricas ou não e podem ser expressões matemáticas simples ou não (Morettin, 2014), (Morettin, Chiann e Montoril, 2014), (Montoril, Morettin e Chiann, 2018)).

Segundo Denault et al. (2021), elas são funções matemáticas que possuem muita utilidade e ajudam-nos a conduzir transformadas, do tipo transformada de Fourier ou transformada de Haar, dentre outras.

Brassarote (2014) vinculou que uma das várias aplicações das wavelets tem sido em áreas de análise de sinais de equações diferenciais, isto deve-se principalmente pela sua capacidade de localização no tempo-frequência contrariamente ao que ocorre com a análise de Fourier, trazendo desta forma uma perspectiva diferente e inovadora no tratamento de um banco de dados de uma função não estacionária.

Segundo (Morettin, 2014), as wavelets satisfazem algumas propriedades indispensáveis, nomeadamente:

P1. $\int_{-\infty}^{\infty} \Psi(t)dt = 0$ (Admissibilidade);

P.2. $\int_{-\infty}^{\infty} \Psi(t)dt < \infty$

P.3. $C_{\Psi} = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega < \infty$ em que $\Psi(\omega)$ é a transformada de Fourier de $\Psi(t)$

P.4. $\int_{-\infty}^{\infty} |\Psi(t)|^2 \, dt = 1 \; \int_{-\infty}^{\infty} |\Psi(\omega)|^2 \; d\omega = 2\pi$

P.5 Os primeiros $r-1$ momentos de Ψ anulam-se, isto é: $\int_{-\infty}^{\infty} t^j \Psi(t)dt = 0, j = 0, 1, \dots, r-1$ para algum $r \leq 1$ e $\int_{-\infty}^{\infty} |t^r \Psi(t)| < \infty$.

Em que o valor de r está ligado ao grau de suavidade (regularidade) de Ψ: quanto maior for o r, mais suave será Ψ.

O mais indicado e que as wavelets tenham suporte compacto, e no caso em que a wavelet tem suporte compacto, o valor de r está relacionado com o suporte da wavelet (Morettin, 2014).

2.4.2 Transformada de wavelet

A análise das wavelets incorpora uma função de wavelet mãe que segundo (Silva et al., 2009), esta wavelet mãe tem algumas especificações importantes, e uma função com media zero que decai bruscamente de forma oscilatória em formato de ondas, e as suas transformadas surgem da wavelet mãe pela superposição de versões dilatadas e transladadas da wavelet.

A transformada de wavelet e capaz de fornecer a informação de tempo e de frequência simultaneamente, consequentemente dando a representação de frequência-tempo da série, além disso, são uteis para que se possa observar algumas características ou informação referente a um sinal, que estejam presentes nelas, mas que por algum motivo não podem ser observadas em um dado domínio (Brassarote, 2014).

A principal diferença que existe entre as bases de Fourier e as wavelet está no fato de que as bases de Fourier são localizadas em frequência, porém, não no tempo (Morettin, 2014), mas por outro lado, a transformada de wavelet nos permite representar função como conjunto de coeficientes (Denault et al., 2021).

Ademais, pequenas mudanças em algumas observações podem criar mudanças em todas as componentes de uma expansão de Fourier, enquanto que o mesmo não acontece ao trabalharmos com uma expansão em series de wavelet (Morettin, 2014).

Ainda na perspectiva de Morettin (2014), a ideia fundamental tanto na análise de Fourier quanto na análise de wavelet, e de aproximar-se uma função por meio de combinação linear de senos e cossenos e wavelet, respectivamente.

Todo o conjunto de wavelets e gerado com base numa wavelet designada de wavelet mãe (Ferreira, Safadi e Ferreira, 2020). Segundo Daubechies (1992), as transformadas de wavelet são baseadas no processo de escalar e transformar no tempo a wavelet mãe em wavelet filha. A equação 2.2, descreve o comportamento da wavelet:

$$(T^{waw}f)(ab) = |a|^{-\frac{1}{2}} \int f(t)\Psi\left(\frac{t-b}{a}\right)dt, \qquad 2.2$$

Em que *a* é o fator de escala, *b* é o fator de translação no tempo e Ψ é a wavelet mãe.

A equação 2.2 descreve de forma generalizada uma forma de se adquirir wavelets filhas, baseado em escalar e transladar a wavelet mãe. Ademais, segundo (Biazon e Bianchi, 2020) a função de wavelet mãe satisfaz a seguinte condição da equação 2.3:

$$\int \Psi(t)dt = 0 \qquad 2.3$$

Na equação 2.3 e apresentado um exemplo retirado do livro de Daubechies (1992), em que se pode ter uma ilustração visual da escala e translação no tempo das wavelet, com base na utilização da segunda derivada da Função Gaussiana, descrita pela equação 2.4:

$$\Psi(t) = (1 - t^2)e^{-\frac{t^2}{2}}. \qquad 2.4$$

A análise de wavelets e uma nova forma de estudar series temporais que teve o seu início em 1980 (Ferreira, Safadi e Ferreira, 2020). A transformada de wavelets e uma ferramenta poderosa para detectar as mudanças sutis em um sinal (Das e Kumar, 2021), e tem a capacidade de decompor e de representar um conjunto de dados que se dispõe e se descreve ao longo do tempo, de modo que o pesquisador possa analisar esta função em diferentes escalas da frequência e do tempo (Morettin, 1999).

Geralmente, os pesquisadores tem mudado das transformadas de Fourier para a transformada de wavelets pois de acordo com Wrobel e (Galvão et al., 2001), a transformada de Fourier, e ideal para fazer análise de sinais de diagnósticos estacionários, enquanto que a transformada de wavelets e mais adequada para os casos em que os sinais são não estacionários, no entanto, na perspectiva de Daubechies (1992) uma transformada de wavelets vem a fornecer localização por tempo e frequência, e servem como uma alternativa melhor em relação as transformações de Fourier em muitos algoritmos de processamento de sinais. As transformadas de wavelet são geralmente um método mais adotado entre os pesquisadores pelo fato de detectarem mudanças bruscas no tempo e na frequência ao mesmo tempo que oferecem

especificidade ((Pittner; Kamarthi, 1999), (Kamarthi; Kumara; Cohen, 2000)). Ademais, na perspectiva de (Biazon; Bianchi, 2020), as transformadas de wavelet são classificadas em duas categorias: Transformadas Continuas de wavelets e Transformadas Discretas de wavelets.

2.4.3 Transformada Discreta de wavelet

A transformada discreta de wavelet, considera apenas variações discretas dos parâmetros a e b. Para a toma-se números positivos e negativos elevados a um parâmetro de escala fixo $a > 1$, que e $a = a_0^m$, e o parâmetro m determina a largura das wavelets (Daubechies, 1992). O parâmetro de translação b também depende do parâmetro m. Wavelets estreitas (alta frequência) são transladadas por pequenos passos para cobrir toda a faixa no tempo, enquanto wavelets largas (baixa frequência) são transladadas por passos maiores. Portanto para discretizar b e escolhido $b = nb_0a_0^m$, onde $b_0 > 0$ e fixo, e $n \in Z$. Então, podemos descrever as wavelets classificadas discretas como:

$$\Psi_{m,n}(x) = a_0^{-\frac{m}{2}} \Psi(a_2^m x - nb_0). \qquad 2.5$$

Então assume-se um Ψ razoável e a_0 e b_0 apropriados, a reconstrução da funções pode ser feita por:

$$f = \sum_{m,n} (f, \Psi_{m,n}) \Psi_{m,n}. \qquad 2.6$$

Que pode ser traduzido como sendo a função de reconstrução a soma de todos os coeficientes de wavelet $(f,\ \Psi_{m,n})$ para cada m e n, sobre a inversa da wavelet $\Psi_{m,n}$.

2.4.4 Algoritmo Piramidal

A transformada de wavelets discreta e implementada pela aplicação do algoritmo piramidal (Mallat, 1989), o pesquisador implementou um modo rápido e confiável de se fazer uma análise de multi resolução via sinal piramidal.

O algoritmo faz menção ao processo de obtenção de aproximações e detalhamento de um dado, de modo que a aproximação corresponde a uma representação de baixa frequência do sinal em estudo, e por outro lado, o detalhamento corresponde a diferença entre duas aproximações sucessivas do sinal original ((Reis e Silva, 2004), (Reis e Silva, 2005)), ou seja, o algoritmo piramidal efetua cálculos usando os filtros passa baixo (chamados de wavelet pai, hk) e passa alto (chamados de wavelet mãe, gk) das funções wavelets cujos coeficientes são dados por (Morettin, 1999):

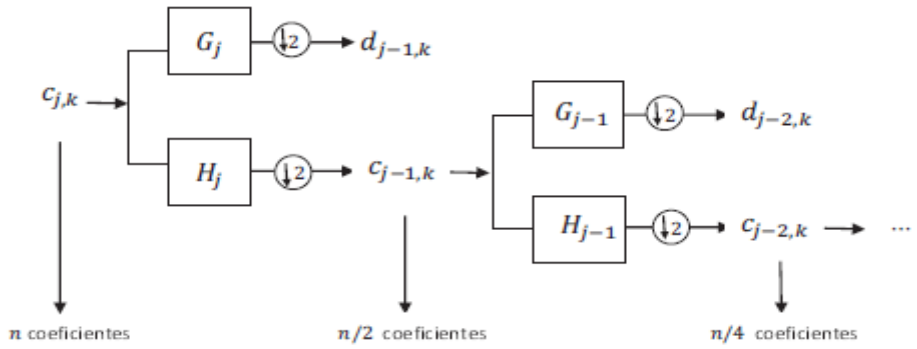
$$h_k = \sqrt{2} \int_{-\infty}^{+\infty} \phi(t)\phi(2t-k)dt \qquad 2.7$$

$$g_k = \sqrt{2} \int_{-\infty}^{+\infty} \Psi(t)\phi(2t-k)dt \qquad 2.8$$

Este algoritmo calcula a transformada discreta a partir dos coeficientes suaves. A transformada de wavelet discreta e calculada usando filtros passa baixa H_j e filtros passa alta G_j e uma decimação designada por *downsampling* (indicada por ↓ 2) ou *upsampling* (indicado por ↑ 2) para os processos de decomposição e reconstrução respectivamente, esse processo de *downsampling* exclui amostras pares ou impares de modo que com o decorrer do processo, verifica-se uma redução dos coeficientes do nível j+1 e consequentemente ha perda de informação (Negri e Souza, 2012).

A Figura 6 ilustra o processo de decomposição da transformada de wavelet discreta decimada, pelos filtros passa baixa e passa alta, onde pode-se obter os detalhes de alta frequência de um sinal pela diferença de informação de um nível j do algoritmo e um mais refinado designado por j+1 (Negri e Souza, 2012), em que H_j e o filtro passa baixa, e o G_j e o filtro passa alta. O ↓ 2 indica o *downsampling*.

Figura 6- Decomposição da transformada de wavelet discreta pelo algoritmo piramidal



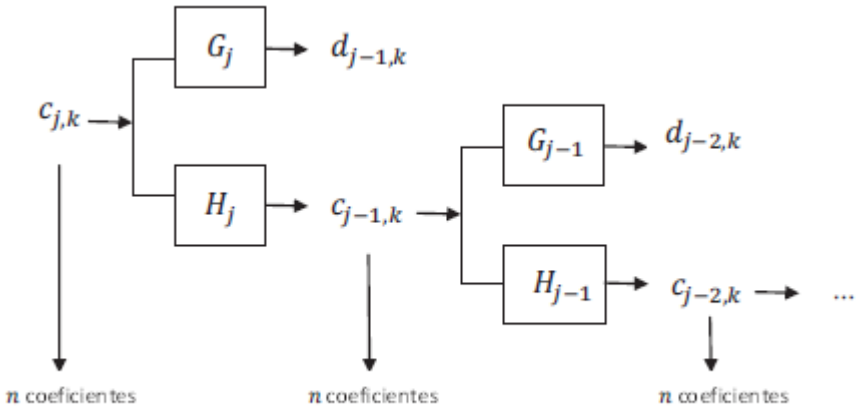
Fonte: Negri e Souza (2012)

2.4.5 Transformada de wavelet discreta não decimada

De acordo com Percival e Walden (2000), a transformada de wavelet não decimada é um processo no qual é feita a implementação da transformada discreta decimada mas com uma pequena modificação na implementação, pois o processo de *downsampling* é eliminado, ou seja, na transformada wavelet não decimada não ocorre perda de informação.

Segundo Nason e Silverman (1995), em cada nível, os filtros passa baixa e passa alta sofrem uma modificação e são inseridos zeros. Portanto, os coeficientes suaves $c_{j,k}$ e os coeficientes de detalhes $d_{j,k}$ podemos obter a partir de $c_j + 1, k$ mas sem precisar de realizar o downsampling(Negri e Sousa, 2012), e a Figura 7 mostra-nos o cálculo da transformada wavelet não decimada de um sinal.

Figura 7- Decomposição da transformada de wavelet não decimada



Fonte: Negri e Souza (2012)

A vantagem que existe em se desenvolver um estudo com recurso á transformada wavelet não decimada reside no fato desta ser invariante à translação, porque ela toma em consideração os elementos pares e ímpares. Desenvolvemos esta pesquisa com base na transformada wavelet não decimada.

Na perspectiva de Ferreira, Safadi e Ferreira (2020), se tomarmos as funções de escala de wavelet dadas por $\phi\, e\, \psi$ respectivamente, podemos representar o vetor de dados $y = y_0, y_1, \dots, y_{m-1}$ com o tamanho *m*, como uma função *g* em termos de deslocamento da função de escala em algum nível de multi-resolução *h* de modo que $j - 1 < log_2 m \leq j$

$$g(x) = \sum_{k=0}^{m-1} y_k \phi_{j,k}(x), \qquad 2.9$$

Em que

$$\phi_{h,k}(x) = 2^{\frac{j}{2}} \phi\left(2^j (x - k)\right). \qquad 2.10$$

A função de interpolação de *g(x)* pode ser escrita novamente, segundo a equação 2.11 (Ferreira, Safadi e Ferreira, 2020):

$$g(x) = \sum_{k=0}^{m-1} c_{j_0 k}\phi_{j_0 k}(x) + \sum_{j=j_0}^{j-1}\sum_{k=0}^{2^{n-1}} d_{jk}2^{\frac{j}{2}}\psi\left(2^j(x-k)\right), \qquad 2.11$$

Onde

$$\phi_{j_0,k}(x) = 2^{\frac{j_0}{2}}\phi\left(2^{j_0}(x-k)\right) \text{ e} \qquad (2.12)$$

$$\psi_{jk}(x) = 2^{\frac{j}{2}}\psi\left(2^j(x-k)\right), \qquad (2.13)$$

Em que $j = j_0, \dots, j-1; k = 0, 1, \dots, m-1$.

O vetor y da transformada de wavelet discreta não decimada é representado pelos coeficientes: $c_{h_{0,}k}$ e $d_{jk}, j = j_0, \dots, j-1; k = 0, \dots, m-1$.

2.4.6 Wavelets de Daubechies

As wavelets de Daubechies pertencem a família de wavelets ortogonais e caracterizam-se pela quantidade de momentos nulos de acordo com um suporte estabelecido, além disso, definem uma transformação de wavelets discretas, pois de acordo com (Rashid, Amin e Lone, 2020), a wavelet de Daubechies pode tornar a analise discreta de wavelets praticável.

É vantajoso trabalhar com funções de wavelets e funções de escala de Daubechies como bases funcionais para a aproximação multi resolução, pois conseguem alterar de forma radical a formulação e a solução das equações de movimento (Nastos e Saravanos, 2021).

As wavelets de Daubechies são definidas pelo número de momentos nulos (m), o que corresponde ao comprimento do filtro (2m) (Rashid, Amin e Lone, 2020).

2.4.7 Escalograma

Um sinal no domínio do tempo e possível ser transformado com a aplicação das transformadas de wavelet, isso permite ao pesquisador fazer uma análise analítica nas propriedades ou características ocultas que descrevem o sinal (MEDAIYESE et al., 2022).

O escalograma é um gráfico que mostra a energia de cada sequenciamento, em cada nível de resolução, e na perspectiva de (Magrini, Domingues e Junior, 2017), os escalogramas baseados na wavelet, são capazes de preservar as estruturas de energia presentes no sinal original, ainda que exista perda de energia.

A transformada discreta de wavelet nos fornece um escalograma que mostra a distribuição de energia no sinal no domínio da escala do tempo (Medaiyese et al., 2022).

A capacidade de detectar componentes diferentes e uma característica importante das wavelets, e diferentes componentes resultarão em picos visíveis no escalograma (Ferreira, Safadi e Ferreira, 2020). Ademais, esses diferentes componentes podem ser extraídos do sinal pela divisão dos coeficientes da wavelet pelos diversos conjuntos pertencentes ao mesmo pico (Ferreira; Safadi; Ferreira, 2020).

A energia na transformada discreta nao decimada de wavelet pode ser calculada por meio da formula (Gencay, Selcuk e Whitcher, 2002).

$$E(j) = \sum_{k=0}^{n} d_{j,k}^{2} \ , \qquad 2.14$$

Para $j = 1, 2, \dots, J$

2.5 Análise de Agrupamento

Analisar agrupamento e o mesmo que usar variadas técnicas computacionais com a finalidade de classificar objetos com base nas suas características. Em várias pesquisas agrupam-se objetos que possuam caraterísticas semelhantes, a ideia fundamental e de analisar em grupos por similaridades, pois este critério permite fazer analise numa mesma classe de objetos em que já foram quantificados e agrupados previamente por meio de medidas de proximidade que podem envolver tantas medidas de similaridades quanto as medidas de dissimilaridade (Ferreira, 2008). Duas formas de agrupamentos podem ser usadas, método de agrupamentos hierárquicos aglomerativos e método de agrupamento por divisão (Manly, 2008), (Manly e Alberto, 2016).

Segundo (Fernando, 2006) os métodos estatísticos procuram organizar os objetos em grupos homogêneos, aplicando para isso o conceito de similaridade. Ainda na perspectiva de (Fernando, 2006), obtém-se a similaridade por meio de coeficientes, em que a escolha desses coeficientes de similaridade depende da escala de mensuração da variável. Portanto, quanto maior for o valor obtido pela similaridade, mais semelhante serão os objetos.

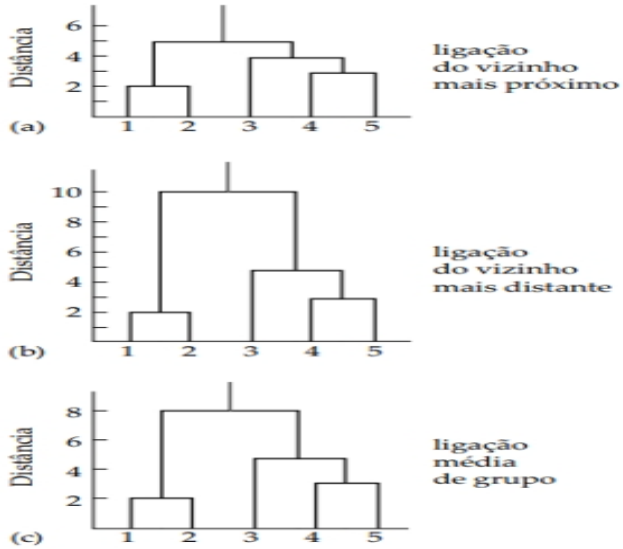
Ainda na perspectiva de Fernando (2006), os métodos hierárquicos aglomerativos seguem um algoritmo específico:

a) procuram-se os dois objetos mais similares na matriz de similaridade;
b) os objetos *i* e *j* são retirados e formam um grupo, em que se eliminam a linha e a coluna correspondente *i* e *j*;
c) são definidas uma linha e uma coluna que são obtidas pelas distâncias entre o grupo *i j* e os objetos restantes, de acordo com o procedimento do algoritmo adotado;

d) os passos anteriores são repetidos *n-1* vezes, de modo que todos os *n* objetos tenham um grupo até o fim do algoritmo.

Uma forma de representar os métodos hierárquicos aglomerativos e por meio de representação gráfica e por dendrogramas como ilustra a Figura 8, portanto, são de fácil entendimento.

Figura 8- Exemplos de dendogramas de análise de agrupamentos de cinco objetos.



Fonte: Manly(2008)

Nesta pesquisa usou-se técnicas hierárquicas pois produzem um dendrogramas, e o método começa com o cálculo das distâncias de cada objeto a todos os outros objetos. Quando se emprega técnicas hierárquicas aglomerativas na formação de agrupamentos, todos os objetos começam sozinhos em grupos de um, e os próximos grupos são formados gradualmente até que finalmente todos os objetos estejam em algum grupo (MANLY, 2008) e (MANLY; ALBERTO, 2016). Portanto, as distancias de cada objeto a todos objetos serão calculadas usando a função de distância de Mahalanobis.

Segundo (YANG; DELPHA, 2022), a distância de Mahalanobis de uma amostra e calculada considerando dois vetores de amostra X, Y provenientes de mesma distribuição e é dada por:

$$d_M(X) = \sqrt{(X-\mu)^T \Sigma^{-1} (X-\mu)}. \qquad 2.15$$

Em que μ é um vetor médio, e Σ é matriz de covariância. Portanto, a distância de Mahalanobis é sempre um valor não negativo.

A vantagem de se trabalhar com a distância de Mahalanobis é que as distâncias são calculadas em unidades de desvio, partindo da média do grupo (Ferreira, 2014), e além disso, é

uma distância em que o cálculo toma em consideração a variabilidade, em vez de tratar todos os dados valores da mesma forma quando se procede ao cálculo da distância ao ponto central (Mclachlan, 1999).

2.6 Expoente de Hurst

Ao longo da pesquisa definiram-se 5 métodos diferentes para determinar o expoente de Hurst, e os valores obtidos para cada nível de decomposição da transformada de wavelet foram utilizados na formação dos grupos para a análise de similaridades. O expoente de Hurst e definido no intervalo (0,1), e é estritamente menor do que 0,5 em sequencias que se apresentam anticorrelacionadas de forma grosseira, e se situa no intervalo de 0,5 < H < 1 quando as sequencias são positivamente correlacionadas. Para os casos em que o expoente de Hurst e igual a meio, ou seja $H = 0.5$, pode-se concluir que a sequência apresenta padrão aleatório (Rout et al., 2022).

No estudo que se pretende desenvolver, irá se fazer uso do expoente de Hurst de modo a identificar similaridades nas sequencias que se propõe. Ademais, será feita a estimativa do expoente de Hurst com base em cinco métodos distintos, que serão descritos de forma resumida a seguir.

Cada uma das metodologias de estimativa do expoente de Hurst será aplicado em cada um dos 6 níveis, em cada uma das sequencias em estudo, sem uma ordem especifica para cada método.

2.6.1 Análise R/S

Esta análise e descrita no artigo de (Maftei, Barbulescu e Carsteanu, 2016). Uma das formas de se calcular o expoente de Hurst e por meio da análise R/S. Portanto, pode-se estudar a dependência de longo alcance com base num algoritmo que divide a serie temporal em d subséries com um comprimento m (Maftei, Barbulescu e Carsteanu, 2016).

Ainda na perspectiva de (Maftei, Barbulescu e Carsteanu, 2016), na análise R/S, para cada série de $n = 1, 2, 3, \dots, d$:

a) inicialmente encontra-se a média (E_n) e o desvio padrão (S_n);
b) a seguir se normaliza a base de dados(X_{in}) subtraindo-se a média das sub-séries como se mostra a seguir: $Z_{in} = X_{in} - E_{in}, i = 1, \dots, m$;
c) a seguir cria-se uma série temporal cumulativa designada por Y_{in}(Maftei, Barbulescu e Carsteanu, 2016) $Y_{in} = \sum_{j=1}^{i} Z_{jn},\ i = 1, \dots, m$;

d) em seguida encontra-se o intervalo ajustado (Ferreira, Safadi e Ferreira, 2020)

$$R(t,k) = \max_{0\le i\le k}[Y_{t+i} - Y_t - \frac{i}{k}(Y_{t+k} - Y_t)] - \min_{0\le i\le k}[Y_{t+i} - Y_t - \frac{i}{k}(Y_{t+k} - Y_t)]; \quad (2.16)$$

e) redimensiona-se o intervalo R_n/S_n;

f) e calcula-se o valor médio do intervalo ré escalonado para todas as subséries de comprimento *m.*

$$(R/S)_m = \frac{1}{d}\sum_{n=1}^{d} R_n / S_n \,. \quad (2.17)$$

Ademais, Maftei, Barbulescu e Carsteanu (2016), afirmam que na prática na análise R/S o expoente de Hurst pode ser estimado como a inclinação do gráfico log/log de *t.*

E um fator importante a ser considerado é que ao em vez de usar o desvio padrão para normalizar a amostra, a análise considera o seguinte desvio (Ferreira, Safadi e Ferreira (2020):

$$S(t,k) = \sqrt{k^{-1}\sum_{i=t+1}^{t+k}(X_i - \bar{X}_{t,k})^2}\,, \quad 2.18$$

em que

$$\bar{X}_{t,k} = k^{-1}\sum_{i=t+1}^{t+k} X_i\,. \quad 2.19$$

Portanto, a proporção padronizada será dada pelo quociente entre R(t,k) e $S(t,k)$:

$$Q(t,k) = \frac{R(t,k)}{S(t,k)} \quad 2.20$$

Que segundo Ferreira et al. (2020) é costumeiramente conhecido como intervalo ajustado ré escalonado ou simplesmente estatística R/S. Cujas hipóteses testadas são as seguintes:

$$\begin{cases} H_0: ausência\ de\ dependência\ de\ longo\ alcance \\ H_1: presença\ de\ dependência\ de\ longo\ alcance \end{cases}$$

Pretende-se executar o passo a passo proposto pela metodologia e, ao se implementar esta análise, os valores de Hurst serão obtidos em cada nível, e posteriormente far-se-á uma análise de agrupamento.

2.6.2 Método de variância agregada

Este método é também descrito no artigo de Maftei, Barbulescu e Carsteanu (2016), em que diz que, se considerarmos uma série com o tamanho N, ela é dividida em *d* subséries de comprimento *m*, e para cada série agregada, composta por uma média que é dada por:

$$X^m(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i, \qquad 2.21$$

e a variância amostral é dada por:

$$VarX^m = \frac{1}{d} \sum_{k=1}^{d} (X^{(m)}(k) - \bar{X})^2 \qquad 2.22$$

Maftei, Barbulescu e Carsteanu (2016), afirmam ainda que para valores sucessivos de *m*, a variância amostral é computada em um gráfico log-log contra *m*. E os mínimos quadrados serão ajustados em uma linha para todos os pontos do gráfico, e o coeficiente de Hurst será calculado tendo em conta a inclinação da linha reta que é dada por *2H-2*.

2.6.3 Método de variância agregada diferenciada

Esta metodologia foi descrita no trabalho de Teverovsky e Taqqu (1995), em que examinaram o efeito de certos tipos de não estacionariedade na detecção de dependência de longo alcance na estimativa de Hurst de parâmetro H, quando se aplica o estimador da variância. A finalidade da aplicação deste método é de se distinguir o caso em que H está próximo de 0.5 e existe uma certa tendência diferente de zero, dos casos em que H se distancia de 0.5, ou por outra, acaba sendo significativamente maior do que 0.5.

Segundo Teverovsky e Taqqu (1995), para se detectar uma dependência de longo alcance é necessário analisar-se a variação da amostra de uma série temporal $X = \{X(i),\ i = 1,2, \dots, \}$, em vários níveis de agregação *m*, na série agregada de ordem *m*. Ao tomarmos a equação que se segue,

$$X^m(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i\,, \qquad 2.23$$

Com $k = 1, 2, \dots, d$

Portanto, a variação da amostra da série agregada X^m será dada por:

$$\widehat{var}X^m = \frac{1}{N/m}\sum_{k=1}^{N/m}\{X^{(m)}(k)\}^2 - \left\{\frac{1}{\frac{N}{m}}\sum_{k=1}^{\frac{N}{m}}X^{(m)}(k)\right\}^2 . \qquad 2.24$$

Segundo Ferreira, Safadi e Ferreira (2020), o método da variação agregada é um estimador do tipo variância em que se pode adquirir as estimativas por meio do logaritmo da diferença de primeira ordem. Ao considerarmos a sequência dada por $m_1, m_2, \ldots$ dos valores de m, e tomarmos a diferença, obtêm-se o seguinte (Teverovsky e Taqqu, 1995):

$$\frac{d\widehat{var}\ X^m}{dm} \approx \beta C_2 m^{\beta-1}. \qquad 2.25$$

E assume-se que

$$\Delta\ \widehat{var}\ X^{(m)} \approx \frac{d\widehat{var}\ X^{(m)}}{dm}\Delta m\,, \qquad 2.26$$

Logo

$$\log\left(\Delta\widehat{var}\ X^{(m)}\right) \approx \log\left(\frac{d\widehat{var}\ X^{(m)}}{dm}\right) + \log\Delta m\,. \qquad 2.27$$

Analogamente segundo Ferreira, Safadi e ferreira (2020), é necessário considerar que por outro lado

$$\frac{d}{dm}Var[\bar{X}_m(k)] \approx (2H-2)Cm^{2H-3}. \qquad 2.28$$

Uma vez que irá se trabalhar com gráficos log-log Ferreira, Safadi e Ferreira(2020), sugerem as seguintes aproximações:

$$log\Delta Var[\bar{X}_m(k)] = (2H-3)\log m + \log(2H-2)C + \log m + C_1 = (2H-2)\log m + C_2\,. \qquad 2.29$$

Ao se aplicar este método, espera-se que obtenha-se igualmente uma reta com a inclinação igual a $2H-2$.

2.6.4 Método dos momentos absolutos

Os momentos absolutos são calculados para a série agregada (Maftei, Barbulescu e Carsteanu, 2016), e o n-ésimo momento absoluto é dado por:

$$AM_n^{(m)} = \frac{1}{(N/m)}\sum_{k=1}^{(N/m)}|\bar{X}_m(k) - \bar{X}_N|^n\,, \qquad 2.30$$

em que $AM_n^{(m)}$ é assintoticamente proporcional a $m^n(H-1)$.

Neste método as estimativas de Hurst serão calculadas partindo do cálculo de $AM_n^{(m)}$ tomando em conta diferentes valores de *m* e de seguida gera-se um gráfico log-log contra *m.* O que se espera ao aplicar-se esta metodologia é que os pontos fossem espalhados ao longo de uma linha reta, com inclinação dada por $n(H-1)$.

2.6.5 Método Peng

Ferreira, Safadi e Ferreira (2020), fizeram uma descrição muito bem detalhada deste método, em que estabelecem duas etapas fundamentais para a execução da metodologia.

1º Passo: deve-se observar cada bloco de certo tamanho *m* e calcular a soma parcial dentro dos blocos segundo a equação 2.31.

$$Y(k)^m = \sum_{t=(k-1)m+1}^{km} X(t)\,, \qquad 2.31$$

Em que $k = 1, 2, \dots, (N/m)$

2º Passo: esta etapa compreende a fase de ajustamento de uma linha de regressão do tipo $g = a + bk$, que irá nos ajudar a calcular a variância do resíduo, que é dada pela equação 2.32

$$s_r^{(m)} = \frac{1}{m}\sum_{k=1}^{N/m}(Y(k)^m - a - bk)^2. \qquad 2.32$$

3º Passo: esboçar os gráficos de $logs_r^{(m)}\ versus \log m$

4º Passo: Por fim, a inclinação a ser obtida deve ser igual a *2H.*

2.7 Regularização

Na maior parte dos estudos recentes, para um conjunto de variáveis $X_1, X_2, \dots, X_P$, sua relação com a variável resposta pode ser descrita através do modelo linear padrão que e dado por (James et al., 2013):

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \varepsilon \qquad 2.33$$

Na configuração de regressão o modelo linear padrão tem a vantagem em termos de inferência e em situações de modelagem no mundo real, e tem se apresentado como uma opção bastante competitiva comparativamente com os métodos não lineares (James et al., 2013). Portanto, os modelos de regressão linear simples podem ser melhorados através da substituição do ajuste simples de mínimos quadrados por um outro ajuste alternativo. Estes ajustes

alternativos têm a vantagem de proporcionar melhor precisão de previsão e melhor interpretação do modelo (James et al., 2013).

Ademais, uma melhor precisão na previsão do modelo e possível desde que a relação entre a variável resposta e os preditores sejam aproximadamente lineares, os mínimos quadrados terão as estimativas com um viés muito baixo. A implementação de penalização nos modelos ajustados e com o intuito de resolver problemas de *"overfitting"*. Portanto, incluir penalizações aos coeficientes estimados reduz de forma significativa a variância e aumenta muito pouco o viés, o que proporciona uma melhoria na precisão na qual pode-se prever a variável resposta.

Os métodos de penalização que serão usados ao longo deste trabalho, envolvem o uso de mínimos quadrados para ajustar o modelo de regressão. Portanto, as duas técnicas mais conhecidas para reduzir os coeficientes de regressão em direção a zero estão a regressão ridge e o lasso. Ademais, o procedimento de ajuste de mínimos quadrados estima os $\beta_1, \beta_2, \ldots, \beta_p$ usando os valores que minimizam (James et al., 2013).

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 . \qquad 2.34$$

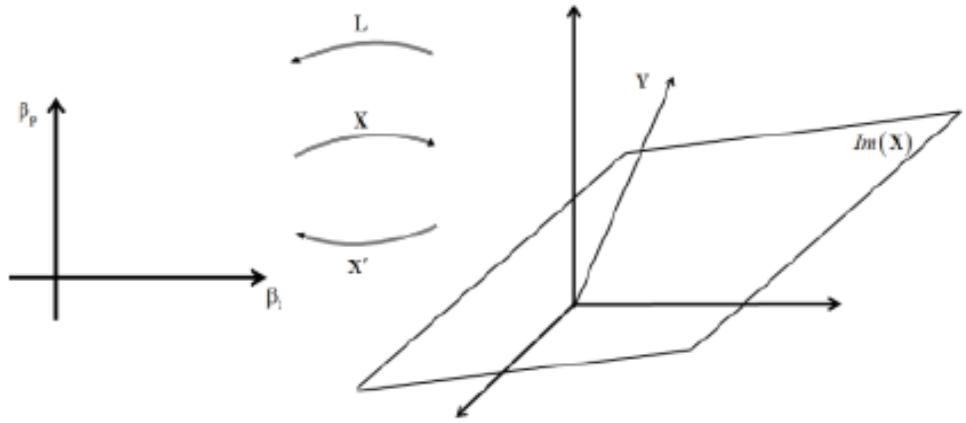
2.7.1 Regularização Ridge

A regressão ridge e uma metodologia que foi proposta pela primeira vez em 1970 por Arthur Hoerl e Robert Kennard, como forma de melhorar o teorema de Gauss Markov atribuindo-se ao teorema o estimador denominado ridge. O estimador ridge e um estimador de encolhimento, e foi desenvolvido por Hoerl e Kennard (1970). O estimador de encolhimento ridge foi desenvolvido a partir da aplicação de uma regularização ou penalização no método de quadrados mínimos.

Hoerl e Kennard (1970), começam considerando uma regressão linear dada por $y = X\beta + \varepsilon$, no qual existe um modelo k que é o subespaço *lm*(X)(Hoerl e Kennard, 1970). Tomando um certo vetor y, a projeção ortogonal deste vetor y em *lm*(X), é que dá origem ao estimador de mínimos quadrados (Pereira, 2017).

Inclui-se uma regularização considerando que as possíveis estimativas se encontram a uma certa distância r da projeção ortogonal (Pereira, 2017) como ilustra a Figura 9.

Figura 9- Interceptação do vetor y



Fonte: Pereira (2013)

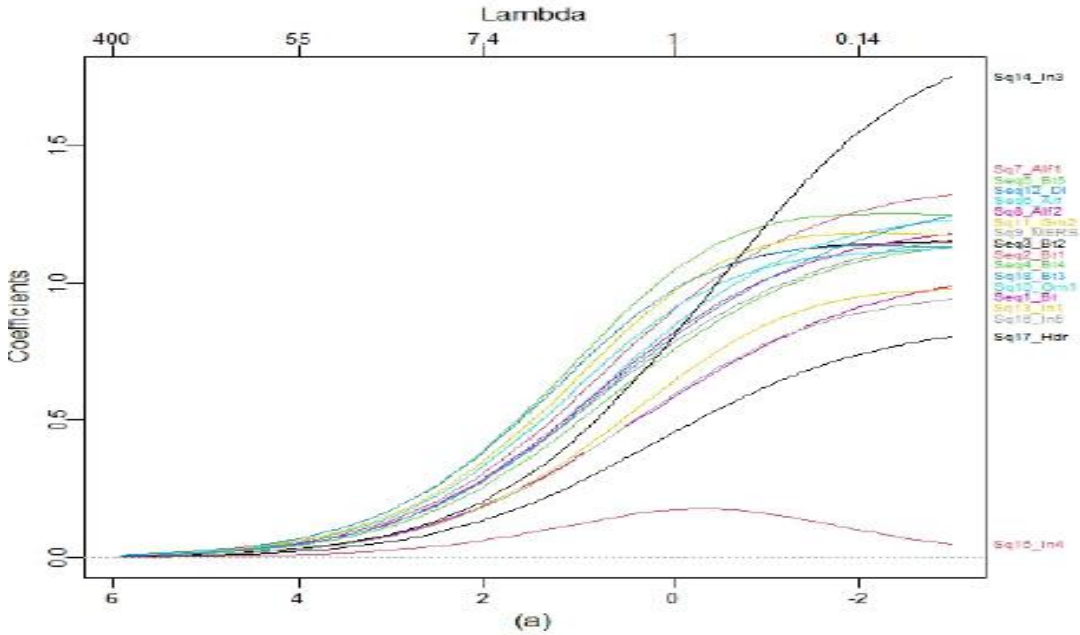
Segundo (James et al., 2013) a regressão ridge busca estimativas de coeficientes que se ajustem bem aos dados, deste modo os β são os valores que minimizam (James et al., 2013)

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2, \qquad 2.35$$

Onde $\lambda \geq 0$ é um parâmetro de sintonia, a ser determinado separadamente.

Na equação 2.35,a penalidade de encolhimento é aplicada aos $\beta_1, \beta_2, \ldots, \beta_p$, mas não para a interceptação β_0 como mostra a Figura 10.

Figura 10- Aplicação do método Ridge na base de dados em estudo



Fonte: A autora (2023)

2.7.2 Regularização Lasso

O método Lasso foi proposto por Tibshirani (1996), é uma metodologia dedicada ao menor encolhimento absoluto e a seleção de variáveis. A metodologia é descrita no artigo de (Tibshirani, 1996).

Se considerarmos $\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p\right)^T$, a estimativa do lasso $\left(\hat{\alpha},\ \hat{\beta}\right)$ é definida por:

$$\left(\hat{\alpha}, \hat{\beta}\right) = argmin\left\{\sum_{i=1}^{N}\left(y_i - \alpha - \sum_j \beta_j x_{ij}\right)^2\right\}. \qquad 2.36$$

Na condição de que $\Sigma_j \left|\beta_j\right| \leq t$. E que $t \geq 0$, portanto o encolhimento é causado por valores de $t < t_0$, observando que, alguns coeficientes podem também ser iguais a zero. O parâmetro $t \geq 0$ controla a contradição aplicada as estimativas (Tibshirani, 1996).

Considerando $\hat{\beta}_j^0$ estimativas dos mínimos quadrados, toma-se $t_0 = \left|\hat{\beta}_j^0\right|$ (Tibshirani, 1996), ideia do método lasso é baseado na proposta de (Breiman, 1993).

$$\sum_{i=1}^{N}\left(y_i - \alpha - \sum_j c_j \hat{\beta}_j^0 x_{ij}\right)^2, \qquad 2.37$$

Condicionado a $c_j \leq 0$ e o $\sum c_j \geq t$.

A regressão lasso é uma abordagem mais recente comparativamente a ridge (James et al., 2013).

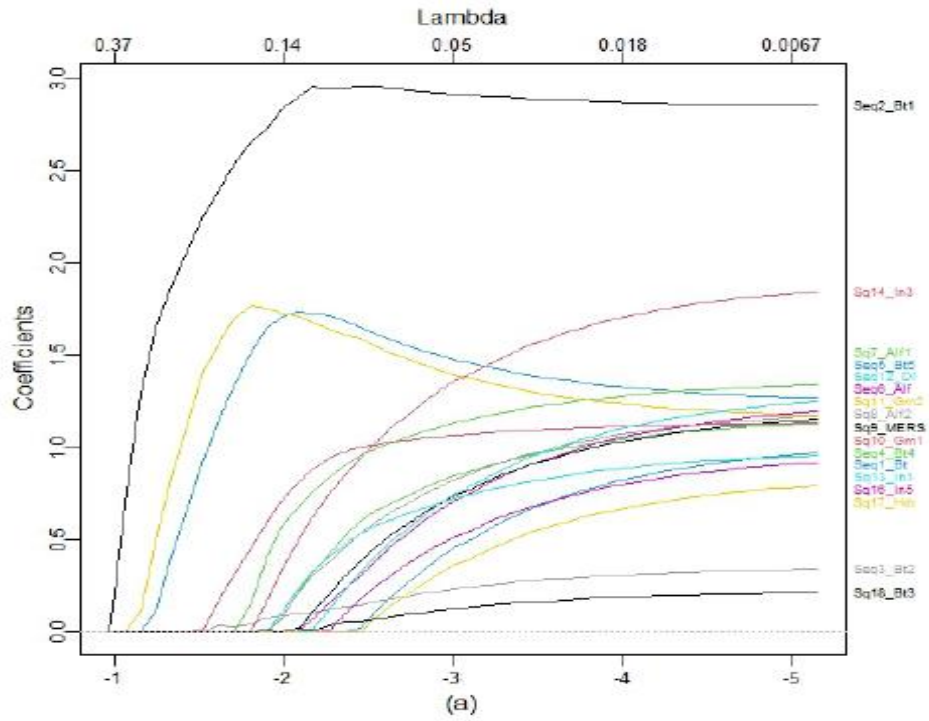
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right| = RSS + \lambda \sum_{j=1}^{p}\left|\beta_j\right| . \qquad 2.38$$

O lasso e uma penalização L_1, ou seja, a primeira norma de um vetor de coeficientes β e dado por $\beta_1 = \left|\beta_j\right|$. Pode-se observar que a equação 2.39 e similar a equação 2.36 da regressão ridge, com o diferencial de que a regressão ridge tem o termo de penalização dado por β_j^2 , enquanto que a regressão lasso tem o termo de penalização dado por $\left|\beta_j\right|$ (James et al., 2013). A regressão lasso faz a seleção de variáveis (Tibshirani, 1996), além disso, os agrupamentos feitos por meio da aplicação do método lasso apresentam mais facilidade de interpretação comparativamente ao método ridge.

Ademais, a regressão ridge apresenta uma melhor performance em relação a regressão lasso, contudo, a lasso e promissora pelo fato de fazer a seleção de variáveis diante de uma base de dados muito grande (Tibshirani, 1996).

Aplicou-se o método lasso a base de dados em estudo, e o resultado encontra-se na Figura 11.

Figura 11- Aplicação do método Lasso na base de dados em estudo



Fonte: A autora (2023)

Pode-se observar pela Figura 11, que o método lasso pode ser útil para a formação de grupos.

2.7.3 Elastic Net

Para uma abordagem mais clara relacionada com os métodos de regressão lasso e ridge, (James et al., 2013) exemplifica tomando em consideração um caso especial em que $n = p$, e X seja uma matriz com 1 na sua diagonal e os outros elementos da matriz compostos por zeros.

Com base nestas suposições os mínimos quadrados são simplificados para encontrar $\beta_1, \beta_2, \dots, \beta_p$ que minimiza:

$$\sum_{j=1}^{p} (y_i - \beta_j)^2. \qquad 2.39$$

Para este caso, teremos a solução dos mínimos quadrados dado por $\beta_j = y_i$. E considerando a regressão ridge, neste caso equivaleria a encontrar $\beta_1, \beta_2, \dots, \beta_p$ de modo que:

$$\sum_{j=1}^{p} (y_i - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \qquad 2.40$$

É minimizado, e o lasso equivale a encontrar os coeficientes tais que:

$$\sum_{j=1}^{p}\left(y_i - \beta_j\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j|, \qquad 2.41$$

e minimizado. Três limitações relacionadas ao método lasso são apontadas por (Zou et al., 2005) nomeadamente:

a) quando o número de variáveis p e maior em relação ao número de observações n, o método lasso seleciona no máximo n variáveis antes de saturar. E esta característica e limitante para um método que faz seleção de variáveis;

b) caso existam grupos de variáveis com correlações entre pares muito altas, o método lasso seleciona apenas uma variável do grupo independentemente de qual seja a variável;

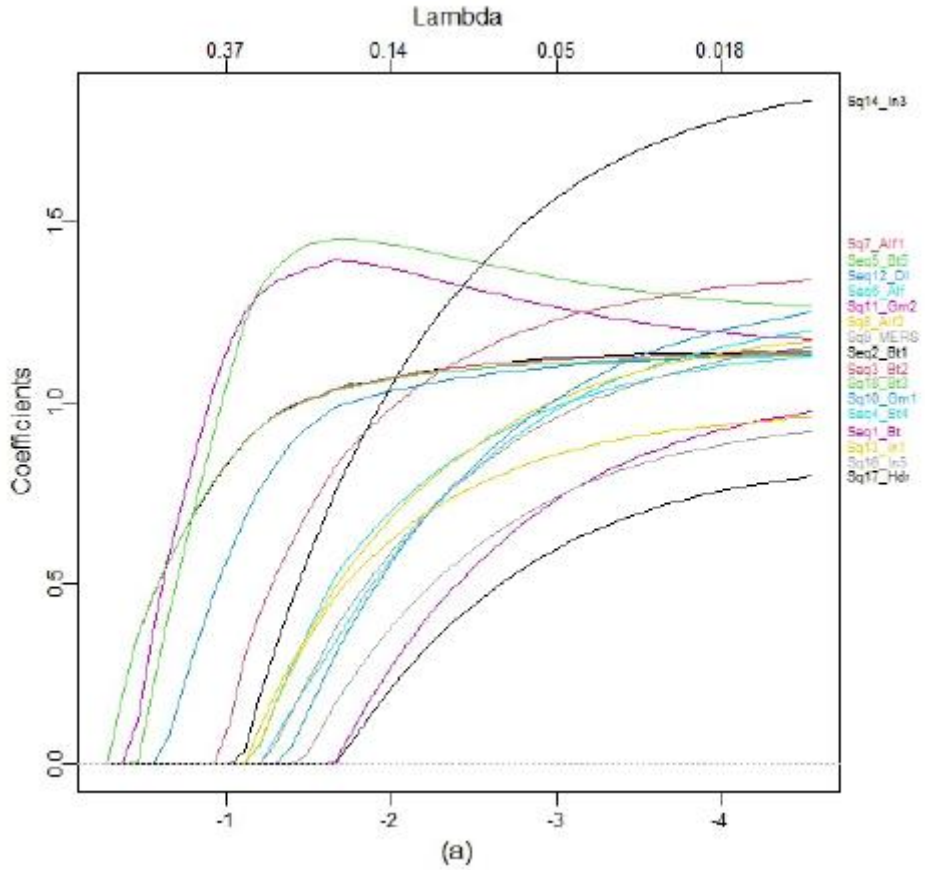
c) nos casos em que n > p, se existir altas correlações entre os preditores, a performance de predição da regressão ridge e melhor do que a performance da regressão lasso (Tibshirani, 1996).

No entanto, de forma a ultrapassar as limitações dos métodos de regressão lasso e ridge, o método elastic net junta o melhor dos dois métodos e inclui um novo hiperparâmetro, o α selecionado juntamente com o λ de modo a otimizar a performance do modelo. E tomando em consideração o exemplo anterior teremos:

$$\hat{\beta} = \sum_{j=1}^{p}\left(y_i - \beta_j\right)^2 + \lambda\left[(1-\alpha)\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|\right]. \qquad 2.42$$

Aplicou-se o método Elastic net na base de dados em estudo, e o resultado pode-se observar na Figura 12. Observa-se que o método elastic net e aplicável para a formação de grupos.

Figura 12- Aplicação do método Elastic net na base de dados em estudo



Fonte: A autora (2023)

REFERÊNCIAS

ALBERTS, B. et al. **Biologia molecular da celula**. [S.l.]: Artmed Editora, 2017.

ANDERSEN, A.; GODOY, E. **Infodemia em tempos de pandemia: batalhas invisiveis com baixas imensuraveis.** Revista Memorare, v. 7, n. 2, p. 184–198, 2020.

AZEVEDO, A. M. et al. **Divergencia genetica e importancia de caracteres morfologicos em genotipos de couve**. Horticultura Brasileira, SciELO Brasil, v. 32, p. 48–54, 2014.

BIAZON, V.; BIANCHI, R. **Gated recurrent unit networks and discrete wavelet transforms applied to forecasting and trading in the stock market**. In: Anais do XVII Encontro Nacional de Inteligencia Artificial e Computacional. Porto Alegre, RS, Brasil: SBC, 2020. p. 650–661. ISSN 2763-9061. Disponivel em: <https://sol.sbc.org.br/index.php/eniac/article/view/12167>.

BRASSAROTE, G. d. O. N. **Analise multiescala de series temporais do efeito da cintilação ionosférica nos sinais de satélite gps a partir de wavelets não decimadas**. Universidade Estadual Paulista (Unesp), 2014.

BREIMAN, L. **Hinging hyperplanes for regression, classification, and function approximation**. IEEE Transactions on Information Theory, IEEE, v. 39, n. 3, p. 999–1013, 1993.

CARDOSO-BERENSZTEJN, A. et al. **Aspectos ecocardiográficos em pacientes com polineuropatia amiloidotica familiar com a mutação val30met: Padrão da função diastólica**. Revista Brasileira de Neurologia e Psiquiatria, v. 18, n. 1, 2014.

DAS, S.; KUMAR, A. **Long-term dependency between sovereign bonds and sectoral indices of india: Evidence using hurst exponent and wavelet analysis**. Managerial Finance, Emerald Publishing Limited, v. 47, n. 10, p. 1448–1464, 2021.

DAUBECHIES, I. **Ten lectures on wavelets**. [S.l.]: SIAM, 1992.

DENAULT, W. R. et al. **A fast wavelet-based functional association analysis replicates several susceptibility loci for birth weight in a norwegian population**. BMC genomics, BioMed Central, v. 22, n. 1, p. 1–9, 2021.

FERNANDO, F. **Introdução a análise de agrupamentos**. [S.l.]: Unesp, 2006.

FERREIRA, D. F. **Estatistica multivariada**. [S.l.]: Editora Ufla Lavras, 2008.

FERREIRA, L.; LIMA, R. **Evaluation of genome similarities using the non-decimated wavelet transform**. Genetics and Molecular Research, Genetics and Molecular Research, v. 16, n. 3, 2017.

FERREIRA, L. M.; SAFADI, T.; FERREIRA, J. L. **Evaluation of genome similarities using a wavelet-domain approach**. Revista da Sociedade Brasileira de Medicina Tropical, SciELO Brasil, v. 53, 2020.

GALVAO, R. K. H. et al. **Estudo comparativo sobre filtragem de sinais instrumentais usando transformadas de fourier e wavelet**. Quimica Nova, SciELO Brasil, v. 24, p. 874–884, 2001.

GENCAY, R.; SELCUK, F.; WHITCHER, B. **Discrete wavelet transforms. An introduction to wavelets and other filtering methods in Finance and economics**, Elsevier Amsterdam, p. 96–160, 2002.

GRIFFITHS, A. J. et al. **Introdução a genética. In: Introdução a genética.** [S.l.: s.n.], 2006. p. 743–743.

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

JAMES, G. et al. **An Introduction to Statistical Learning with Applications in R**. [S.l.]: Springer, 2013.

KAMARTHI, S.; KUMARA, S.; COHEN, P. **Flank wear estimation in turning through wavelet representation of acoustic emission signals**. J. Manuf. Sci. Eng., v. 122, n. 1, p. 12–19, 2000.

LI GUIGUI, G. Y. et al. **Exploring the medication pattern of chinese medicine for peptic ulcer based on data mining**. Journal of Healthcare Engineering, Hindawi, v. 2021, 2021.

MAFTEI, C.; BARBULESCU, A.; CARSTEANU, A. A**. Long-range dependence in the time series of tai, ta river discharges.** Hydrological Sciences Journal, Taylor & Francis, v. 61, n. 9, p. 1740–1747, 2016.

MAGRINI, L. A.; DOMINGUES, M. O.; JUNIOR, O. **M. Analise tempo escala de séries temporais de geofısica espacial com lacunas: estudo de caso**. Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 5, n. 1, 2017.

MALLAT, S. G. A **Theory for multiresolution signal decomposition: the wavelet representation.** IEEE transactions on pattern analysis and machine intelligence, Ieee, v. 11, n. 7, p. 674–693, 1989.

MANLY, B. **Multivariate statistical methods: A primer. in: Carmona, sic (trad.). Metodos Estatisticos Multivariados: Uma introducao**. 3a ed. Porto Alegre: Bookman, 2008.

MANLY, B.; ALBERTO, J. **Multivariate statistical methods: a primer**. Florida. [S.l.]: CRC Press, 2016.

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

MEDAIYESE, O. O. et al. **Wavelet transform analytics for rf-based uav detection and identification system using machine learning**. Pervasive and Mobile Computing, Elsevier, v. 82, p. 101569, 2022.

MITTAL, R.; NI, R.; SEO, J.-H. **The flow physics of covid-19.** Journal of fluid Mechanics, Cambridge University Press, v. 894, p. F2, 2020.

MONTORIL, M. H.; MORETTIN, P. A.; CHIANN, C. **Wavelet estimation of functional coefficient regression models**. International Journal of Wavelets, Multiresolution and Information Processing, World Scientific, v. 16, n. 01, p. 1850004, 2018.

MORETTIN, P. A. **Ondas e Ondaletas** Vol. 23. [S.l.]: Edusp, 1999.

MORETTIN, P. A. **Ondas e ondaletas: da análise de fourier a análise de ondaletas de series temporais**. Edusp, 2014.

MORETTIN, P. A.; CHIANN, C.; MONTORIL, M. H. **Wavelet estimation of functional-coefficient regression models**. In: 29th International Workshop on Statistical Modelling. [S.l.: s.n.], 2014. p. 231.

NASON, G. P.; SILVERMAN, B. W. **The stationary wavelet transform and some statistical applications. In: Wavelets and statistics**. [S.l.]: Springer, 1995. p. 281–299.

NASTOS, C. V.; SARAVANOS, D. A. **Multiresolution daubechies finite wavelet domain method for transient dynamic wave analysis in elastic solids**. International Journal for Numerical Methods in Engineering, Wiley Online Library, v. 122, n. 23, p. 7078–7100, 2021.

NEGRI, T. T.; SOUZA, E. M. de. **Comparação das transformadas wavelet decimada e não-decimada para detecção do efeito do multi caminho de sinais gps e aplicacao no monitoramento de estruturas**. 2012

NETO, J. P. d. A. et al. **Analysis and comparison between regression models for temperature estimation of solar collectors operating with nanofuids**. http://www. abmec. org. br/congressos-e-outros-eventos, 2019.

NING, B. et al. **Liposome-mediated detection of sars-cov-2 rna-positive extracellular vesicles in plasma**. Nature nanotechnology, Nature Publishing Group UK London, v. 16, n. 9, p. 1039–1044, 2021.

PERCIVAL, D. B.; WALDEN, A. T. **Wavelet methods for time series analysis**. [S.l.]: Cambridge university press, v. 4. 2000.

PEREIRA, L. d. S. **Geometria dos métodos de regressão lars, lasso e elastic net com uma aplicação em seleção genômica**. Universidade Federal de Lavras, 2017.

PITTNER, S.; KAMARTHI, S. V. **Feature extraction from wavelet coefficients for pattern recognition tasks**. IEEE Transactions on pattern analysis and machine intelligence, IEEE, v. 21, n. 1, p. 83–88, 1999.

PLATTO, S. et al. **History of the covid-19 pandemic: Origin, explosion, worldwide spreading**. Biochemical and biophysical research communications, Elsevier, v. 538, p. 14–23, 2021.

RASHID, O.; AMIN, A.; LONE, M. R. **Performance analysis of dwt families.** In: IEEE. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). [S.l.], 2020. p. 1457–1463.

REIS, A. J. R.; SILVA, A. P. **Aplicação da transformada wavelet discreta na previsão de carga a curto prazo via redes neurais**. Sba: Controle & Automacao Sociedade Brasileira de Automatica, SciELO Brasil, v. 15, p. 101–108, 2004.

REIS, A. R.; SILVA, A. A. D. **Feature extraction via multiresolution analysis for short-term load forecasting**. IEEE Transactions on power systems, IEEE, v. 20, n. 1, p. 189–198, 2005.

ROUT, R. K. et al. **Feature-extraction and analysis based on spatial distribution of amino acids for sars-cov-2 protein sequences**. Computers in biology and medicine, Elsevier, v. 141, p. 105024, 2022.

SAINI, S.; DEWAN, L. **Application of discrete wavelet transform for analysis of genomic sequences of**. Mycobacterium tuberculosis, 2016.

SILVA, L. H. A. da et al. **Virus respiratorio sincicial humano e metapneumovirus humano**. Clinical and Biomedical Research, v. 29, n. 2, 2009.

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, Wiley Online Library, v. 18, n. 3, p. 279–304, 1997.

TIBSHIRANI, R**. Regression shrinkage and selection via the lasso.** Journal of the Royal Statistical Society Series B: Statistical Methodology, Oxford University Press, v. 58, n. 1, p. 267–288, 1996.

TIBSHIRANI, R. J.; HOEFLING, H.; TIBSHIRANI, R. **Nearly-isotonic regression. Technometrics**, Taylor & Francis, v. 53, n. 1, p. 54–61, 2011.

WOO, E. J. et al. Association of receipt of the ad26. cov2. s covid-19 **vaccine with presumptive guillain-barre syndrome**. Jama, American Medical Association, v. 326, n. 16, p. 1606–1613, 2021.

YANG, J.; DELPHA, C. **An incipient fault diagnosis methodology using local mahalanobis distance: Detection process based on empirical probability density estimation. Signal Processing,** Elsevier, v. 190, p. 108308, 2022.

ZAHA, A.; FERREIRA, H.; PASSAGLIA, L. **Biologia molecular basica**. 4ª edição. Editora Artmed, 2014.

ZOU, H.; HASTIE, T. et al. **Addendum: regularization and variable selection via the elastic net.** Journal-royal statistical society series b statistical methodology, Blackwell Publishing Ltd, v. 67, n. 5, p. 768, 2005.

SEGUNDA PARTE: ARTIGOS

ARTIGO 1 - REGRESSÃO PENALIZADA NO ESTUDO DE SIMILARIDADES DE GENOMAS DE VÍRUS DAS FAMÍLIAS CORONAVIRIDAE E PARAMYXOVIRIDAE – *Publicado na revista Contemporânea*

Contemporânea
Contemporary Journal
3(8): 12000-12017, 2023
ISSN: 2447-0961

Artigo

REGRESSÃO PENALIZADA NO ESTUDO DE SIMILARIDADES DE GENOMAS DE VÍRUS DAS FAMÍLIAS CORONAVIRIDAE E PARAMYXOVIRIDAE

PENALIZED REGRESSION IN THE STUDY OF GENOME SIMILARITIES OF VIRUSES OF THE FAMILIES CORONAVIRIDAE AND PARAMYXOVIRIDAE

DOI: 10.56083/RCV3N8-113
Recebimento do original: 17/07/2023
Aceitação para publicação: 17/08/2023

Dulcídia Carlos Guezimane Ernesto
Doutoranda em Estatística e Experimentação Agropecuária pela Universidade Federal de Lavras (UFLA)
Instituição: Universidade Federal de Lavras (UFLA)
Endereço: Avenida Bueno da Fonseca, 543, Inácio Valentim, Cidade de Lavras – MG, CEP: 37200-000
E-mail: dcg.ernesto@gmail.com

Leila Maria Ferreira
Pós-Doutora em Estatística e Experimentação Agropecuária pela Universidade Federal de Lavras (UFLA)
Instituição: Universidade Federal de Lavras (UFLA)
Endereço: Avenida Bueno da Fonseca, 543, Inácio Valentim, Cidade de Lavras – MG, CEP: 37200-000
E-mail: leilamaria2003@gmail.com

Thelma Sáfadi
Pós-Doutora em Ciências Exatas e da Terra pela Georgia Institute of Technology (Georgia TECH), Pós-Doutora em Inferência Bayesiana pela Universidade de São Paulo (USP), Pós-Doutora em Séries Temporais pela Universidade Carlos III de Madrid (UC3M)
Instituição: Universidade Federal de Lavras (UFLA)
Endereço: Avenida Bueno da Fonseca, 543, Inácio Valentim, Cidade de Lavras – MG, CEP: 37200-000
E-mail: safadi@ufla.br

RESUMO: Este trabalho teve por objetivo procurar similaridades entre alguns sequenciamentos das famílias Paramyxoviridae e Coronaviridae, com recurso ao método de regressão lasso e ridge sob o domínio da transformada de wavelet discreta não decimada de Daubechies com 4 momentos nulos. A transformada discreta não decimada de Daubechies foi implementada de

modo a se decompor o conteúdo GC em seis níveis de decomposição, com uma janela deslizante de comprimento n = 100. Conteúdo GC é a proporção de guanina e citosina presentes no genoma de um indivíduo, e por meio do conteúdo GC é possível ter a ancestralidade de um sequenciamento, incluindo informação sobre todos os organismos que são evolutivamente semelhantes a um determinado organismo. As wavelets permitiram que se pudesse decompor o conteúdo GC de cada sequenciamento, e consequentemente obteve-se a distribuição do conteúdo GC aumentando o nível de detalhamento e mostrando detalhes omissos do sinal. A inclusão dos métodos lasso e ridge foi feita com o intuito de se formar agrupamentos, consoante a similaridade dos sequenciamentos em estudo. Ao fim da pesquisa, observou-se que o método lasso teve melhor performance na formação dos grupos.

PALAVRAS-CHAVE: Regressão Lasso, Regressão Ridge, Transformada de Wavelet, Análise de Similaridade.

ABSTRACT: This work aimed to search for similarities between some sequences of the Paramyxoviridae and Coronaviridae families, using the lasso and ridge regression method under the domain of the discrete non-decimated Daubechies wavelet transform with 4 null moments. The nondecimated discrete Daubechies transform was implemented in order to decompose the GC content into six decomposition levels with a sliding window of length n = 100. GC content is the proportion of guanine and cytosine present in an individual's genome, and by means of GC content it is possible to have the ancestry of a sequence, including information about all organisms that are evolutionarily similar to a given organism. The wavelets allowed the GC content of each sequence to be decomposed, and consequently the distribution of the GC content was obtained, increasing the level of detail and showing details missing from the signal. The inclusion of the lasso and ridge methods was done in order to form clusters, depending on the similarity of the sequences under study. At the end of the research, it was observed that the lasso method performed better in forming the clusters.

KEYWORDS: Lasso Regression, Ridge Regression, Wavelet Transform, Similarity Analysis.

1. Introdução

Em pesquisas atuais é indispensável a implementação de métodos estatísticos no processamento de dados, de modo a assegurar a credibilidade das mesmas. A regressão penalizada tem sido um recurso bastante usado na modelagem de modelos que melhor possam se adequar aos dados em diversas áreas de conhecimento pois, num cenário atual em que o pesquisador está sujeito a trabalhar com um número elevado de variáveis preditoras que possam explicar uma variável resposta existe uma certa necessidade de se escolher as melhores variáveis preditoras. O mesmo se aplica na análise de similaridades entre duas sequências genômicas. O método Ridge é o método com um estimador de encolhimento, e a vantagem é que permite uma análise gráfica bastante útil do processo de estimação (Nelo et al., 2019). Por outro lado, o método Lasso é um método que faz seleção de variáveis preditoras, ou seja, o lasso minimiza a soma de quadrados com uma restrição (Tibshrani, 2011). Vários foram os pesquisadores que usaram os métodos de regressão penalizada Lasso ou Ridge. Nelo et al. (2019), aplicaram o método de regressão Lasso, Ridge e regressão Polinomial na previsão de temperatura do fluido e ganho de energia de um sistema solar em operações com nanofluídos, ao fim do estudo a regressão Ridge não apresentou o melhor desempenho pois o Ridge considerou um modelo com todos os preditores e obteve um alto erro de teste e o lasso excluiu alguns preditores e obteve um resultado melhor que o Ridge.

Dougho et al. (2023) desenvolveu uma pesquisa para modelo preditivo de aprendizado de máquina para triagem de aspiração em pacientes hospitalizados com AVC agudo. Neste estudo foram aplicadas as regressões Ridge, Lasso, Elastic net e mais algumas outras metodologias. Ao fim do estudo, o modelo de regressão Ridge foi o modelo com melhor performance, entre todos os modelos.

Alamro et al. (2023), exploraram modelos de aprendizado de máquina para identificar novos biomarcadores da doença de Alzaheimer e possíveis alvos, ou seja, projetaram um método computacional que explora vários métodos de classificação de genes hub e métodos de seleção de recursos no aprendizado de profundo para identificar biomarcadores e alvos. Ao fim do estudo puderam concluir que das metodologias aplicadas, os métodos de seleção de características baseadas nos algoritmos lasso e Ridge tiveram melhor performance.

Cygu et al. (2023) aplicaram métodos de previsão de tempo na comparação de abordagens de aprendizado de máquina para incorporar covariáveis que variam no tempo na previsão de tempo de sobrevivência do câncer. Aplicaram métodos baseado em aprendizado de máquinas como: modelo de aumento de gradiente (gbm), floresta de sobrevivência aleatório, Elastic net, regressão Lasso, regressão Ridge, e foram comparados com o modelo tradicional de riscos proporcionais de Cox. Ao fim da pesquisa, o modelo de aumento de gradiente teve melhor performance.

Neste trabalho, implementou-se a regressão penalizada por meio dos métodos Lasso e Ridge, sob o domínio das transformações Wavelet com o objetivo de procurar similaridades em alguns sequenciamentos genéticos das famílias Paramyxoviridae e Coronaviridae.

2. Material e Métodos

Fez-se a extração dos sequenciamentos genéticos no National Center for Biotechnology Information (NCBI). Como ilustra a tabela 1.

Tabela 1 – Lista de sequenciamentos em estudo.

N° de acesso	*Nome*	*Abreviatura*	*Família*
NC_001906	Henipavirus hendraense	Hendra	Paramyxovírus
NC_003461	Respirovírus humano 1	Infl1	Paramyxovírus
NC_001796	Respirovírus humano 3	Influ3	Paramyxovírus
NC_021928	Vírus da parainfluenza humana 4	Infl4	Paramyxovírus
NC_006430	Vírus parainfluenza5	Infl5	Paramyxovírus
NC_038294	Betacoronavírus Inglaterra 1	Bet1	Coronaviridae
NC_019843	Coronavírus relacionado à síndrome respiratória do Oriente Médio	MERS	Coronaviridae
NC_045512	Síndrome respiratória aguda grave coronavírus 2	Bet3	Coronaviridae
NC_003045	Coronavírus bovino	Bet4	Coronaviridae
NC_006213	Coronavírus humano OC43	Bet2	Coronaviridae
NC_004718	BtRf-BetaCoV/HeB2013	Bet5	Coronaviridae
NC_028752	Alfacoronavírus de camelo	Alf1	Coronaviridae
NC_002645	Coronavírus humano 229E	Alf2	Coronaviridae
NC_048214	Pato Coronavírus	Gama1	Coronaviridae
NC_010800	Turquia Coronavírus	Gama2	Coronaviridae
NC_005831	Coronavírus humano NL63	Alfa	Coronaviridae
NC_006577	Coronavírus humano HKU1	Beta	Coronaviridae
KJ481931.1	Deltacoronavirus	Del	Coronaviridae

Fonte: Elaborado pelas autoras.

Foram analisadas 18 sequências das quais 5 da família Paramyxoviridae e 13 da família Coronaviridae, de seguida 3 fases foram implementadas de modo a se obter o resultado, nomeadamente:

1. Calculou-se a proporção média de Guanina e citosina (conteúdo GC) em cada uma das sequências em estudo;
2. Aplicou-se a transformada de wavelet discreta não decimada da família Daubechie de modo a decompor cada uma das sequências e se obter finalmente a distribuição do conteúdo GC de cada sequência;
3. Aplicou-se o método de regressão penalizada Lasso e o método de regressão penalizada Ridge para a formação de grupos.

Segue a descrição de cada um dos procedimentos

2.1 Conteúdo GC

Na literatura, o conteúdo GC pode ser representado como um valor percentual ou proporcional, e é expresso como a razão que é apresentada na equação 1. Portanto, três pontes de hidrogênio ligam a Citosina a

Guanina, comparativamente às duas ligações que conectam a Adenina á Timina, ademais, organismos geneticamente semelhantes apresentam a mesma distribuição do conteúdo GC (Zaha, 2014). Segundo Saini e Dewan (2016) o conteúdo GC é dado por:

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

Após o cálculo do conteúdo GC, aplicou-se as transformações de wavelet de modo a decompor o conteúdo GC em seis níveis.

2.2 Algoritmo Piramidal de Mallat e as Transformações Wavelet

O algoritmo piramidal de Mallat foi proposto para calcular a transformada discreta de Wavelet. Geralmente as Wavelets são designadas por Ψ e obedecem a duas condições básicas (Morettin, 1999):

1. Condição de admissibilidade*;*

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

2. A função Wavelet deve ter energia unitária;

$$\int_{-\infty}^{+\infty} |\psi(t)|^2 dt = 1.$$

Matematicamente $\psi \in L^2(R)$, e é chamada de Wavelet mãe pelo fato dela poder gerar outras wavelets seja por compressão ou por translação da Wavelet mãe, e é dada por (Morettin, 1999):

$$\psi_{(a,b)}(t) = \psi\frac{(t-b)}{a}\frac{1}{\sqrt{a}},$$

Onde:

a é escala e b é parâmetro de tempo.

E satisfazem outro critério de admissibilidade:

$$C_\psi = \int_{-\infty}^{+\infty} |\psi(t)|^2 dt \ < +\infty.$$

O algoritmo de Mallat é explicado no trabalho de Li et al. (1997), em que detalha o procedimento de decomposição da transformada de Wavelet discreta. A transformada de Wavelet discreta é uma Wavelet natural em que os parâmetros de tempo e de escala são discretos. Ao tomarmos uma determinada sequência de tempo discreto dado por f(n), a transformada discreta de Wavelet é definida pela decomposição multiresolução em tempo discreto que pode ser calculada pelo algoritmo piramidal de decomposição de Mallat:

$$V_n^0 = f(n), n \in N$$
$$V_n^j = \sqrt{2} \sum_{k \in Z} h(k-2n) V_k^{j-1}, j = 1,2,\dots,L$$
$$W_n^j = \sqrt{2} \sum_{k \in Z} g(k-2n) V_k^{j-1}, \qquad j = 1,2,\dots,L$$

Onde:

h e g são respostas de impulso do filtro passa-baixa H e do filtro passa-alta G, respectivamente.

Por outro lado, a decomposição do sinal é feita com base no teorema que é dado por:

Teorema 1: Se o comprimento da escala de sequência V_n^{j-1} é N, e a resposta ao impulso h e g têm r e s amostras diferentes de zero, respectivamente, j = 1,2,...,L, então o comprimento de V_n^j e W_n^j são $\left(\frac{N+1}{2}\right)+\left(\frac{r-1}{2}\right)$ e $\left(\frac{N+1}{2}\right)+\left(\frac{s-1}{2}\right)$ respectivamente. A prova deste teorema pode ser encontrada no trabalho de Li et al. (1997)

Ademais, segundo Humbe et al. (2009), o processo de decomposição de sinais por meio da transformada de Wavelets, é possível devido a esses filtros passa- baixa e passa-alta. Na perspectiva de Tibuleac et al. (2003), a principal função das Wavelets é de decompor e localizar um sinal no domínio do tempo e de frequência, e a vantagem de se aplicar a técnica de transformada discreta de Wavelet em pesquisas, é a capacidade de extração de características e detecção de distúrbios que podem existir em sinais que não são estacionários. No entanto, neste trabalho usou-se a transformada discreta de Wavelet não decimada, pois esta garante que durante a decomposição não haja perda de informação. Segundo Tsai et al. (2013), no algoritmo da transformada de Wavelet discreta não decimada, os seus coeficientes de detalhe (alta frequência, passa-alta) e os coeficientes de aproximação (baixa frequência, passa baixa), em cada um dos níveis, têm o mesmo comprimento do sinal original. Aplicar a transformada discreta não decimada nos permitiu ter uma decomposição muito simples, e um alto nível de detalhamento na observação dos níveis pois, o algoritmo aplica a transformada em cada ponto da imagem e conserva os coeficientes de detalhamento, ao mesmo tempo que faz uso dos coeficientes de aproximação para o nível a seguir, e este processo de decomposição se repete conforme os níveis de interesse para o estudo.

Ao longo deste estudo aplicou-se a transformada de Wavelet discreta não decimada de Daubechies com 4 momentos nulos para decompor os

sequenciamentos em estudo. A formação de clusters foi exequível com base na aplicação do método de regressão penalizada lasso.

2.3 Regressão Penalizada

Os modelos de regressão penalizada aplicam uma penalização sobre os parâmetros de modo a regularizar problemas de *Overfiting* e de *Underfitting* em modelos de regressão linear. Portanto, a regressão de Ridge segundo Friedman, Hastie, Tibshirani (2010), é conhecida pelo fato de encolher os coeficientes de preditores correlacionados entre si, e por consequência, permite que eles peguem emprestado forçosamente um do outro e, ela inclui todos os preditores no seu modelo final. E por outro lado a regressão Lasso penaliza os parâmetros, seleciona as variáveis e zera estimativas de alguns dos seus parâmetros (Friedman, Hastie, Tibshirani, 2010).

2.4 Regressão Ridge

Este método de regressão é descrito no trabalho de Hoerl e Kennard (1970), em que se considera a regressão Ridge como a regressão de encolhimento e exerce um papel crucial pois na análise de dados possibilita a resolução do problema de multicolinearidade (Roozbeh e Mahdi, 2015). Ademais, ainda no trabalho de Hoerl e Kennard (1970) definiu-se o estimador de Ridge através de uma penalização no método de mínimos quadrados. Portanto, a regressão Ridge tem sido designada em outras literaturas por penalização l_2, pelo fato de penalizar os coeficientes quadráticos e tornar a soma dos erros quadráticos muito pequena quando procura as estimativas dos coeficientes que melhor se ajustam aos dados. Ademais, na regressão Ridge a forma de encolhimento depende basicamente da correlação dos

preditores (Tibshrani, 1996). Portanto, a regressão Ridge pode ser designada por:

$$SEQ_{l_2} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_I\right)^2 + \lambda \sum_{j=1}^{p} \beta_j.^2$$

Na perspectiva de Hastie, Tibshirani e Friedman (2013), o termo $l_2 = \lambda \sum_{j=1}^{p} {\beta_j}^2$, é pequeno quando os preditores $\beta_1, \beta_2, \ldots, \beta_p$ apresentam-se muito próximos de zero, o que possibilita a redução das estimativas do β_j para zero.

Na equação (3), temos que $\lambda \geq 0$ é o parâmetro de penalização, regularização ou ajuste do modelo, e é calculado de forma separada pela validação cruzada (Hastie, Tibshirani e Friedman, 2013). E o p determina o número de variáveis explicativas.

Por outro lado, no método Ridge a penalização aplica-se nos $\beta_1, \beta_2, \ldots, \beta_p$, pois tem-se por finalidade a redução da associação estimada em cada uma das variáveis explicativas com a variável resposta (Friedman, Hastie e Tibshirani, 2010). Por outro lado, é importante realçar que o parâmetro do ajuste λ controla a força dos parâmetros na equação e tem capacidade de assumir valores maiores ou iguais a zero.

A regressão Ridge possibilita a vantagem de ter representação gráfica, que segundo (Hoerl e Kennard, 1970) é chamado de Ridge trace.

2.5 Regressão Lasso

Na perspectiva de Tibshrani (1996) o Lasso é uma metodologia que além de fazer a seleção de variáveis, também faz penalização dos parâmetros buscando aumentar a acurácia da previsão e tornando o modelo mais facilmente interpretável. Segundo Xei et al (2023), é o método de regressão penalizada que usa a penalidade l_1, para restringir os coeficientes

de regressão (λ) em direção a zero e obter uma solução esparça, ou por outra, é a metodologia que tem a capacidade de zerar algumas estimativas dos coeficientes, excluindo dessa forma, certas variáveis do modelo (Tibshirani, 1996).

Esta metodologia é descrita no artigo de Tibshirani(1996), em que inicialmente considera-se uma determinada base de dados, (x^i, y_i) com $i = 1, 2, \dots, N$, onde $x^i = (x_{i1}, \dots, x_{ip})^T$são as variáveis preditoras e y_i são as variáveis resposta. Poartanto, neste método é importante realçar que se assume as observações y_i como sendo condicionalmente independentes, dados os valores de x_{ij}. E por outro lado, assume-se também que os x_{ij} são padronizados de modo que $\sum_i x_{ij}/N = 0$, $\sum_i {x^2}_{ij}/N = 1$.

Deste modo, tem-se que $\beta = (\beta_1, \beta_2, \dots, \hat{\beta}_p)^T$, e a estimativa de Lasso $(\hat{\alpha}, \hat{\beta})$ é definida por:

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

Em que:

$t \geq 0$ é um parâmetro de ajustamento. Portanto, para todo t, a solução para α é $\hat{\alpha} = \bar{y}$. Pode-se assumir sem perda de generalidade que, $\bar{y} = 0$ e consequentemente omite-se o α.

Uma das diferenças existentes entre o método Lasso e o Ridge, é que o modelo de penalização Lasso dá a possibilidade de se fazer um aumento dos coeficientes, ao contrário do método Ridge que utiliza a soma dos valores absolutos.

Para a questão de análise de clusters o método Lasso mostra-se mais eficaz que o método Ridge pelo fato de apresentar gráfico com a formação de grupos mais clara, é possível pelo método de Lasso verificar as sequências que são similares, e o nível de importância dos grupos de similaridade (genomas mais importantes estão de esquerda para a direita), e o método

Lasso faz exclusão de variáveis, e pelo gráfico observa-se a exclusão do genoma de Coronavírus relacionado à síndrome respiratória do Oriente Médio (MERS).

3. Resultados e Discussão

Após o cálculo da proporção de guanina e citosina (conteúdo GC), obteve-se a tabela 2, na qual pode-se observar que o conteúdo GC varia de 0.30 a 0.50. Esta variação da proporção do conteúdo GC é explicada por Saini e Dewan(2016), pois geralmente o conteúdo GC varia de 0.25 a 0.75, o que significa que as sequências em estudo tem baixo teor de conteúdo GC e consequentemente são moléculas instável por ter a composição genética composta por uma fita de mRNA (RNA mensageiro).

Tabela 2 – Proporção média de Guanina e Citosina (conteúdo GC) em cada sequenciamento.

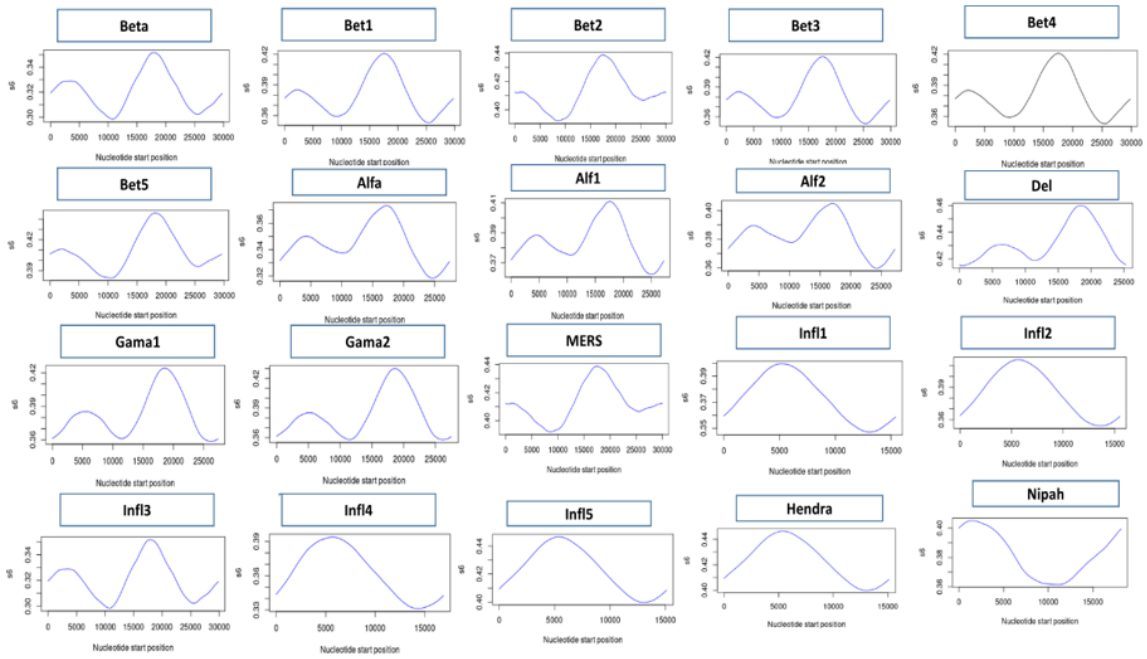
Nome do sequenciamento	**Conteúdo GC**
Henipavirus hendraense	0.4226682
Respirovírus humano 1	0.3724359
Respirovírus humano 3	0.3205908
Vírus da parainfluenza humana 4	0.3623035
Vírus parainfluenza5	0.4226682
Betacoronavírus Inglaterra 1	0.3797278
Coronavírus relacionado à síndrome respiratória do Oriente Médio	0.4123643
Síndrome respiratória aguda grave coronavírus 2	0.3797278
Coronavírus bovino	0.3797278
Coronavírus humano OC43	0.4123643
BtRf-BetaCoV/HeB2013	0.4076166
Alfacoronavírus de camelo	0.3841323
Coronavírus humano 229E	0.3826189
Pato Coronavírus	0.3822709
Turquia Coronavírus	0.382521
Coronavírus humano NL63	0.3446086
Coronavírus humano HKU1	0.3205908
Deltacoronavirus	0.4317878

Fonte: Elaborado pelas autoras.

Ademais, aplicou-se as transformações Wavelet de modo a decompor cada uma das sequencias em 6 níveis, e encontrar-se a real distribuição do conteúdo GC ao longo dos sequenciamentos como ilustra o gráfico 1. Pelo gráfico 1 observa-se que alguns sequenciamentos apresentam distribuição

do conteúdo GC muito similar. E mais adiante, o gráfico 3 irá mostrar quais os grupos que apresentam genomas evolutivamente similares.

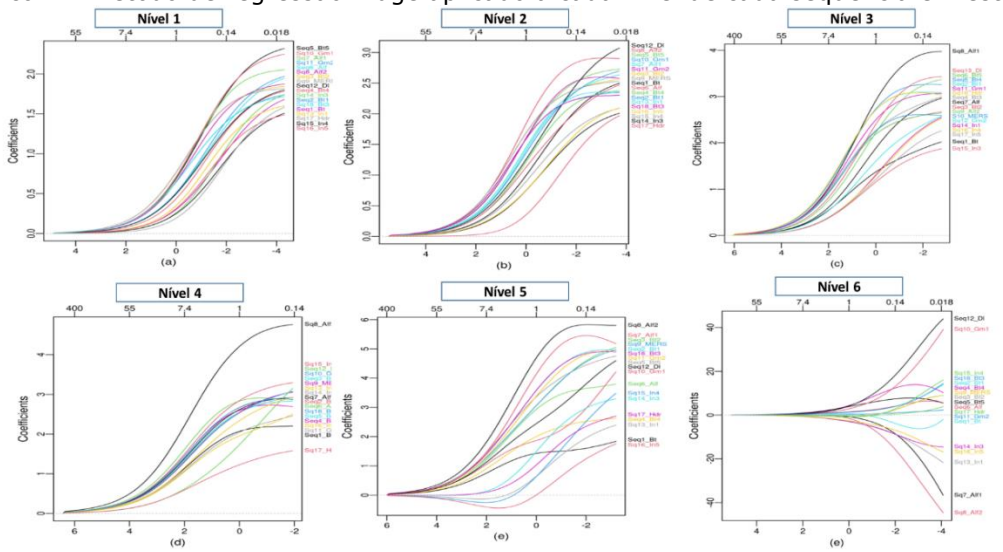
Figura 1 – Distribuição do conteúdo GC em cada sequenciamento.



Fonte: Elaborado pelas autoras.

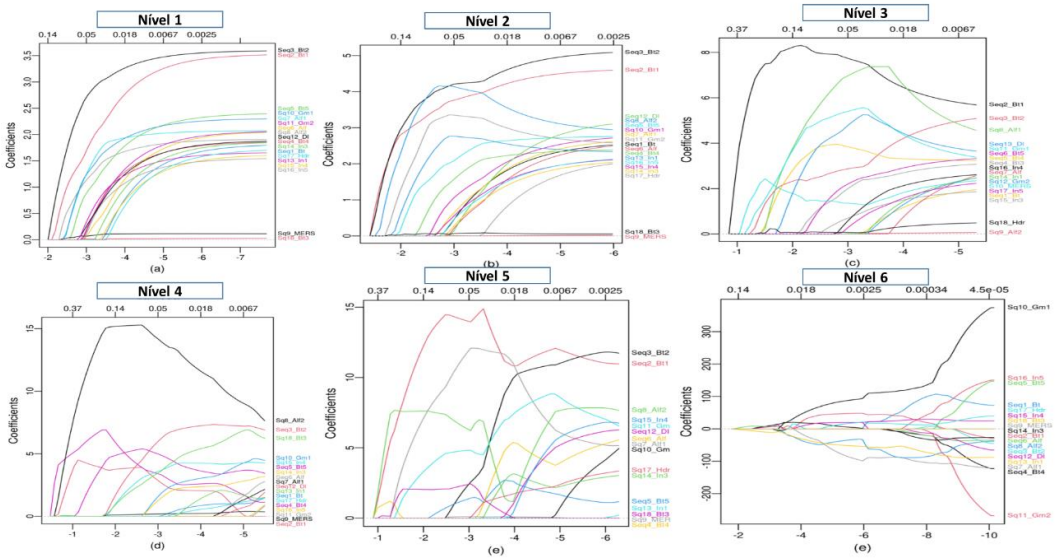
Após a decomposição das sequências pela transformada discreta não decimada de Daubechies, aplicou-se de forma separada os métodos Ridge e Lasso de modo a fazer análise de agrupamento como ilustram os gráficos 2 e 3 respectivamente.

Gráfico 2 – Método de regressão Ridge aplicado a cada nível de cada sequência em estudo.



Fonte: Elaborado pelas autoras.

Gráfico 3 – Método de regressão Lasso aplicado a cada nível de cada sequência em estudo.



Fonte: Elaborado pelas autoras.

Pode-se observar que aplicada a regressão penalizada, nos métodos Ridge e Lasso como ilustram os gráficos 2 e 3 respectivamente, o método Ridge não permitiu que fosse possível a formação de grupos de modo a se verificar agrupamentos similares, comparativamente ao método de

regressão Lasso, este apresentou uma melhor performance na formação de grupos e os agrupamentos formados estão apresentados na tabela 3.

Tabela 3 – Formação de grupos pelo método Lasso.

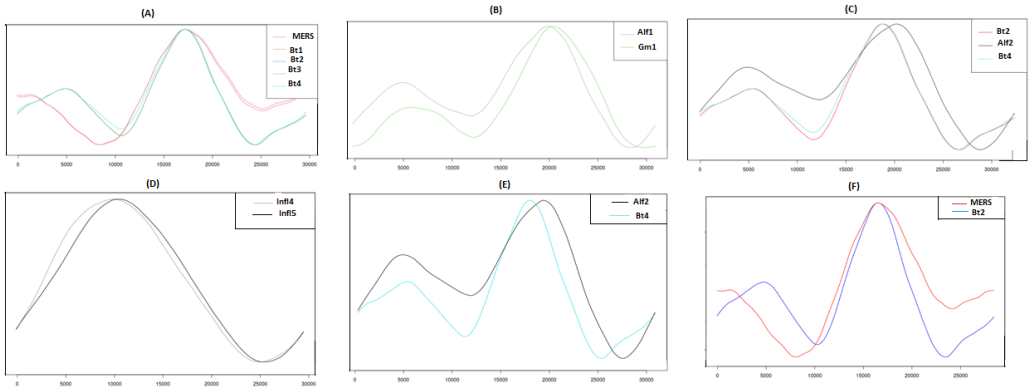
Nível	**Grupos formados**
1º Nível (a)	❖ MERS, Alf2 ❖ Gm1, Alf1 ❖ Gm2, Bt4, Infl1, Del, Bt2 ❖ Infl5, Alf
2º Nível (b)	❖ Bt2, Bt1 ❖ Bt3, Gm2 ❖ Infl4, Alf1 ❖ Inf5, Bt ❖ Alf, Del
3º Nível (c)	❖ Bt3, Bt5 ❖ Bt4, Alf1 ❖ Alf, Alf2
4º Nível (d)	❖ Bt1, Alf2 ❖ Bt2, MERS, Infl4, Bt3 ❖ Gm1, Infl3 ❖ Alf, Alf1 ❖ Infl5, Gm2
5º Nível (e)	❖ Alf2, Bt1, Alf, Bt3 ❖ Alf, Gm2 ❖ Bt2, MERS

Fonte: Elaborado pelas autoras.

É importante realçar que a leitura a ser feita para a formação de grupos por sequenciamentos similares, é a partir do momento em que as linhas saem da mesma origem no eixo das abcissas (eixo x), pertencem ao mesmo grupo. E linhas que saem de forma isolada, são cepas que foram isoladas pela metodologia. O método lasso apresentou melhor performance em relação ao método Ridge, pois no método Ridge não houve formação de grupos. No entanto, o Lasso tende a isolar muitas variáveis, e o último nível de decomposição não possibilita que se faça uma leitura nítida dos agrupamentos formados. Alguns padrões se repetem ao longo dos níveis, nomeadamente: Bet2 e MERS, Bt1 e Alf2, tendem a aparecer juntas nos níveis 4 e 5. Das cepas pertencentes a família Paramyxoviridae a única que não se juntou com as demais, ou seja, que se isolou em todos os níveis é a Henipavirus hendraense (Hendra). Um fator interessante que pode explicar esse isolamento do Henipavirus hendraense(Hendra) em relação as demais

cepas da família Paramyxovírus, é que segundo o site Virus taxónomy, existe uma característica que distingue o henipavírus dos demais vírus da família Paramyxoviridae, são os longos 3´-UTRs dos mRNAs, que o tornam um gene mais longo que os outros da mesma família. Portanto, o gráfico 4 vem mostrar claramente o quão próximo são algumas cepas.

Figura 4 – verificando o grau de aproximação de alguns sequenciamentos.



Fonte: Elaborado pelas autoras.

Pelo gráfico 4 observa-se que algumas distribuições do conteúdo GC apresentam um grau de similaridade tão grande, que praticamente se sobrepõe uma da outra, como é o caso de Bet2 e MERS, Bt4 e Bt3, Alf1 e Gm1, e por fim Infl4 e Infl5.

4. Conclusão

Durante o estudo pôde-se comprovar a flexibilidade e eficiência da transformada de Wavelet discreta não decimada de Daubechies, além do fato da regressão Lasso ter mostrado melhor performance que a regressão Ridge na análise de similaridades em sequenciamentos da família Coronaviridae e Paramyxoviridae. No entanto, na formação de grupos, houve uma evidência clara de que em nenhum momento o Hendra Henipavírus se agrupou com as demais cepas do Parainfluenza bem como do Coronaviridae.

Referências

NETO, A.; ROCHA, J. P; Costa, P. A.; et al. Analysis and comparison between regression models for temperature estimation of solar collectors operating with nanofuids. In: IBERO-LATIN-AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGINEERING, CILAMCE- ABMEC, XL., 11-14 nov. 2019, Natal/RN, Brazil. Proceedings […], Natal/RN, Brazil, 2019.

DOUGHO, P.; Il FILHO, S.; KIM, M. S.; KIM, T. Y.; CHOI, J. H.; LEE, S. E.; HONG, D.; KIM, M. C. Modelo preditivo de aprendizado de máquina para triagem de aspiração em pacientes hospitalizados com AVC agudo. *Sci Rep.* **13**, 7835 (2023). https://doi-org.ez26.periodicos.capes.gov.br/10.1038/s41598-023-34999-8.

ALAMRO, H.; THAFAR, M. A.; ALBARADEI, S.; GOJOBORI, T.; ESSACK, M. GAO, X. Explorando modelos de aprendizado de máquina para identificar novos biomarcadores da doença de Alzheimer e possíveis alvos. Scientific Reports. **13**, 4979 (2023). https://doi-org.ez26.periodicos.capes.gov.br/10.1038/s41598-023-30904-5.

CYGU, S.; SEOW, H.; DUSHOFF, J.; BOLKER, B. M. Comparando abordagens de aprendizado de máquina para incorporar covariáveis que variam no tempo na previsão do tempo de sobrevivência do câncer. Scientific Reports. **13**, 1370 (2023). https://doi-org.ez26.periodicos.capes.gov.br/10.1038/s41598-023-28393-7.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010; 33(1):1-22. PMID: 20808728; PMCID: PMC2929880.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. An Introduction to Statistical Learning: with applications in r. 1. ed. Nova Iorque: Springer, 2013.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, JSTOR, v. 58, p. 267-288, 1996.

HOERL, A. E.; KENNARD, R. W. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 12(1): 55-67. 1970. DOI: 10.1080/00401706.1970.10488634.

ROOZBEH, M.; ARASHI, M. New Ridge Regression Estimator in Semiparametric Regression Models.Communications In: Statistics - Simulation and Computation, [s.l.]. 2015. 45(10), 3683-3715. Informa UK Limited. http://dx.doi.org/10.1080/03610918.2014.953685.

SAINI, S., e DEWAN, L. Application of discrete wavelet transform for analysis of genomic sequences of Mycobacterium tuberculosis. *Springer Plus jornal. 2016. 5(64).* DOI 10.1186/s40064-016-1668-9.

TIBSHIRANI, R. Regression Shrinkage and Selection via The Lasso: A Retrospective, *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2011. **73**(3). 273–282. https://doi.org/10.1111/ j.1467-9868.2011.00771.x.

DOHERTY, T., DEMPSTER, E., HANNON, E. *et al.* Uma comparação de metodologias de seleção de recursos e algoritmos de aprendizado no desenvolvimento de um estimador de comprimento de telômero baseado em metilação de DNA. BMC Bioinformatics. 2023. **24**(178). https://doi-org.ez26.periodicos.capes.gov.br/10.1186/s12859-023-05282-4.

ZAHA, A. Biologia Molecular Básica. 5.ed. [S.l.]: Artmed, 2014.

MORETTIN, P. A., Waves and Wavelets: from Fourier analysis to wavelet analysis. EDUSP, 1999.

HUMBE, V.; GORNALE, S. S.; MAGAR, G.; MANZA, R.; KALE, K. V. Fingerprint Image De-noising through Decimated and Un-decimated Wavelet Transforms (WT). 2009 Conferência Internacional sobre Futuro Computador e Comunicação, Kuala Lumpar, Malásia, 2009, pp. 500-504, doi: 10.1109/ICFCC.2009.101.

LI, X.; LI, H.; WANG, F.; DING, F. A remark on the mallat pyramidal algorithm of wavelet analysis wavelet analysis, Communications in Nonlinear Science and Numerical Simulation. 1997. 2(4), 240-243. https://doi.org/10.1016/S1007-5704(97)90010-1.

TIBULEAC, I. M.; HERRIN, E. T.; BRITTON, J. M.; SHUMWAY, R.; ROSCA, A. C. Determinação Automática dos Tempos de Chegada da Fase Sísmi*ca* Secundária Utilizando Transformadas Wavelet. Cartas de Pesquisa Sismológica. 2003, 74 (6): 884–892. Doi: https://doi.org/10.1785/gssrl.74.6.884.

TSAI, D. Y.; MATSUYAMA, E.; CHEN, H. M. Melhorando a qualidade da imagem em imagens médicas usando um método combinado de transformação wavelet indecimada e mapeamento de coeficiente wavelet. International Journal of Biomedical Imaging. 2013. https://doi.org/10.1155/2013/797924.

XIE, Y.; SHI, H.; HAN, B. Bioinformatic analysis of underlying mechanisms of Kawasaki disease via Weighted Gene Correlation Network Analysis (WGCNA) and the Least Absolute Shrinkage and Selection Operator method (LASSO) regression model. BMC Pediatrics. 2023. 23(90). https://doi.org/10.1186/s12887-023-03896-4.

ARTIGO 2 - THE USE OF AN ELASTIC NET METHOD IN THE STUDY OF GENOME SIMILARITIES OF CORONAVIRIDAE AND PARAMYXOVIRIDAE VIRUSES– *Submetido na Biometrical Journal*

The use of an elastic net method in the study of genome similarities of Coronaviridae and Paramyxoviridae viruses

Abstract

In this work, we propose the implementation of an elastic net method with the aid of the nondecimated Daubechies discrete wavelet transform at 6 levels of decomposition and four null moments to analyze the formation of clusters in the families Coronaviridae and Paramyxoviridae. The nondecimated discrete wavelet transform allows signal decomposition and consequently a higher level of signal detail, revealing missed details throughout the decomposition. The formation of clusters was performed by including the elastic net method because it allows the formation of groups between correlated variables as it uses the best of the ridge method and the lasso method. This study aimed to analyze similarities accounting for the wavelet transform and the elastic net; therefore, the research involved three essential steps: the extraction of sequences from the NCBI database, the decomposition of each of the sequences including the GC content using the elastic net, and finally, the formation of clusters performed using the elastic net. In conclusion, in addition to the wavelet transform being very effective and accurate during the decomposition of the GC content, we revealed that human parainfluenza viruses 1, 3, 4 and 5 are similar to the less severe variants of the coronavirusy (OC43, 229E, NL63, HKU1).

Key words: Elastic net; GC content; Genetic sequencing; Wavelet transform;

1 Introduction

The first cases of the new coronavirus (COVID-19) occurred in the city of Wuhan, China, in December 2019 (Campus and Costa, 2020), and the patients initially presented symptoms similar to pneumonia, which evolved and led to death during a short period of time. In some cases, the coronavirus caused lung lesions involving the destruction of the lung parenchyma and extensive consolidation and interstitial inflammation (Campo and Costa, 2020). Eventually, it was found that variants of the new coronavirus developed that could cause different symptoms, which ranged from those of the common cold to the more severe symptoms associated with severe acute respiratory syndrome.

Coronaviridae and Paramyxoviridae are two families responsible for acute respiratory tract infections ranging from the mildest to the most severe (Silva et al., 2009 and Sales et al., 2020) and are often transmitted by contact with an infected individual through body fluids, excretions, secretions and salivary droplets (Sales et al., 2020); patients infected by viruses of these two families often present symptoms of a common cold, accompanied by fever, sore throat, headache and other symptoms depending on the type of virus.

In addition, it is believed that the study of similarities between genetic sequences may enable the discovery of vaccines for various diseases. The implementation of statistical methods that allow efficient and accurate data processing in the study of the similarities of genetic sequencing is essential to produce good quality results.

In addition, it is believed that the study of similarities between genetic sequences may enable the discovery of vaccines for various diseases. The implementation of statistical methods that allow

efficient and accurate data processing in the study of the similarities of genetic sequencing is essential to produce good quality results.

Bernatz et al. (2023) conducted a radiomics study of primary squamous cell carcinoma of head and neck(SCCHN) tumors to evaluate the performance of the treatment model and the importance of the developed resource in the treatment and prognosis of the disease; the main objective was to increase the interpretability of the models and classify the features based on their predictive importance. In this study, patients were stratified according to the type of treatment they received, cross-validation was performed with 100 iterations, the patients were classified, prognostic signatures were identified and intercorrelated using the elastic net and forest random survival methods, and the models were compared with clinical parameters to select the image resources with the highest rating.

Perez et al. (2023) investigated changes in fetal heart rate and neuraxial analgesia at birth based on a machine learning approach. In the study, they performed the comparison using a principal components regression model with tree-based random forest, ridge regression, multiple regression, a general additive model and elastic net regression; the precision was then evaluated with the square error medium. In this study, a retrospective analysis was performed on 1077 healthy pregnant women receiving neuraxial analgesia during labor. The tree-based random forest model ultimately showed better performance in data processing.

Malakouti (2023) developed a study to improve the performance of machine learning algorithms in predicting the speed and power of a SCADA system and compared the performance of the algorithms AdaBoost, Light Gradient Boosting Machine, gradient boosting regressor (GBR), lasso regression and Elastic Net. They concluded that the lasso regression and the elastic net method take less time to predict wind speed and energy generation.

Sonnweber et al. (2023) performed risk stratification based on supervised and unsupervised learning analyses of phenotyping in pulmonary arterial hypertension; seven parameters identified by elastic net modeling constituted a highly predictive mortality risk signature, and the elastic net signature demonstrated greater prognostic accuracy compared to five established risk scores. Furthermore, for multiparametric survival modeling, the elastic net model was trained using the IBK cohort, and the ideal lambda parameter was identified during the cross-validation with 200 replicates. Therefore, the Cox elastic net of proportional hazards and the distribution around the cluster were applied to establish a multiparameter PAH mortality risk signature and to investigate the phenotypes.

Lu et al. (2023) conducted a study in which stacking set models involving four basic learners, namely, ridge regression, random forest, gradient-increasing decision tree and artificial neural network, and elastic net were proposed to predict the daily number of hospital admissions using historical admission data, air quality data and weather data in Chengdu. The proposed stacking model that considered environmental exposure was more efficient in predicting daily hospital admissions and had practical value for early warning and the allocation of health resources. The model showed that environmental characteristics played an important role in improving the performance of the model, and

it was possible for the researchers to identify high temperatures and high concentrations of gaseous atmospheric polluting agents that could be closely associated with the risk of increasing the prevalence of cerebrovascular disease.

Doherty et al. (2023) developed a study to test and compare a variety of feature selection methods and algorithms in the development of a telemeter length estimator, used the cross-validation of two sets of tests independently for the comparison and found that principal component analysis before elastic net regression led to the best performing estimator when evaluated using a cross-validation analysis.

Ferreira et al. (2018) evaluated the similarity of the Mycobacterium tuberculosis genome by combining the nondecimated discrete wavelet transform and the elastic net, which was precise and capable of regrouping at each level of decomposition.

The present study aimed to use elastic net in the nondecimated Dubechie wavelet transform domain for the analysis of similarities between the sequences of genomes of the families Coronaviridae and Paramyxoviridae. Furthermore, the study was developed based on the previously cited study by Ferreira et al. (2018); the major difference is that the Mycobacterium tuberculosis genome was from a family of bacteria with more than 4 million nucleotides.

2 Materials and methods

Sequences from the complete genomes of the Coronaviridae and Paramyxoviridae families were extracted from the National Center for Biotechnology Information (NCBI) database, as shown in Table 1.

Table 1 Description of each genome and its origin

Sequence	Abbreviation	Family
Betacoronavírus of England I	Bt1	Coronaviridae
Middle East respiratory syndrome-related coronavirus	MERS	Coronaviridae
Coronavírus bovine	Bt4	Coronaviridae
Coronavírus human OC43	Bt2	Coronaviridae
SARS coronavirus Tor2	Bt5	Coronaviridae
Alfacoronavírus camel	Alf1	Coronaviridae
Coronavírus human 229E	Alf2	Coronaviridae
Pato coronavirus	Gm1	Coronaviridae
Turquia coronavirus	Gm2	Coronaviridae
Human coronavirus NL63	Alpha	Coronaviridae
Coronavírus humano HKU1	Bt	Coronaviridae
Severe acute respiratory syndrome coronavirus 2	Bt3	Coronaviridae
Deltacoronavirus	Del	Coronaviridae
Vírus da parainfluenza humana 1	Inf1	Paramyxoviridae
Human for influenza virus 3	Inf3	Paramyxoviridae
Vírus da parainfluenza humana 4	Inf4	Paramyxoviridae
Vírus da parainfluenza humana 5	Inf5	Paramyxoviridae
Hendra henipavirus	Hydra	Paramyxoviridae

The steps of the analysis are presented below:

1st Stage: After the extraction of the sequences, the GC content was analyzed using a sliding window of 100 base pairs.

The GC content was used to map the composition of the genome and understand the evolution of its coding sequence (Zaha, 2014). Therefore, it is calculated as the ratio of the sum of guanine (G) and cytosine (C) bases to the sum of adenine (A), guanine (G), cytosine (C) and threonine (T) bases (Ferreira et al., 2020). Determining the GC ratio helps in the identification of regions rich in genes in the genome because regions rich in GC content have many protein-coding genes (Summer et al. (1993), Aissani and Bernardi, 1991), and the GC content was calculated as follows:

$$GC_{content} = \frac{nG+nC}{nA+nG+nC+nT}, \quad (1)$$

in whichnA, nC, nG andnT are the number of nucleotides with the nitrogenous bases adenine, cytosine, guanine and thymine, respectively. According to Saini and Dewan (2016), the higher the proportion of GC bases, the greater the likelihood of the existence of many protein-coding genes in the genome. Therefore, GC content analysis is essential in the study of similarities because, according to Zaha (2014), genetically similar organisms tend to have similar GC content distributions throughout their genetic sequences.

2nd Stage: The nondecimated Daubechies discrete wavelet transform was applied for decomposition.

According to Bhuvaneswary et al. (2019), the transform of wavelets has the special function of decomposing a signal in terms of groups of basic functions that involve time and frequency. Therefore, wavelet theory defines a variety of basic functions, and for the case of orthogonal wavelets, signal decomposition is performed by the construction of mirror filters.

In this study, we chose to work with the Daubechies discrete wavelet transform with four null moments. Furthermore, from the perspective of Bhuvaneswary et al. (2019), Daubechie's wavelets are a family of orthogonal wavelets that process a wavelet transform separated and characterized by a certain amount of null moments for a certain support. In addition, when working with orthogonal wavelets, they define a wavelet remodeling approach that is fast and is used to modify the wavelet transform (Bhuvaneswary et al, 2019), and this remodeling is what allows studies to be conducted to assess sequence similarity in this context.

Moreover, the nondecimated wavelet transform is a process in which the decimated discrete transform is implemented, but with a small modification in the implementation, because the *downsampling* process is eliminated, i.e., in the nondecimated wavelet transform, it does not occur, resulting in a loss of information (Percival and Walden, 2000). According to Nason and Silverman (1995), at each level, the low-pass and high-pass filters are modified, and zeros are inserted.

From the perspective of Ferreira et al. (2020), if we take the wavelet scale functions given by ϕ *e* ψ,, we can represent the data vector $y = y_0, y_1, \dots, y_{m-1}$ of size *m* as a function *g* in terms of the displacement of the scale function at some multiresolution level *h* so that $j - 1 < log_2 m \leq j$

$$g(x) = \sum_{k=0}^{m-1} y_k \phi_{j,k}(x), \quad (2)$$

where $\phi_{h,k}(x) = 2^{\frac{j}{2}} \phi(2^j(x-k))$.

A decomposition was performed for each sequence in which there was a mother wavelet and a father wavelet, which had coefficients as follows:

$$g(x) = \sum_{k=0}^{m-1} c_{j_0} \phi_{j_0,k}(x) + \sum_{j=j_0}^{j-1} \sum_{k=0}^{2^{n-1}} d_{jk} 2^2 \psi\left(2^j(x-k)\right) \quad (3)$$

where

$\phi_{j_0,k}(x) = 2^{\frac{j_0}{2}} \phi\left(2^{j_0}(x-k)\right)$;

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

$j = j_0, \ldots, j-1; k = 0, 1, \ldots, m-1.$

The vector y of the nondecimal discrete wavelet transform is represented by the coefficients $c_{h_{0,}k}$ and $d_{jk}, j = j_0, \ldots, j-1; k = 0, \ldots, m-1.$

In addition, the nondecimated discrete wavelet transform is advantageous because it has the following properties (Farnazehdehkordi et al., 2022):

1. The transform enables the relationship between information to be determined at a given point in time at different scales; that is, the transformation makes it easy to locate important information at each of the levels for the same timepoint.
2. It does not affect the reconstruction of the original signal in the same way that the signal cannot be corrupted (nonaliased).
3. There is little chance of neglecting key features during the feature extraction step.
4. The outputs of this transform do not depend on the choice of signal origin because there is no sensitivity, no beginning and no distinct end.

After the implementation of Daubechie´s nondecimated discrete wavelet transform to decompose the signal into six levels, cluster analysis was performed using the elastic net method.

3rd Stage: After decomposition, the elastic net methodology was applied to each of the levels so that similar sequences formed clusters.

This methodology is described in detail in the work by Zou and Hastie (2005), where in principle it is considered that the elastic net is the combination of the penalties l_1 (lasso) and penalty l_2 (ridge). Therefore, it combines the advantages of the lasso and ridge methods. According to Wang et al. (2022), the elastic net solves the issue of the selection of grouped variables that are not known. The ridge regression estimator (with all predictors) is given by (Zou and Hastie, 2005):

$$\hat{\beta} = \arg\min\left\{\sum_{i=1}^{n} (Y_i - X\beta)^2 + \lambda \sum_j \beta_j^2\right\} \quad (4)$$

In turn, the lasso regression estimator with all its predictors is given by (Zou and Hastie, 2005):

$$\hat{\beta} = \arg\min\left\{\sum_{i=1}^{n} (Y_i - X\beta)^2 + \mu \sum_j |\beta_j|\right\} \quad (5)$$

Therefore, the elastic net estimator is the combination of lasso and ridge (Zou and Hastie, 2005):

$$\widehat{\beta} = \arg\min\left\{\sum_{i=1}^{n}(Y_i - X\beta)^2 + \lambda\left((1-\alpha)\sum_{j}^{n}|\beta_j| + \alpha\sum_{j}\beta_j^2\right)\right\} \quad (6)$$

3 Results

After the extraction of the complete genomes, the GC content was calculated for each of the sequences under study, as shown in Table 2.

Table 2 NCBI accession data and guanine and cytosine coefficients

NCBI Access Code	Sequence	GC Content
NC_038294	Bt1	0.3205908
NC_019843	MER	0.4117764
NC_006213	Bt2	0.36788
NC_003045	Bt4	0.36788
NC_004718	Bt5	0.4076166
NC_028752	Alf1	0.3446086
NC_002645	Alf2	0.3841323
NC_048214	Gm1	0.3826189
NC_010800	Gm2	0.4123643
NC_005831	Alpha	0.3927362
NC_006577	Bt	0.382521
NC_045512	Bt3	0.36788
KJ481931.1	Del	0.4317878
NC_003461.1	Infl1	0.3724359
NC_001796.2	Infl3	0.3452335
NC_021928.1	Infl4	0.3623035
NC_006430.1	Infl5	0.4226682
NC_001906.3	Hydra	0.394428

The proportions of GC content range from 0.30 to 0.45. Throughout the sequencing, the GC content has a highly variable distribution, thus exhibiting a mosaic with varying percentages; in some points, the content is higher, while in others, it has the lowest percentage of chromosome staining (Terrence and Haussler, 2003). Next, the GC content of each of the genetic sequences was decomposed into six levels of decomposition using the nondecimated Daubechie's discrete wavelet transform, as shown in Figure 1.

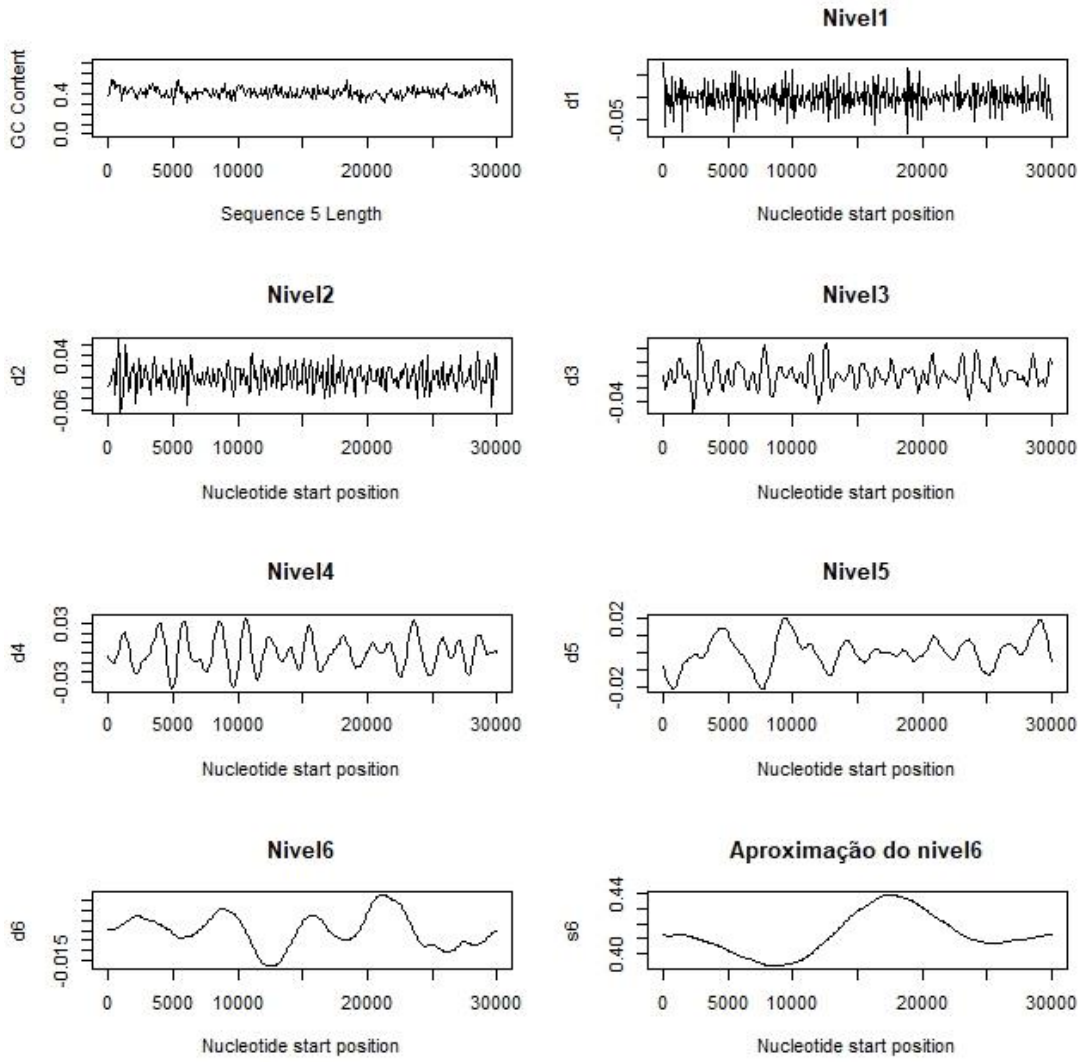


Figure 1 Decomposition of the GC content sequence for the MERS

Figure 1 exemplifies the decomposition of the GC content sequence for the Middle East respiratory syndrome-related coronavirus genome of the Coronaviridae family into six levels of decomposition. As shown in the last graph of Figure 1, the approximation of the last level of decomposition more comprehensively shows the distribution of the GC content of each of the sequences. The same technique was implemented for all sequences under study.

After the decomposition of each of the sequences, the elastic net method was applied to each of the levels of decomposition to analyze the formation of clusters of sequences that may be evolutionarily similar, as illustrated in Figure 2.

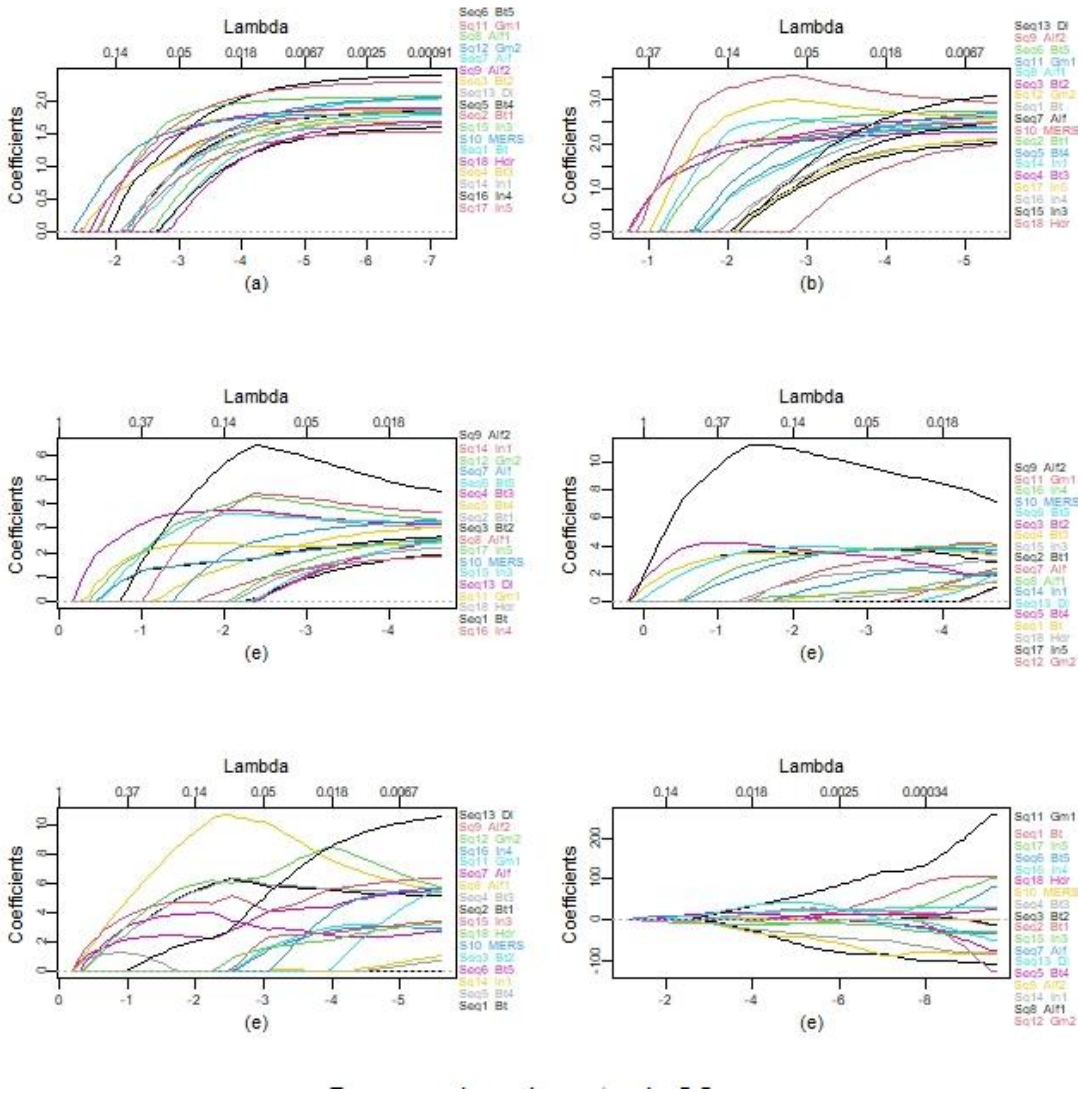


Figure 2 Formation of clusters using elastic net

Therefore, as shown in Figure 2, taking the sequences that apparently come together on the x-axis, Table 3 was prepared to reflect all the information linked in Figure 2. Table 3 summarizes all the groups formed in each of the levels, the isolated sequencing, and the excluded sequencing, after applying the elastic net method in each of the levels of decomposition of the GC content.

Table 3 Groupings formed along the levels

Level	Groups formed	Isolated strains	Excluded
Level 1	1. MERS, Bt2 2. Bt1, Bt3 3. Alf1, Gm1 4. Del, Bt4 5. Alf, Gm2, 6. Infl4, Bt	Alf2, Bet5, Infl5, Infl3, Hdr, infl1	None

Level 2	1. MERS, Bt3, Bt1, Bt2 2. Bt5, Influ1 3. Alf1, Gm1, Bt4 4. Bt, Alf, Infl5, Infl3, Del	Alf, Gm2, Hdr	Infl4
Level 3	1. HDR, Bt1 2. Bt2, Alf2, Bt4 3. Infl4, Infl5 4. Infl1, Bt,Alf,Gm2	MERS, Gm1, Alf1, Del, Bt3, Bt5, Infl3	None
Level 4	1. Alf2, Bt4, Bt3, Bt1 2. Bt2, MERS 3. Alf1, Gm1 4. Bt, Infl1 5. Gm2, Infl5	Bt5, Infl4, Infl3,Hdr, Alf, Del	None
Level 5	1. Alf2, Alf, Bt3, Bt1 2. MERS, Bt2 3. Infl4, Gm1 4. Infl1, Bt4	Del, Gm2, Bt5¸Hdr, Infl3, Alf1,Bt	Infl5

Table 3 shows that at each level, new groups formed, but there are strains that repeatedly grouped throughout the levels. However, it was also observed that in the first levels of decomposition, clusters that were clearer and easier to analyze formed. The final levels, namely, the 5th and 6th levels, were less delineated making analysis difficult.

4 Discussion

The common cold has been considered a mild disease regarding the symptoms, and according to Turner (1997) and Kirkpatrick (1996), the viruses that cause the common cold are most often the Parainfluenza virus, the influenza virus, the rhinovirus in 40% of cases, and the coronavirus in 10% to 20% of cases. Furthermore, according to Hasöksüz et al. (2020), seven variants of the coronavirus are responsible for infecting humans; however, four of them are considered to be the causes of the common cold in immunocompetent individuals, namely, human coronavirus OC43 (Bt2), human coronavirus 229E (Alf2), human coronavirus NL63 (Alf) and human coronavirus HKU1 (Bt).

Table 3 shows that there is a clear tendency for parainfluenza virus strains to associate with less severe coronavirus variants such as Alf2, Alf, Bt2 and Bt in the formation of groups along the levels, including (Inf4 and Bt) at the 1st level, (Bt, Alf, Inf5, Inf3) at the 2nd level, (Inf1, Bt, Alf) at the 3rd level, (Bt, Inf1) at the 4th level, and (Inf1 and Bt4) at the 5th level.

Table 3 also shows that the MERS variant is associated only with Betacoronavirus variants, namely, (MERS, Bt2) at the 1st and 5th levels and (MERS, Bt3 and Bt1) at the 2nd level. According to

Liang et al. (2020), the MERS variant is derived from Betacoronavirus and has a zoonotic origin, i.e., it has had hosts such as camels and other animals in recent decades.

Other coronavirus variants that were associated along the levels included the alphacoronavirus and gammacoronavirus variants, such as (Alf1, Gm1) at the 1st, 2nd and 4th levels and (Alf, Gm2) at the 1st and 3rd levels. This relationship of similarity between the alphacoronavirus and gammacoronavirus variants can be explained by the existence of similar mutations, including the N501Y mutation and the S106/G107/F108del mutation, which are targeted by vaccines developed by Oxford-AstraZeneca and Pfizer. According to Michelon (2021), to combat these two variants, it is necessary for the vaccine to have high serum concentrations to achieve neutralization.

Conclusion

In addition to the fact that the wavelet transform proved to be very effective and accurate in the decomposition of the GC content in each of the sequences under study, increasing the level of detail, the elastic net methodology implemented for the formation of clusters made it possible to conclude that strains of the Paramyxoviridae family in human parainfluenza viruses 1, 3, 4 and 5 are similar to the less severe variants of human coronavirus (OC43, 229E, NL63, HKU1), namely, Bt, Alf, Alf2 and Bt2, which all cause infections with symptoms of the common cold. Moreover, the elastic net applied to the studied sequences showed better performance at levels 1 to 5.

References

1. Aïssani, B. and Bernardi, G. (1991). CpG islands, genes and isochores in vertebrate genomes, Gene. **106**(2): 185 – 195. https://doi.org/10.1016/0378-1119(91)90198-K
2. 2. Al Sulaiman, K. et al. (2023). The clinical outcomes of critically ill patients with COVID-19 co-infected with other respiratory viruses: a multicenter cohort study. BMC Infect Dis. **23**, 75. https://doi-org.ez26.periodicos.capes.gov.br/10.1186/s12879-023-08010-8
3. Bernatz, S. et al.(2023). Radiomics for Therapy-Specific Head and Neck Squamous Cell Carcinoma Survival Prognosis (Part I). BMC Imaging Medical. **23**, 71. https://doi-org.ez26.periodicos.capes.gov.br/10.1186/s12880-023-01034-1.
4. Bhuvaneswary, N. et al. (2019). Design of Parallel Pipelined Architecture For Wavelet Based Image Compression Using Daubechie's. IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). Tamilnadu, India, **2019**, 1-5, doi: 10.1109/INCOS45849.2019.8951432.
5. Berensztejn, A. C. et al. (2014). Echocardiographic aspects in patients with familial amyloid polyneuropathy with the val30met mutation: pattern of diastolic function. Brazilian Journal of Neurology and Psychiatry. **18**(1):13-23.

6. Campos, N. G. and Da Costa, R. F. (2020). Lung changes caused by the new Coronavirus (COVID-19) and the use of invasive mechanical ventilation. Journal Health Biology Science. **8**(1):1-3. doi: 10.12662/2317-3076jhbs.v8i1.3185.p1-3.2020.
7. Da Silva, A. H. A. et al. (2009). Human respiratory syncytial virus and human metapneumovirus**.** Rev HCPA. **29**(2):139-146.
8. De Souza SGA, Da Silva SP, Da Costa MAS, et al. SARS-CoV, MERS-CoV and SARS-CoV-2 infections in pregnancy and fetal development. J Gynecol Obstet Hum Reprod. 2020 Jun 26;49(10):101846. Doi: 10.1016/j.jogoh.2020.101846.
9. Doherty, T. *et al.* Uma comparação de metodologias de seleção de recursos e algoritmos de aprendizado no desenvolvimento de um estimador de comprimento de telômero baseado em metilação de DNA. BMC Bioinformatics. 2023. 24(178). https://doi-org.ez26.periodicos.capes.gov.br/10.1186/s12859-023-05282-4.
10. Ernesto, D. C. G. et al. (2023). Penalized regression in the study of genome similarities of viruses from the coronaviridae and Paramyxoviridae families. Contemporary. **3**(8): 12000-12017. DOI: 10.56083/RCV3N8-113.
11. Farias, L. P. G. et al. (2020). Chest tomographic manifestations in symptomatic respiratory patients with COVID-19. Radiol Bras. **53**(4): 255.
12. Farzanehdehkordi, M. et al. (2022) Ghaffaripour S, Tirdad K, Dela Cruz A, Sadeghian A. A wavelet feature-based neural network approach to estimate electrical arc characteristics. Electric Power Systems Research. 208. https://doi.org/10.1016/j.epsr.2022.107893.
13. Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. **33**(1):1-22.
14. Hastie, T., Tibshirani, R., Friedman, J. (2013). An Introduction to Statistical Learning: with applications in r. 1ª ed. Nova Iorque: Springer.
15. Hoerl, A. E., Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 55-67.
16. Hasöksüz, M., Kiliç, S., Saraç, F. (2020). Coronaviruses e SARS-COV-2. Turk J Med Sci.
17. Kirkpatrick, G. L. (1996). The common cold. Prim Care**. 23**:657-75.
18. Liang, Y. et al. (2020). Highlight of Immune Pathogenic Response and Hematopathologic Effect in SARS-CoV, MERS-CoV, and SARS-Cov-2 Infection. Frontiers immunology.**11**. https://doi.org/10.3389/fimmu.2020.01022.
19. Lu, X. and Qiu, H. (2023). Explainable prediction of daily hospitalizations for cerebrovascular disease using stacked ensemble learning. BMC Medical Informatics and Decision Making. **23**: 59. https://doi-org.ez26.periodicos.capes.gov.br/10.1186/s12911-023-02159-7.
20. Malakouti, S. M. (2023). Improving wind speed prediction and power production of SCADA system with ensemble method and 10-fold cross-validation. Case studies in chemical and environmental engineering. **8**, 100351.

21. Medina, J. and Carrillo, I. (2018). Spectral and spatial evaluation of decimated and undecimated TDW wavelet transform for OrbView-2 satellite image fusion. 13th Iberian Conference on Information Systems and Technologies (CISTI*)*, Cáceres, Espanha. 1-6, doi: 10.23919/CISTI.2018.8399418.
22. Percival, D. B.(2000). Walden AT. Wavelet Methods for Time Series Analysis, Cram-bridge University Press.
23. Perez, E., Gutierrez, J. J., Molano, B. (2023). Fetal heart rate changes and neuraxial labor analgesia: a machine learning approach. BMC Gravidez Parto. **23**, 329. https://doi-org.ez26.periodicos.capes.gov.br/10.1186/s12884-023-05632-3.
24. Roozbeh, M. and Arashi, M. (2015) New Ridge Regression Estimator in Semiparametric Regression Models.Communications In Statistics - Simulation And Computation. **45**(10): 3683-3715. http://dx.doi.org/10.1080/03610918.2014.953685.
25. Saini, S. and Dewan, L. (2016). Application of discrete wavelet transform for analysis of genomic sequences of Mycobacterium tuberculosis. Springer Plus jornal. **5**(64). DOI 10.1186/s40064-016-1668-9.
26. Sales, E. M. P. (2020). Physiotherapy, functionality and covid-19: integrative review. Esp Notebooks. Ceará. **14**(1): 68 – 73.
27. Sonnweber, T. et al. (2023). The Combination of Supervised and Unsupervised Learning-Based Risk Stratification and Phenotyping in Pulmonary Arterial Hypertension - a Long-Term Retrospective Multicenter Study. BMC Pulm Med. **23**, 143. https://doi-org.ez26.periodicos.capes.gov.br/10.1186/s12890-023-02427-2.
28. Sumner, A. T. et al. (1993). The distribution of genes on chromosomes: a cytological approach. Journal of Molecular Evolution. **37**: 117-122. https://doi.org/10.1007/BF02407346.
29. Sun, Y. et al. (2023). Prediction of hot spots at protein-DNA binding interfaces based on discrete wavelet transform and wavelet packet transform. BMC Bioinformatics. **24**(129). https://doi-org.ez26.periodicos.capes.gov.br/10.1186/s12859-023-05263-7.
30. Terrence, S. F. and Haussler, D. (2003). Integration of the cytogenetic map with the draft human genome sequence. Human Molecular Genetics. **12**(9): 1037–1044. https://doi.org/10.1093/hmg/ddg113.
31. Tibshirani, R. (2011). Regression Shrinkage and Selection via The Lasso: A Retrospective, Journal of the Royal Statistical Society Series B: Statistical Methodology. **73**(3): 273–282. https://doi.org/10.1111/ j.1467-9868.2011.00771.x.
32. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. **58**: 267-288.
33. Turner, R. B. (1997). Epidemiology, pathogenesis, and treatment of the common cold. Ann Allergy Asthma Immunol**. 78**: 531-39.

34. Wang, W. et al. (2022). A Robust Variable Selection Method for Sparse Online Regression via Elastic Net Penalty. Mathematics. **10**(16):2985. https://doi.org/10.3390/math10162985
35. Zaha, A. (2014). Basic Molecular Biology. 5.ed. [S.l.]: Artmed
36. Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). **67**: 301–320.

ARTIGO 3 - ANALYSIS OF SIMILARITIES IN RESPIRATORY TRACT GENOMES: An approach using the Hurst exponent and the non-decimated discrete wavelet transform – *Artigo publicado na Revista Caderno Pedagógico*

Analysis of similarities in respiratory tract genomes: an approach using the Hurst exponent and the non-decimated discrete wavelet transform

Análise das semelhanças nos genomas do trato respiratório: uma abordagem usando o expoente de Hurst e a transformação discreta de wavelet não-dizimada

DOI: 10.54033/cadpedv20n9-021

Recebimento dos originais: 08/11/2023
Aceitação para publicação: 11/12/2023

Dulcídia Carlos Guezimane Ernesto
Doctoral Student in Agricultural Statistics and Experimentation
Institution: Universidade Púnguè and Universidade Federal de Lavras
Address: Av. Bueno da Fonseca, 543, Inácio Valentim, Lavras – MG,
CEP: 37200-000
E-mail: dcg.ernesto@gmail.com or dcg.franque@gmail.com

Leila Maria Ferreira
PhD in Agricultural Statistics and Experimentation
Institution: Universidade Federal de Lavras
Address: Av. Bueno da Fonseca, 543, Inácio Valentim, Lavras – MG,
CEP: 37200-000
E-mail: leilamaria2003@gmail.com

Thelma Sáfadi
PhD in Statistics
Institution: Universidade Federal de Lavras
Address: Av. Bueno da Fonseca, 543, Inácio Valentim, Lavras – MG,
CEP: 37200-000
E-mail: safadi@ufla.br

ABSTRACT

The implementation of statistical methods that can make the similarity study process more efficient has been very common in current research, because some methods such as wavelets make the processing of data manipulation faster and with the ability to process thousands of nucleotides in a genomic sequencing, and consequently better estimates are obtained. Thus, this work aimed to analyze some genomes of the Coronaviridae family and the Paramyxoviridae family, looking for similarities through signal decomposition considering the non-decimated discrete wavelet transform techniques. The entire genome sequences of the families were initially extracted from the NCBI website and then pattern analysis was performed using a signal processing method, based on the GC

content. For each GC sequence, decomposition was performed using the six-level discrete non-decimated Daubechies wavelet transform. Then, the Hurst exponent was calculated for each level of decomposition and the sequences were verified with similar patterns. After the cluster analysis, it was possible to obtain that the pair Bet1 and MERS proved to be similar in almost all methods, the other pair of sequences that proved to be similar by the absolute moments methods and by the aggregated variance method was the pair Gamma1 and Del, the group composed of Influ1, Influ4, Influ3 and Hendra was similar by the absolute moments methods, aggregated variance, and in the R/S analysis method, there was a substitution of Influ4 by Influ5. At the end of the study, it was possible to conclude that the non-decimated discrete wavelet transform allowed us to decompose each one of the sequences and, consequently, a more detailed study of similarity can be carried out, and it was possible to obtain, by some methods, strains of Coronaviridae that do not resemble the strains of Paramyxoviridae.

Keywords: genetic sequencing, Hurst exponent, similarity, wavelets.

RESUMO

A implementação de métodos estatísticos que podem tornar o processo de estudo de similaridade mais eficiente tem sido muito comum na pesquisa atual, porque alguns métodos, como as ondas, tornam o processamento de manipulação de dados mais rápido e com a capacidade de processar milhares de nucleotídeos em uma sequência genômica, e consequentemente, melhores estimativas são obtidas. Assim, este trabalho teve como objetivo analisar alguns genomas da família Coronaviridae e da família Paramyxoviridae, procurando semelhanças através da decomposição de sinal considerando as técnicas de transformação de ondas discretas não-decimadas. Todas as sequências genômicas das famílias foram inicialmente extraídas do site do NCBI e, em seguida, a análise de padrões foi realizada usando um método de processamento de sinais, com base no conteúdo GC. Para cada sequência de GC, a decomposição foi realizada usando a transformação de onda de Daubechies discreta de seis níveis e não-dizimada. Então, o expoente de Hurst foi calculado para cada nível de decomposição e as sequências foram verificadas com padrões semelhantes. Após a análise de cluster, foi possível obter que o par Bet1 e Mers se mostraram semelhantes em quase todos os métodos, o outro par de sequências que se mostraram semelhantes pelos métodos de momentos absolutos e pelo método de variância agregada foi o par Gamma1 e Del, o grupo composto de Influ1, Influ4, Influ3 e Hendra foi semelhante pelos métodos de momentos absolutos, variância agregada, e no método de análise R/S, houve uma substituição de Influ4 por Influ5. No final do estudo, foi possível concluir que a transformação discreta de ondas não-dizimadas permitiu decompor cada uma das sequências e, consequentemente, um estudo mais detalhado de similaridade pode ser realizado, e foi possível obter, por alguns métodos, linhagens de Coronaviridae que não se assemelham às linhagens de Paramyxoviridae.

Palavras-chave: sequenciamento genético, exponente de Hurst, similaridade, wavelets.

1 INTRODUCTION

Finding similarities in the sequencing of some evolutionarily similar species of viruses could be a way for scientists to find the formula for a preventive vaccine, or for the development of a cure for a disease. Therefore, the implementation of statistical methods that can make the process of studying similarities more efficient is welcome in the sense that the faster and with the capacity to process thousands of nucleotides in a genomic sequence, the better estimates will be obtained. Furthermore, if the researcher intends to identify emerging viruses and bacteria and work with a very large genomic database, it is essential to implement a clustering methodology that is reliable for genomic sequences in real time.

Several methodologies have been used by researchers to study genetic sequencing; however, wavelets have become an effective and efficient alternative for processing genetic sequences in real time, to study similarities, but specifically the method of estimating the exponent of Hurst in this type of study involving respiratory tract virus genomes has been very little used to date. In this work, 5 different methods were applied to estimate the Hurst exponent to study similarities in the genomes of respiratory tract viruses of the families Coronaviridae and Paramyxoviridae. This methodology has already been applied in the work by Ferreira et al. (2020) finding similarities in tuberculosis strains.

There are many approaches so far involving the Hurst exponent and wavelets. Das and Kumar (2021) carried out research in the financial area with the objective of optimizing a portfolio and showing that by combining the Hurst exponent and wavelet analysis they could help in increasing portfolio returns. In this study the authors use the Hurst exponent and wavelets to analyze the long-term dependencies between sovereign bonds and sector indices in India.

Furthermore, Arouxet et al. (2022) analyzed the impact of covid19 on cryptocurrencies using the Hurst exponent and wavelets. Their research examines long-term memory of return and volatility using seven-currency high-frequency time series. This analysis mainly focused on two periods, namely: the pre-covid19 period and the subsequent pandemic period, implemented the wavelet transform as it allowed to have more robust estimators of the Hurst exponent.

Considering all above, this study in general aims to analyse some genomes of the Coronaviridae family and the Paramyxoviridae family looking for similarities through signal decomposition with the support of the non-decimated discrete transform techniques of the Daubechie wavelet. Furthermore, the study seeks to apply the R/S, Peng methods, aggregate variation, differentiated aggregate variation, and the method of absolute moments to analyze similarities between the genomes of the Coronaviridae and Paramyxoviridae family.

2 MATERIAL AND METHODS

Complete genome sequences of the Coronaviridae and Paramyxoviridae virus families obtained from the National Center for Biotechnology Information (NCBI) database were considered. Therefore, it is essential to know the virus that circulates in the population infected by both Sars-Cov-2 and parainfluenza so that researchers can track its mutations and map the circulating strains and later develop more efficient methods of combating diseases from these viruses. Table 1 shows information related to each of the sequences in the NCBI, the first column is the identification of the access number on the platform, the second column is the nomenclature of each sequence, within the third column is the abbreviation created to facilitate the elaboration of the dendrograms, the fourth column shows the GC coefficient rate per sequence, and the last column shows the number of nucleotides for each sequence.

Table 1. Genomes Description of the Paramyxoviridae and Coronaviridae families

NCBI Access	Name of the Virus	Sequence	Total GC Content Rate	Nucleotide number
NC_038294	Betacoronavirus England 1	Bet1	0.3205908	14676
NC_019843	Middle East respiratory syn-drome-related coronavirus	MERS	0.4117764	14676
NC_006213	Human coronavirus OC43	Bet2	0.36788	14676
NC_003045	Bovine coronavirus	Bet4	0.36788	14676
NC_004718	SARS coronavirus TOR2	Bet5	0.4076166	14676
NC_028752	Camel alpha coronavirus	Alf1	0.3446086	14676
NC_002645	Human coronarirus 229E	Alf2	0.3841323	14676
NC_048214	Coronavirus duck	Gama1	0.3826189	14676
NC_010800	Turkey coronavirus	Gama2	0.4123643	14676
NC_005831	Coronavirus humano NL63	Alfa	0.3927362	14676

NCBI Access	Name of the Virus	Sequence	Total GC Content Rate	Nucleotide number
NC_006577	Human Coronarirus HKU1	Beta	0.382521	14676
KJ481931.1	Deltacoronavirus	Del	0.4317878	14676
NC_003461.1	Human parainfluenza virus 1	Influ1	0.3724359	14701
NC_001796.2	Human parainfluenza virus 3	Influ3	0.3452335	14676
NC_021928.1	Human parainfluenza virus 4a viral cRNA	Influ4	0.3623035	14676
NC_006430.1	Parainfluenzavirus 5 strain W3A	Influ5	0.4226682	14676
NC_001906.3	Hendrahenipavirus	Hendra	0.394428	14676

Source: The authors.

Furthermore, the Table 1 show, the sequence classes belonging to the Coronaviridade family: Beta, Bet1, Bet2, Bet3, Bet4, Bet5, MERS, Alpha, Alf1, Alf2, Del, Gamma1, Gamma2, and those that belong to the Paramyxoviridae family: Influ1, Influ3, Influ4, Influ5 and Hendra.

Therefore, the variants belonging to the Coronaviridae family that have been the subject of research are divided in two groups, the variants that have humans and vertebrate animals as hosts, and those that have only humans as hosts. However, the Bet1 variant, which was isolated in England (de Groot et al. 2013), the MERS variant which originated from the HCov-EMC strain and was isolated in 2012 in a patient in Sudita Arabia (van Boheemen et al. 2012), the Bet2 variant originated from the ATCC VR-759 strain (St-Jean et al. 2004), the Bet4 variant isolated from BCoV-ENT (Chouljenko et al. 2001), the Bet5 variant isolated from Tor2 (He et al. 2004), the Alf1 variant isolated from camel/Riyadh/Ry141/2015 and that originated from MERS-CoV in Sudita Arabia.

The Alf1 variant have three camel coronavirus species, and that according to Sabir et al. (2016) MERS-CoV and a human strain of CoV 229E species circulated simultaneously with high prevalence causing respiratory tract co-infections in dromedary camels, and camels shared the three covid species with humans.

The Alf2 variant, the Gamma1 variant dominant in birds such as duck, isolated from DK/GD/27/2014 (Zhuang et al. 2015), the Gamma2 variant isolated from an outbreak of enteritis in young turkeys (Gomaa et al. 2008), and the Del variant, originating in pigs in the United States from the

SDCV/USA/Illinois121/2014 strain (Marthaler et al. 2014), are variants with human and vertebrate animal hosts.

While the variants: Alpha, which has a complete genome sequence, which indicates that the virus does not come from recombination's (van der Hoek et al. 2004) and the Beta variant, first diagnosed in Shenzhen, China in a71-year-old patient with pneumonia (Woo et al. 2005), have their occurrence only in human hosts.

On the other hand, the Paramyxoviridae family is also composed of Influ1 which is originated from the Washington strain (Newman et al. 2002), Influ3, Influ5 which originated from strain W3A, and Hendra which originated in Australia and was diagnosed for the first time in an outbreak that caused the death of horses (Wang et al. 2000). Furthermore, these strains of the Paramyxoviridae family occur in human hosts and vertebrate animals, however, Influ4 is the only strain that occurs only in humans and had its origin in the W-25 strain (Komada et al. 2011).

For each genome, sequences of the GC content were determined, with a sliding window of size n=100. Using the Daubechies wavelet with 4 null moments, the non-decimated wavelet transform was applied to the GC sequences, obtaining decomposition in six levels of resolution for each sequence.

To evaluate the similarity between the sequences under study, the Hurst exponent was calculated at each level of decomposition using five different methods, namely: the aggregated variance method, the differentiated aggregated variance method, the absolute moment's method, the Peng method, and the R/S analysis method. Finally, a cluster analysis was performed based on the Mahalanobis distance. All analyses were performed using R software (R Core Team 2017).

Following is presented all the methodology used in the analysis.

2.1 THE GC CONTENT

GC content is often used to map genome composition and understand the evolution of its coding sequence (Saini and Dewan 2016). To obtain the signal referring to the genomes under study, the GC content was estimated with a

sliding window of size n = 100. The GC content is calculated as the ratio of the sum of Guanine (G) and Cytosine (C) bases to the sum of the bases Adenine (A), Guanine (G), Cytosine (C) and Tinine (T), given by equation 1.

$$GC_{content} = \frac{nG+nC}{nA+nG+nC+nT}, \quad (1)$$

Where

nA, nG,nC and nT represent the number of nucleotides of the bases of A, C, G, and T respectively. The proportions obtained according to Saini and Dewan (2016) can be interpreted, in which GC-rich regions indicate the inclusion of many protein-coding genes, that is, by determining the GC proportion, gene-rich regions in the genome are identified and at the same time as the signals of each sequence are obtained.

After obtaining the signal of each sequence, the decomposition phase followed according to the non-decimated wavelet transform.

2.2 WAVELETS

The wavelets, according to Denault et al. (2021), are mathematical functions that help to conduct transforms, such as the Fourier transform or Haar transform, among others. Brassarote et al. (2015) linked that one of the several applications of wavelets has been in areas of analysis of differential equation signals, this is mainly due to their ability to localize in time-frequency contrary to what occurs with Fourier analysis, bringing these forms a different and innovative perspective on the treatment of a non-stationary function database.

According to Abreu and Speckbacher (2022) wavelets satisfy the indispensable condition of admissibility (equation 2):

$$C_{\psi} = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega < \infty. \quad (2)$$

According to Alshammari et al. (2023), there are two types of wavelets, namely father wavelets and mother wavelets, which are represented in equations 3 and 4 respectively:

$$\phi_{j,k} = 2^{\left(\frac{-j}{2}\right)}\phi\left(t - \frac{2^j k}{2j}\right), \quad (3)$$

$$\psi_{j,k} = 2^{\left(\frac{-j}{2}\right)}\psi\left(t - \frac{2^j k}{2j}\right),$$

where

j = 1,2,3, …, J, and J are the levels of decomposition.

$$\int \phi(t)dt = 1\ ,\ \int \psi\ (t)\ dt = o \quad (4)$$

It is recommended that wavelets have compact support. Still from the perspective of Alshammari et al. (2023), the general mathematical model is represented in equation 5:

$$F_{j,k} = \int \psi_{j,k} f(t)dt \quad (5)$$

2.2.1 Daubechies Wavelets

Daubechies wavelets belong to the family of orthogonal wavelets and are characterized by the number of null moments according to an established support, in addition, they define a transformation of discrete wavelets, according to Rashid et al. (2020), the Daubechies wavelet can make discrete wavelet analysis feasible. They are defined by the number of null moments (m), which corresponds to the filter length (2m) (Rashid et al. 2020). It is advantageous to work with wavelet functions and Daubechies scale functions as functional bases for the multi-resolution approximation as they can radically change the formulation and solution of the equations of motion (Nastos and Saravanos 2021).

2.2.2 Non-Decimated Discrete Wavelet Transform

According to Wang et al. (2019) the non-decimated discrete wavelet transform not only retains the properties of the decimal wavelet transform, but also has the property of being invariant to displacement, that is, the non-

decimated discrete wavelet transform does not use the down sampling operation during the signal decomposition process, it inserts zeros every two coefficients in order to make the filter expansion in the high-pass and low-pass filtering process.

According to Yu et al. (2021), at each decomposition level of the non-decimated discrete wavelet transform, the scale filter ($H_0\ (z)$) and the wavelet filter ($H_1(z)$) are upsampled by filling zeros for each response coefficient to the impulse of the filters. The coefficients of filters at a certain level j are calculated by equations 6 and 7.

$$d_j^{2n}(-k) = d_{j-1}^{n}(k) * h_{0j}(k) \quad (6)$$

$$d_j^{2n+1}(-k) = d_{j-1}^{n}(k) * h_{1j}(k) \quad (7)$$

2.3 HURST EXPONENT

According to Soterroni et al. (2008), the Hurst exponent (H) is a tool capable of providing information related to correlation and persistence in a time series.

According to Rout et al. (2022), the Hurst exponent is defined in the range (0.1) and is strictly less than 0.5 in sequences that are anticorrelated and is in the range of 0.5 < H < 1 when the sequences are positively correlated. For cases where the Hurst exponent is equal to 0.5, it can be concluded that the sequence presents random patterns (Rout et al. 2022).

The Hurst exponent can be estimated considering 5 different techniques:

2.3.1 R/S Analysis

According to Lara-Musule et al. (2021), the R/S analysis measures the amplitude of the deviations of the partial sums from the mean rescaled by the standard deviation of the series. The R/S analysis method is described by Bărbulescu et al. (2010) and by Lara-Musule et al. (2021). Therefore, according to Lara-Musule et al (2021), when we take a time series $Z_N = Z_i$ with length N, we consider a subsequence $X_{NS} = X_i$ with length N_S where $N_S < N$. Therefore, the R/S analysis for X_{NS} is calculated as follows:

$$(R/S) = \frac{1}{\sigma_S(N_S)}\left\{\max_{1 \le i \le M} \sum_{k=1}^{i}(X_k - \bar{X}_{Ns}) - \min_{1 \le i \le M} \sum_{k=1}^{i}(X_k - \bar{X}_{Ns})\right\}. \quad (8)$$

Where

$\bar{X}_{NS}$ is the mean of the subsequence and $\sigma_S(N_S)$ is the standard deviation of the sample, which are defined as:

$$\bar{X}_{Ns} = \frac{1}{N_S}\sum_{i=1}^{N_S} X_i \quad (9)$$

and

$$\sigma_S(N_S) = \sqrt{\frac{1}{N_S}\sum_{k=1}^{N_S}(X_k - \bar{X}_{Ns})^2} \quad (10)$$

Still from the perspective of Lara-Musule et al. (2021) the R/S analysis is guided by the power law, which states that:

$$(R/S) = aN_S^H \quad (11)$$

Where

a is a constant and H is the Hurst exponent. A log-log plot of (R / S) versus $N_S \in (N_{(S,min)}, N_{(S,max)})$, gives a straight line with slope H. Therefore, if the plot is a straight line with a slope H = 0.5, means that the time series data are independent. Furthermore, if H > 0.5 indicates that the time series is persistent and at the same time presents a long-term autocorrelation. And if H < 0.5, it means that the autocorrelations in the signal are anti-persistent.

2.3.2 Aggregate Variance Method

The aggregate variance method is based on the self-similarity property of the samples of a process (Aldea and Tarniceriu 2013). When considering X a time series with length N, it is divided into d subseries of length m, and for each aggregate series, composed by an average given by Aldea and Tarniceriu (2013):

$$X^{\boldsymbol{m}}(\boldsymbol{k}) = \frac{1}{m}\sum_{i=(k-1)m+1}^{km} X_i,\ k = 1, 2, \dots, d \quad (12)$$

and by the sample variance given by Bărbulescu et al. (2010):

$$VarX^m = \frac{1}{d}\sum_{k=1}^{d}(X^{(m)}(k) - \bar{X})^2. \quad (13)$$

Bărbulescu et al. (2010) also state that for successive values of m, the sample variance is computed in a log-log plot against m. And the least squares will be fitted on a line for all points on the graph, and the Hurst coefficient will be calculated considering the slope of the straight line which is given by 2H-2.

2.3.3 Differentiated Aggregate Variance Method

This methodology was described by Teverovsky and Taqqu (1997), in which they examined the effect of certain types of non-stationarity on the detection of long-range dependence on Hurst's estimate of parameter H, when the variance estimator is applied. The purpose of applying this method is to distinguish the case in which H is close to 0.5 and there is a certain non-zero trend, from the cases in which H is far from 0.5, or otherwise, ends up being significantly greater than 0.5.

According to Teverovsky and Taqqu (1997), to detect a long-range dependence it is necessary to analyse the sample variation of a time series $X = \{X_i, i = 1, 2, \dots\}$, at several levels. of aggregation m, in the aggregate series of order m. By taking the following equation,

$$X^{\boldsymbol{m}}(\boldsymbol{k}) = \frac{1}{m}\sum_{i=(k-1)m+1}^{km} X_i,\ k = 1, 2, \dots, d \quad (14)$$

Therefore, the sample variation of the aggregate series X^m will be given by:

$$\widehat{\text{var}}\, X^{(m)} = \frac{1}{N/m}\sum_{k=1}^{N/m}\{X^{(m)}(k)\}^2 - \left\{\frac{1}{N/m}\sum_{k=1}^{N/m}X^{(m)}(k)\right\}^2 \quad (15)$$

The aggregate variance method can acquire the estimates through the logarithm of the first order difference. When considering the sequence given by m_1, m_2 from the values of m, and taking the difference, we obtain the following (Teverovsky and Taqqu 1997):

$$\frac{d\widehat{var}X^{(m)}}{dm} \approx \beta C_2 m^{\beta-1} \qquad (16)$$

and it is assumed that

$$\Delta\widehat{var}X^{(m)} \approx \frac{d\widehat{var}X^{(m)}}{dm}\Delta m$$

Then

$$\log\left(\Delta\widehat{var}X^{(m)}\right) \approx \log\left(\frac{d\widehat{var}X^{(m)}}{dm}\right) + \log\Delta m$$

Analogously ist necessary to consider that

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

Since we will work with log-log graphs, the following approximations will be used:

$$log\Delta Var[\bar{X}_m(k)] = (2H-3)\log m + \log(2H-2)C + \log m + C_1 = (2H-2)\log m + C_2 \qquad (17)$$

When applying this method, it is expected that a line with slope equal to 2H-2 will also be obtained.

2.3.4 Absolute Moments Method

From the perspective of Garcin (2022), this method allows the researcher to make a good assessment of the model specification, and is a generalization of the aggregate variance method as it uses the same principle as X^m (Aldea and Tarniceriu 2013).

The absolute moments are calculated for the aggregate series (Bărbulescu et al., 2010), and the nth absolute moment is given by:

$$AM_n^{(m)} = \frac{1}{(N/m)} \sum_{k=1}^{(N/m)} |\bar{X}_m(k) - \bar{X}_N|^n \quad (18)$$

where

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

In this method the estimates of the Hurst exponent will be calculated starting from the calculation of $AM_n^{(m)}$ considering different values of m and generating a log-log plot against m. What is expected when applying this methodology is that the points are spread along a straight line, with slope given by n(H-1).

2.3.5 Peng Method

This method is discussed in the literature by Adler et al. (1998), in which four fundamental steps can be applied:

1st Step: observe each block of size m and calculate the partial sum within the blocks according to equation 19.

$$Y(k)^m = \sum_{t=(k-1)m+1}^{km} X(t),\ k = 1, 2, \dots, (N/m) \quad (19)$$

2nd Step: this step comprises the adjustment phase of a regression line of the type $g = a + bk$, which will help us to calculate the variance of the residual, which is given by equation 20.

$$s_r^{(m)} = \frac{1}{m}\sum_{k=1}^{N/m}(Y(k)^m - a - bk)^2 \quad (20)$$

Step 3: Sketch the graphs of $logs_r^{(m)} versus \log m$

4th Step: Finally, the slope to be obtained must be equal to 2H.

Cluster Analysis

Cluster analysis is the classification of objects by groups, that is, each group must contain objects with similar characteristics, according to some statistical distance function.

From the perspective of Manly (2008), two forms of grouping can be used, the hierarchical and division grouping method.

According to Frei (2006), hierarchical methods follow a specific algorithm:

Search for the two most similar objects in the similarity matrix.

Objects i and j are removed and form a group, in which the corresponding row and column i and j are eliminated.

A row and a column are defined, which are obtained by the distances between the group i j and the remaining objects, according to the adopted algorithm procedure.

The previous steps are repeated n-1 times, so that all n objects have a group until the end of the algorithm.

Furthermore, hierarchical techniques were used, as they allow the construction of dendrograms, and the implementation of these hierarchical techniques begins with the calculation of the distances of each object to all other objects. The distances from each object to all objects will be calculated using the mahalanobis distance function.

According to Yang and Delpha (2022), the Mahalanobis distance of a sample is calculated considering two sample vectors X, Y from the same distribution and is given by:

$$d_M(X) = \sqrt{(X - \mu)^T \Sigma^{-1} (X - \mu)} \quad (21)$$

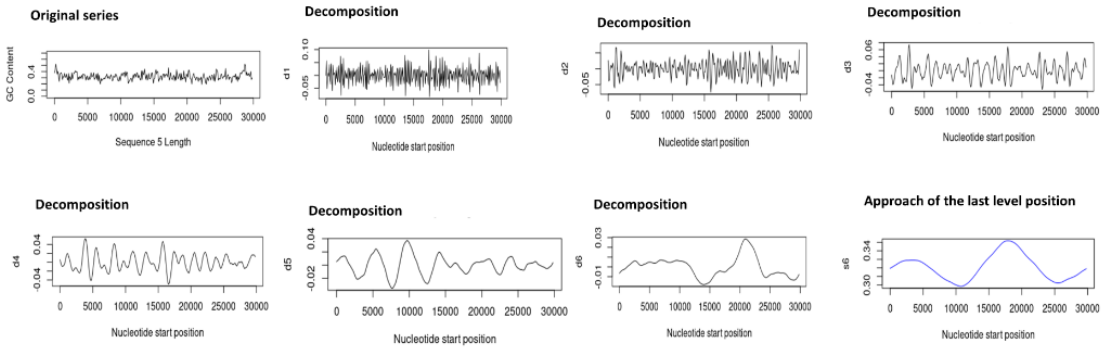
where

μ is a mean vector, and Σ is the covariance matrix. Therefore, the Mahalanobis distance is always a non-negative value, which is calculated in units of deviations, and takes variability into account (McLachlan 1999).

3 RESULTS

Table 1 summarizes the important aspects of the analyzed genomes. The decomposition of the GC content sequence is exemplified by Betacorona virus England and is shown in Figure 1.

Figure 1: Decomposition of the human coronavirus HKU1(Beta) sequence



Source: The authors.

Table 2 illustrates the Hurst exponent values for the aggregated absolute value method for the decomposition levels of each GC sequence under study.

Table 2. Hurst Exponent Values by absolute moments method

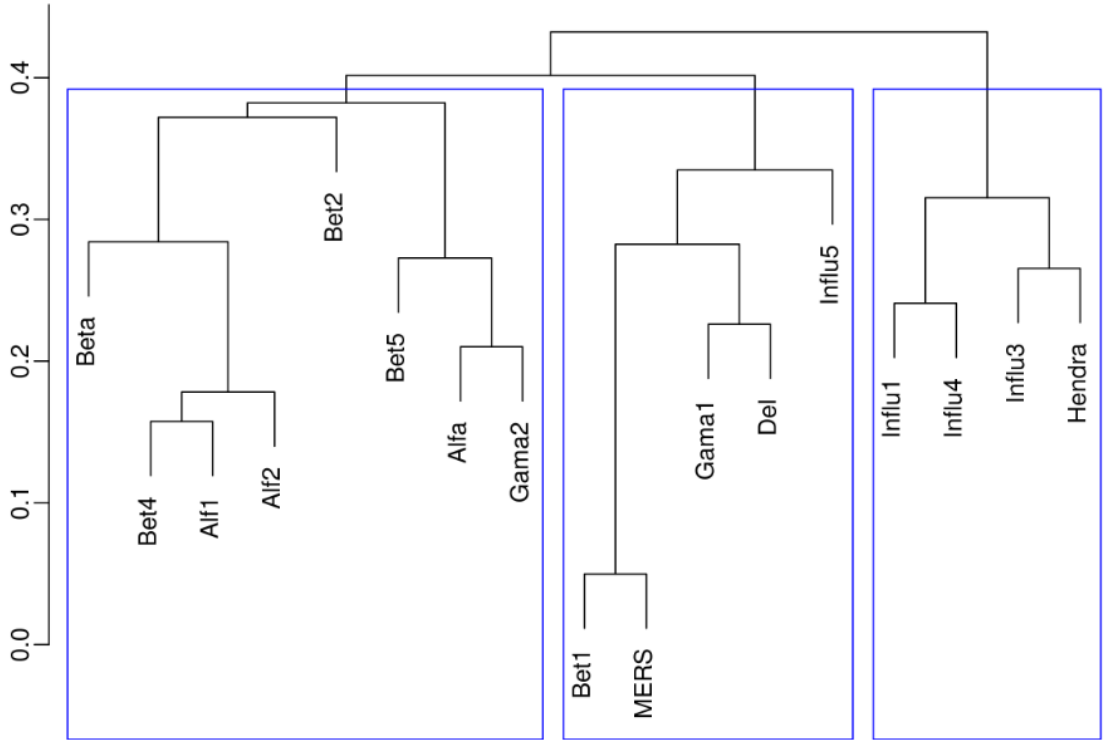
Sequences	Level1	Level2	Level3	Level4	Level5	Level6
Beta	0.15925	0.06039	0.03276	0.03453	0.18273	0.83592
Bet1	0.11744	0.14182	0.07244	-0.02129	0.36787	0.73179
Bet2	0.14523	-0.11753	0.10988	-0.10595	0.28212	0.69620
Bet3	0.14523	-0.11753	0.10988	-0.10595	0.28212	0.69620
Bet4	0.11044	0.04285	-0.11222	0.01375	0.27903	0.69817
Bet5	-0.15239	-0.26166	-0.06326	-0.075288	0.25063	0.79646
Alfa	-0.07652	-0.12263	-0.00088	-0.04076	0.28340	0.70964
Alf1	0.06487	0.02972	-0.13859	0.06475	0.23351	0.72348
Alf2	0.08155	0.14401	-0.20865	0.08037	0.18944	0.72569
MERS	0.16466	0.16676	0.07986	-0.02699	0.36907	0.73739
Gama1	1.05926	0.16426	0.21280	-0.12138	0.27105	0.83348
Gama2	-0.11325	-0.04656	0.07619	0.02878	0.33219	0.83523
Del	-0.01649	0.28606	0.06015	-0.04101	0.31838	0.86559
Influ1	-0.03829	-0.17001	0.03567	0.07965	0.26689	0.99896
Influ3	-0.03885	0.09411	-0.02732	0.13057	0.25049	1.03683
Influ4	0.067420	-0.17437	0.05549	0.07060	0.46435	1.09032
Influ5	0.19312	0.20963	-0.06607	-0.14383	0.29561	1.01124
Hendra	0.11724	0.06061	-0.10491	0.09032	0.33878	0.99014

Source: The authors.

The same calculation was made for the other techniques. The package Nbclust of R was used and applied in all five methodologies to verify the optimal

number of clusters to be applied in the formation of the groups, and from Figure 2 it can be observed that for the method of aggregated absolute value we obtained a grouping with 3 groups, and some subgroups.

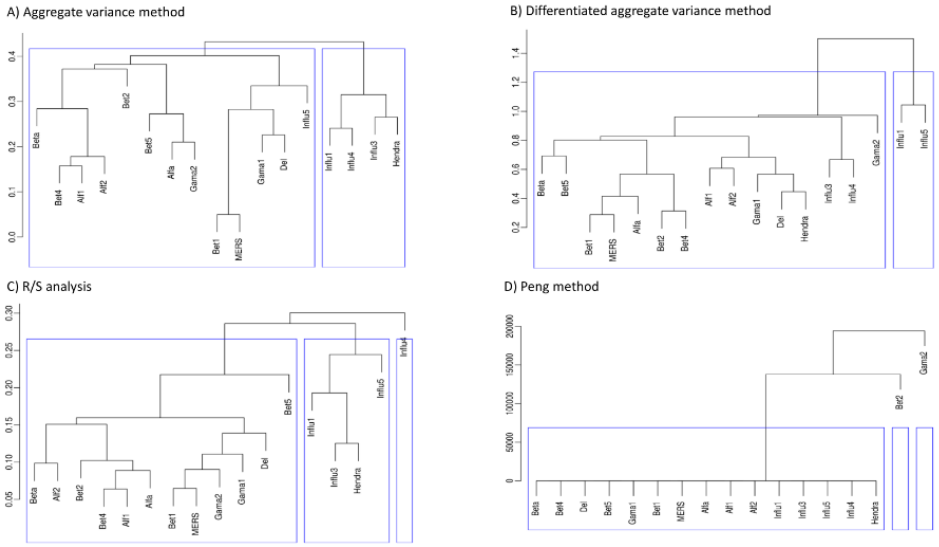
Figure 2: Groupings of sequences by similarities by the method of absolute moments



Source: The authors.

Figure 3 reflects the groups formed by the aggregated variance, differentiated aggregated variance, Peng method and R/S analysis methods, and the aggregated variance and differentiated aggregated variance methods presented the formation of two groups, the R/S analysis presented the formation of three groups and the Peng method presented the formation of four groups. Furthermore, the R/S method and the Peng method had Hurst exponent estimates H>0.5 at levels 4, 5 and 6, which means that the sequences are positively correlated, and levels 1 and 2 presented $H < 0.5$, which indicates the existence of long-range dependence.

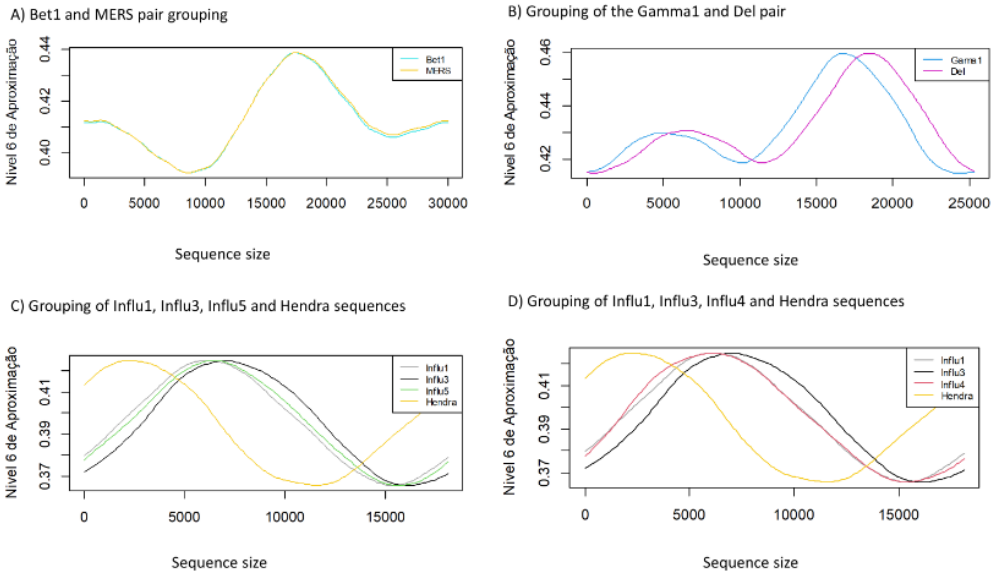
Figure 3: Groupings by other methods



Source: The authors.

The differentiated aggregate variance method presented the estimate of $H > 0.5$ at levels 5 and 6, while at levels 1, 2, 3 and 4, $H < 0.5$, the aggregated variance and aggregated absolute value methods presented the estimates of $H > 0.5$ at level 6, while at levels 1, 2, 3 and 4 (except for the Gamma1 variant) the $H < 0.5$. Finally, the graphs of the last level of approximation of the groups that were similar throughout the application of the techniques were verified, as shown in Figure 4.

Figure 4: Similar groupings that are repeated throughout the techniques used



Source: The authors.

4 DISCUSSION

Figure 1 shows the decomposition process that was carried out at each of the levels where the non-decimal discrete wavelet transform was applied. The Hurst exponent as a measure to verify similar genomic sequences has been previously addressed by Ferreira et al. (2020), in which they apply the 5 techniques discussed here to assess similarities in strains of the tuberculosis genome.

Figure 2 shows the formation of 3 clusters, after the implementation of the method of absolute moments, and in these clusters, we must highlight some patterns that are also repeated in other methodologies. The Bet1 and MERS pair proved to be similar in almost all methods, except for the Peng method. Furthermore, previous studies indicate that the phylogenetic analysis of SARS-CoV-2 belongs to the genus Betacoronavirus that includes the SARS bat types, SARS-Cov and MERS-SAR (Petrosillo et al. 2020). When analyzing the similarities in sequences of the Coronaviridae and Paramyxoviridae families, a high similarity was observed between the Bet1 and MERS pair, and from figure 4 A) the graph of the Bet1 and MERS sequences almost overlap.

The other pair of sequences that proved to be similar by the absolute moment methods and by the aggregated variance method was the pair Gama1 and Del. Previous studies indicate that the Gamma variant is immune to the Coronavac vaccine, therefore, this strain is immune to neutralizing antibodies generated in order to respond to polyclonal stimulation against the variants that circulated before the Gamma variant (Souza et al. 2021), on the other hand , according to Bian et al. (2021), the Delta variant contains mutations L452R, T478K, D614G and P681R that may also affect resistance to specific antibodies and that, according to Hu et al. (2021) after two injections of Coronavac the neutralizing titer decreased 2.7-fold.

The group composed of Influ1, Influ4, Influ3 and Hendra was similar by the methods of absolute moments, aggregated variance, and in the R/S analysis method, there was a substitution of Influ4 by Influ5. Furthermore, according to Wong et al. (2016) Hendra virus entry into the paramyxovirus cell is mediated by the F fusion protein, in response to the binding of a host receptor by the binding protein. Therefore, the structure of Hendra virus clearly shows a very high

similarity with the structure of the parainfluenza virus5. On the other hand, the parainfluenza 5 pair of fusion proteins (F) together with the parainfluenza 3 and Newcastle disease virus binding protein pair behave according to the fusion activation association model (Wong et al. 2016), on the other hand, in Figure 4D) the Influ1 and Influ4 sequences almost overlap. In addition, the human parainfluenza virus types 1, 3 and 4 cause mild, severe, or prolonged infections in the respiratory tract, such as bronchiolitis and pneumonia.

However, Liu et al. (2013) developed studies to explore the epidemiological characteristics and clinical manifestations of parainfluenza virus variants in types 1, 2, 3 and 4 to establish clinical distinctions, and found that the only significant difference in clinical presentation between types of parainfluenza was sputum, and that the most frequent types in patients were Influ 3 and Influ1, this theory is confirmed in research by Denny and Clyde (1986), Glezen et al. (1984) and Murphy et al. (1980), and that the Influ 1 and Influ 3 types present many similar clinical features that somehow differ from parainfluenza 4 (Influ 4). Therefore, one of the aspects to be considered in this research is the fact that it was found that the genetic sequencing of the Influ1 virus is like the genetic sequencing of Influ4, a fact that had not been mentioned in previous research.

5 CONCLUSIONS

The non-decimated Daubechies discrete wavelet transform allowed the decomposition to be performed on each of the GC sequences. Furthermore, when the Hurst exponent estimation techniques were implemented, patterns with greater similarity were repeated throughout the applied methodologies, and the absolute moments and aggregate variance methods presented similar clusters compared to the other implemented techniques. On the other hand, the two virus families, despite having different ancestors, present some similarity in the strains because the methodology used was able to distinguish variants of Coronaviridae and Paramyxoviridae with similarities, which is the case of Influ5, which is identical to the Bet1 and MERS pair, and to the pair Gama1 and Del at least 30%. Not all strains mixed with other variants from other families, in the case of Influ1, it remained separate from Coronaviridae variants in all techniques, apart from the Peng method.

The absolute moment, aggregated variance and R/S analysis methods managed to separate the genomes of the families under study so that not all strains were mixed with other variants from other families, and the Peng method, in turn, did not show satisfactory results. The study shows that some sequences remained more isolated from the others, which is the case of Influ1 that remained separated from the variants of Coronaviridae in all the techniques, except in the Peng method. Furthermore, the methods of absolute moment, aggregated variance and R/S analysis showed better performance than the other applied methodologies.

REFERENCES

ABREU, L. D. and SPECKBACHER, M. Affine density, von Neumann dimension and a problem of Perelomov. *Advances in Mathematics*. 2022, *407*, 108564. https://doi.org/10.1016/j.aim.2022.108564

ADLER, R., FELDMAN, R. and TAQQU, M. *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Springer Science and Business Media. 1998, pp. 534.

ALDEA, R. and TARNICERIU, D. G. Estimating the hurst exponent in motor imagery-based brain computer interface. *7th Conference on Speech Technology and Human-Computer Dialogue, SpeD 2013, Cluj-Napoca, Romania*. 2013, 1–6. https://doi.org/10.1109/SpeD.2013.6682656

ALSHAMMARI, T., et al. Forecasting Stock Volatility Using Wavelet-based Exponential Generalized Autoregressive Conditional Heteroscedasticity Methods. *Intelligent Automation & Soft Computing*. 2023, *35*(3). https://doi.org/10.32604/iasc.2023.024001

AROUXET, M. B., et al. Covid-19 impact on cryptocurrencies: Evidence from a wavelet-based Hurst exponent. *Physica A: Statistical Mechanics and Its Applications*. 2022, *596*, 127170. https://doi.org/10.1016/j.physa.2022.127170

ASSAF, A., et al. Multivariate long memory structure in the cryptocurrency market: The impact of COVID-19. *International Review of Financial Analysis*. 2022, *82*, 102132. https://doi.org/10.1016/j.irfa.2022.102132

BĂRBULESCU, A., SERBAN and MAFTEI, C. Evaluation of Hurst exponent for precipitation time series. Latest trends on computers: Proceedings of the 14th WSEAS international conference on Computers: part of the 14th WSEAS CSCC multiconference – Volume II. 2010, *2*, 7.

BIAN, L., et al. Impact of the Delta variant on vaccine efficacy and response strategies. *Expert Review of Vaccines*. 2021,1. https://doi.org/10.1080/14760584.2021.1976153

BRASSAROTE, G. De O. N., et al. Análise multiescala a partir de wavelets não decimadas: Investigação dos efeitos da cintilação ionosférica nos sinais GPS. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*. 2015, *3*(1). https://doi.org/10.5540/03.2015.003.01.0476

CHOULJENKO, V. N., et al. Comparison of genomic and predicted amino acid sequences of respiratory and enteric bovine coronaviruses isolated from the same animal with fatal shipping pneumonia. *The Journal of General Virology*. 2001, *82*(12), 2927–2933. https://doi.org/10.1099/0022-1317-82-12-2927

DAS, S. and KUMAR, A. Long-term dependency between sovereign bonds and sectoral indices of India: Evidence using Hurst exponent and wavelet analysis. *Managerial Finance*. 2021, *47*(10), 1448–1464. https://doi.org/10.1108/MF-12-2020-0596

De Groot, R. J., et al. Middle East Respiratory Syndrome Coronavirus (MERS-CoV): Announcement of the Coronavirus Study Group. *Journal of Virology*. 2013, *87*(14), 7790–7792. https://doi.org/10.1128/JVI.01244-13

DENAULT, W. R. P., et al. A fast wavelet-based functional association analysis replicates several susceptibility loci for birth weight in a Norwegian population. *BMC Genomics*. 2021, *22*(1), 321. https://doi.org/10.1186/s12864-021-07582-6

DENNY, F. W. and CLYDE, W. A. Acute lower respiratory tract infections in non-hospitalized children. *The Journal of Pediatrics*. 1986, *108*, 635–646. https://doi.org/10.1016/s0022-3476(86)81034-4

FERREIRA, L. M., SÁFADI, T. and FERREIRA, J. L. Evaluation of genome similarities using a wavelet-domain approach. *Revista Da Sociedade Brasileira de Medicina Tropical*. 2020, *53*, e20190470. https://doi.org/10.1590/0037-8682-0470-2019

FREI, F. 2006. *Introdução à análise de agrupamentos: Teoria e prática*. São Paulo: UNESP, pp. 112.

GARCIN, M. A comparison of maximum likelihood and absolute moments for the estimation of Hurst exponents in a stationary framework. *Communications in Nonlinear Science and Numerical Simulation*. 2022, *114*, 106610. https://doi.org/10.1016/j.cnsns.2022.106610

GLEZEN, W. P., et al. Parainfluenza Virus Type 3: Seasonality and Risk of Infection and Reinfection in Young Children. *The Journal of Infectious Diseases*. 1984, *150*(6), 851–857. https://doi.org/10.1093/infdis/150.6.851

GOMAA, M. H., et al. Complete genomic sequence of turkey coronavirus. *Virus Research*. 2008, *135*(2), 237–246. https://doi.org/10.1016/j.virusres.2008.03.020

HE, R., et al. Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochemical and Biophysical Research Communications*. 2004, *316*(2), 476–483. https://doi.org/10.1016/j.bbrc.2004.02.074

HU, J., et al. Reduced neutralization of SARS-CoV-2 B.1.617 variant by inactivated and RBD-subunit vaccine. *bioRxiv*. 2021. https://doi.org/10.1101/2021.07.09.451732

KOMADA, H., et al. Completion of the full-length genome sequence of human parainfluenza virus types 4A and 4B: Sequence analysis of the large protein

genes and gene start, intergenic and end sequences. *Archives of Virology*. 2011, *156*(1), 161–166. https://doi.org/10.1007/s00705-010-0834-6

LARA-MUSULE, A., et al. Diagnosis and Monitoring of Volatile Fatty Acids Production from Raw Cheese Whey by Multiscale Time-Series Analysis. *Applied Sciences*. 2021, *11*(13). https://doi.org/10.3390/app11135803

LI, G. and GUO, Y. Exploring the Medication Pattern of Chinese Medicine for Peptic Ulcer Based on Data Mining. *Journal of Healthcare Engineering*. *2021*, 9072172. https://doi.org/10.1155/2021/9072172

LIU, W.-K., et al. Epidemiology and clinical presentation of the four human parainfluenza virus types. *BMC Infectious Diseases*. 2013, *13*, 28. https://doi.org/10.1186/1471-2334-13-28

MANLY, B. F. J. *Métodos estatísticos multivariados: Uma introdução*. 3th ed. Bookman, 2008.

MARTHALER, D., et al. Complete Genome Sequence of Strain SDCV/USA/Illinois121/2014, a Porcine Deltacoronavirus from the United States. *Genome Announcements*. 2014, *2*(2), e00218-14. https://doi.org/10.1128/genomeA.00218-14

MCLACHLAN, G. J. Mahalanobis distance. *Resonance*. 1999, *4*(6), 20–26. https://doi.org/10.1007/BF02834632

MURPHY, B., et al. Seasonal Pattern of Childhood Viral Lower Respiratory Tract Infections in Melbourne. *Medical Journal of Australia*. 1980, *1*(1), 22–24. https://doi.org/10.5694/j.1326-5377.1980.tb134568.x

NASTOS, C. V. and SARAVANOS, D. A. Multiresolution Daubechies finite wavelet domain method for transient dynamic wave analysis in elastic solids. *International Journal for Numerical Methods in Engineering*. 2021, *122*(23), 7078–7100. https://doi.org/10.1002/nme.6822

NEWMAN, J. T., et al. Sequence analysis of the Washington/1964 strain of human parainfluenza virus type 1 (HPIV1) and recovery and characterization of wild-type recombinant HPIV1 produced by reverse genetics. *Virus Genes*. 2002, *24*(1), 77–92. https://doi.org/10.1023/a:1014042221888

PETROSILLO, N., et al. COVID-19, SARS and MERS: Are they closely related? *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*. 2020, *26*(6), 729–734. https://doi.org/10.1016/j.cmi.2020.03.026

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria: Vienna: R Foundation for Statistical Computing, 2017. Available from: *http://www.R-project.org/*

RASHID, O., AMIN, A. and LONE, M. R. Performance Analysis of DWT Families. *3rd International Conference on Intelligent Sustainable Systems (ICISS)*. 2020, 1457–1463. https://doi.org/10.1109/ICISS49785.2020.9315960

ROUT, R. K., et al. Feature-extraction and analysis based on spatial distribution of amino acids for SARS-CoV-2 Protein sequences. *Computers in Biology and Medicine*. 2022, *141*, 105024. https://doi.org/10.1016/j.compbiomed.2021.105024

SABIR, J. S. M., et al. Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science Express. 2016, 351*(6268), 81–84. https://doi.org/10.1126/science.aac8608

SAINI, S. and DEWAN, L. Application of discrete wavelet transform for analysis of genomic sequences of Mycobacterium tuberculosis. *SpringerPlus*. 2016, *5*(1), 64. https://doi.org/10.1186/s40064-016-1668-9

SOTERRONI, A. C., DOMINGUES, M. O. and RAMOS, F. M. Estimativa do expoente de hurst de séries temporais caóticas por meio da transformada wavelet discreta. *In: Proceedings of theThematic Congresson Dynamics, Control and Applications-DINCON, Presidente Prudente, Brazil. 2008, 7-9*.

SOUZA, W. M., et al. Neutralisation of SARS-CoV-2 lineage P.1 by antibodies elicited through natural SARS-CoV-2 infection or vaccination with an inactivated SARS-CoV-2 vaccine: An immunological study. *The Lancet Microbe*. 2021, *2*(10), e527–e535. https://doi.org/10.1016/S2666-5247(21)00129-4

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*. 1997, *18*(3), 279–304. https://doi.org/10.1111/1467-9892.00050

VAN BOHEEMEN, S., et al. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *MBio*. 2012, *3*(6), e00473-12. https://doi.org/10.1128/mBio.00473-12

VAN DER HOEK, L., et al. Identification of a new human coronavirus. *Nature Medicine*. 2004, *10*(4), 368–373. https://doi.org/10.1038/nm1024

WANG, L. F., et al. The exceptionally large genome of Hendra virus: Support for creation of a new genus within the family Paramyxoviridae. *Journal of Virology*. 2000, *74*(21), 9972–9979. https://doi.org/10.1128/jvi.74.21.9972-9979.2000

WANG, X., et al. The UDWT image denoising method based on the PDE model of a convexity-preserving diffusion function. *EURASIP Journal on Image and Video Processing*. *2019,* (1), 81. https://doi.org/10.1186/s13640-019-0480-1

WONG, J. J. W., et al. Structure and stabilization of the Hendra virus F glycoprotein in its prefusion form. *Proceedings of the National Academy of Sciences*. 2016, *113*(4), 1056–1061. https://doi.org/10.1073/pnas.1523303113

YANG, J. and DELPHA, C. An incipient fault diagnosis methodology using local Mahalanobis distance: Fault isolation and fault severity estimation. *Signal Processing*. 2022, *200*, 108657. https://doi.org/10.1016/j.sigpro.2022.108657

YU, Y., et al. A Two-Stage Wavelet Decomposition Method for Instantaneous Power Quality Indices Estimation Considering Interharmonics and Transient Disturbances. *IEEE Transactions on Instrumentation and Measurement*. 2021, *70*, 1–13. https://doi.org/10.1109/TIM.2021.3052554

ZHUANG, Q.-Y., et al. Genomic Analysis and Surveillance of the Coronavirus Dominant in Ducks in China. *PloS One*, 2015, *10*(6), e0129256. https://doi.org/10.1371/journal.pone.0129256

TERCEIRA PARTE - CONSIDERAÇÕES FINAIS

CONSIDERAÇÕES FINAIS

As técnicas propostas para a análise de agrupamento sob o domínio das wavelets neste trabalho apresentaram resultados muito significativos.

A técnica com recurso aos métodos laço e ridge, teve melhor desempenho o método laço em detrimento ao método ridge, pelo fato do ridge ser um método de encolhimento de variáveis, não teve capacidade de agrupar os genomas por níveis de similaridade. Ao contrário do método laço que diferenciou os sequenciamentos pelo nível de similaridade com recurso a distância euclidiana pois o laço além de fazer seleção de vaiáveis, faz a penalização dos parâmetros segundo a descrição de Tibshirani (1996)

A técnica Elastic net uni o melhor dos métodos laço e ridge pois, à medida que faz o encolhimento, seleciona as variáveis e penaliza os parâmetros. No trabalho com recurso ao elastic net pôde-se observar que os primeiros níveis de decomposição apresentaram melhores resultados na formação dos grupos em comparação com os últimos níveis de decomposição, contrariando o trabalho realizado por Ferreira et. al (2018), em que os últimos níveis apresentaram resultados mais satisfatórios.

No trabalho utilizando o expoente de Hurst, também sob o domínio das wavelets foi possível verificar a formação de grupos para os sequenciamentos em estudo, a semelhança das outras metodologias utilizadas, contudo o expoente de Hurst considera a distância de Mahalanobis para a formação dos grupos.

Nesta metodologia, as técnicas que deram melhores resultados foram o método de variância agregada a semelhança do trabalho de Ferreira et. al (2019), e o método de momentos absolutos.

O grande contributo dos métodos utilizando o expoente de Hurst sob o domínio das wavelets, é pelo fato de não excluir ou restringir qualquer sequenciamento na análise de similaridade, e comparativamente aos outros métodos estudados, os métodos de variância agregada e de momentos absolutos proporcionam uma melhor leitura dos resultados dos agrupamentos, sem ignorar o fato de ser uma nova abordagem para o agrupamento de genomas similares, visto que até então teve sua aplicação maioritariamente em estudos na área de hidráulica.

Ademais, todas as metodologias que são abordadas neste trabalho conferem um rápido processamento mesmo quando se faz análises com uma base de dados muito grande. Ao se implementar as wavelets, detalhes omissos são evidenciados na medida que se faz a decomposição devido ao grau de refinamento durante o processamento dos dados.

Em trabalhos futuros seria interessante a implementação de outras wavelets, e o estudo em outras famílias de vírus que sejam compostas por RNA ou DNA.