DENISE DE ASSIS PAIVA

ANÁLISE DE SIMILARIDADE GENÔMICA ENTRE DIFERENTES CORONAVÍRUS: A CONTRIBUIÇÃO DOS MÉTODOS K-MER E *NATURAL VECTOR*

LAVRAS – MG

2024

DENISE DE ASSIS PAIVA

ANÁLISE DE SIMILARIDADE GENÔMICA ENTRE DIFERENTES CORONAVÍRUS: A CONTRIBUIÇÃO DOS MÉTODOS K-MER E *NATURAL VECTOR*

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

Profª. Drª. Thelma Sáfadi
Orientadora

Profª. Drª. Karla Suemy Clemente Yotoko
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).

Paiva, Denise de Assis

Análise de Similaridade Genômica entre Diferentes Coronavírus: A Contribuição dos Métodos K-mer e *Natural Vector* / Denise de Assis Paiva. 1^a ed. rev., atual. e ampl. – Lavras: UFLA, 2024.

65 p. : il.

Tese(doutorado)–Universidade Federal de Lavras, 2024.
Orientadora: Profª. Drª. Thelma Sáfadi.
Coorientadora: Profª. Drª. Karla Suemy Clemente Yotoko.
Bibliografia.

1. Coronavírus. 2. Alinhamento. 3. Análise de Cluster. I. Sáfadi, Thelma. II. Yotoko, Karla Suemy Clemente. III. Título.

DENISE DE ASSIS PAIVA

ANÁLISE DE SIMILARIDADE GENÔMICA ENTRE DIFERENTES CORONAVÍRUS: A CONTRIBUIÇÃO DOS MÉTODOS K-MER E *NATURAL VECTOR*
GENOMIC SIMILARITY ANALYSIS AMONG DIFFERENT CORONAVIRUSES: THE CONTRIBUTION OF K-MER AND NATURAL VECTOR METHODS

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

APROVADA em 11 de março de 2024.

Prof. Dr. Denismar Alves Nogueira	UNIFAL
Prof. Dr. Paulo Henrique Sales Guimarães	UFLA
Prof. Dr. Renato Ribeiro de Lima	UFLA
Prof. Dr. Tales Jesus Fernandes	UFLA

Profª. Drª. Thelma Sáfadi
Orientadora

Profª. Drª. Karla Suemy Clemente Yotoko
Coorientadora

LAVRAS – MG
2024

Ao meu irmão Téo pelo apoio incondicional em todos os momentos. Você é inspiração e luz.
Com amor, dedico.

AGRADECIMENTOS

Primeiramente, expresso minha gratidão a Deus, pois a jornada com fé é mais suave.

Aos meus familiares, manifesto meu profundo agradecimento por tudo que sempre fizeram por mim, desde a graduação até a realização deste sonho, especialmente à minha mãe, Maria, e ao meu irmão, Téo.

À minha orientadora, Profa. Thelma, reconheço todo o apoio e incentivo desde o momento em que ingressei na Universidade Federal de Lavras (UFLA).

À minha coorientadora, Profa. Karla, agradeço por sua constante atenção e suporte desde o primeiro contato.

Às pessoas especiais na minha vida nessa jornada, Ana, Ariane, Carlos, Fernanda, Lina, Luiz, Isabela, Izabela, Kaio, Marcelo e Paula e em particular, a Nadia.

Agradeço também a todos os professores do departamento de Estatística da UFLA, cujas contribuições foram fundamentais para o meu crescimento tanto profissional quanto pessoal.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

Estudos envolvendo métodos de alinhamento de sequências genômicas existem desde a década de 70. Entretanto, o processo para alinhar essas sequências ainda é relativamente demorado e requer computadores mais potentes, tanto para análise com genomas virais quanto principalmente com genomas de bactérias. Nesse sentido, os métodos livres de alinhamento conseguem superar essa questão alcançando a mesma precisão com um tempo de análise consideravelmente menor. Nesta tese, foram realizados estudos considerando dois métodos livres de alinhamento, um baseado em k-mer e o outro, o *Natural Vector*, na classificação de genomas virais. O método k-mer manteve a precisão e obteve um tempo menor de análise, conseguindo separar corretamente os grupos correspondentes às variantes e linhagens das sequências do SARS-CoV-2, em comparação com o método tradicional de alinhamento. O método *Natural Vector* classificou corretamente diferentes espécies de coronavírus também levando em conta menos tempo. Como cada método se mostrou preciso e o tempo de análise foi um ponto fundamental, verifica-se que ambos se complementam na classificação de novos vírus: o *Natural Vector* identifica corretamente a espécie de coronavírus em estudo, enquanto o k-mer agrupa os vírus dentro das espécies de forma concisa. A classificação rápida das sequências de coronavírus é de suma importância para o controle de epidemias, especialmente em época de surtos virais, pois nessas situações o tempo de análise é crucial.

Palavras-chave: Classificação. Genoma. Método livre de alinhamento. Pandemia.

ABSTRACT

Studies involving genomic sequence alignment methods have existed since the 1970s. However, the process of aligning these sequences remains relatively time-consuming and requires more powerful computers, both for analysis with viral genomes and particularly with bacterial genomes. In this regard, alignment-free methods can overcome this issue by achieving the same level of accuracy with significantly reduced analysis time. This thesis conducted studies considering two alignment-free methods, one based on k-mer and the other, Natural Vector, in viral genome classification. The k-mer method maintained precision and achieved a shorter analysis time, accurately segregating the groups corresponding to variants and lineages of SARS-CoV-2 sequences compared to the traditional alignment method. The Natural Vector method accurately classified different species of coronaviruses while also considering less time. As each method demonstrated precision and analysis time was a critical factor, it is evident that both methods complement each other in classifying new viruses: Natural Vector correctly identifies the species of coronavirus under study, while k-mer succinctly groups the viruses within the species. Swift classification of coronavirus sequences is paramount for epidemic control, especially during viral outbreaks, as analysis time is crucial in such scenarios.

Keywords: Classification. Free alignment method. Genome. Pandemic.

INDICADORES DE IMPACTO

A tese propõe um aplicativo usando o *software* R (pacote *Shiny*) para a classificação de sequências de coronavírus por meio de um método livre de alinhamento baseado em k mer e do método k-médias. Este avanço tecnológico tem o potencial de gerar diversos impactos significativos em diferentes esferas. No âmbito social, o aplicativo pode contribuir para a rápida identificação e caracterização de vírus, sendo crucial em situações de surtos e pandemias, promovendo a saúde pública e a tomada de decisões baseadas em evidências. Em termos tecnológicos, a criação desse aplicativo impulsiona o desenvolvimento de ferramentas avançadas de análise genômica, ampliando o conhecimento científico e as aplicações práticas na área da bioinformática. Do ponto de vista econômico, o uso desse aplicativo pode reduzir custos e tempo associados à análise de sequências virais, beneficiando laboratórios de pesquisa e instituições de saúde. Além disso, a capacidade de classificar rapidamente sequências virais também pode ter impactos culturais, aumentando a conscientização sobre a importância da genômica na medicina e na compreensão das doenças infecciosas, influenciando assim as políticas de saúde e as atitudes em relação à ciência e à tecnologia. Em suma, a criação desse aplicativo representa um avanço significativo que transcende o campo da biologia e estatística, impactando positivamente aspectos sociais, tecnológicos, econômicos e culturais da sociedade.

IMPACT INDICATORS

The thesis proposes an application using R software (Shiny package) for the classification of coronavirus sequences through an alignment-free method based on k-mer and k-means methods. This technological advancement has the potential to generate several significant impacts in various spheres. Socially, the application can contribute to the rapid identification and characterization of viruses, being crucial in outbreak and pandemic situations, promoting public health and evidence-based decision-making. Technologically, the creation of this application drives the development of advanced genomic analysis tools, expanding scientific knowledge and practical applications in the field of bioinformatics. From an economic perspective, the use of this application can reduce costs and time associated with viral sequence analysis, benefiting research laboratories and healthcare institutions. Additionally, the ability to quickly classify viral sequences can also have cultural impacts, increasing awareness of the importance of genomics in medicine and understanding infectious diseases, thus influencing health policies and attitudes towards science and technology. In summary, the creation of this application represents a significant advancement that transcends the fields of biology and statistics, positively impacting social, technological, economic, and cultural aspects of society.

LISTA DE FIGURAS

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO

Os seres humanos tiveram um papel importante em zoonoses emergentes ao longo dos séculos, com atividades como agricultura, desmatamento e urbanização (Foster, 2018). Além disso, o surgimento e o ressurgimento de doenças infecciosas virulentas apresentam uma grande ameaça à saúde pública (Gao, 2018). Isso resulta em números expressivos de sequências do genoma viral nos bancos de dados públicos, desempenhando um papel importante na classificação de vírus. Essas sequências são a única forma de classificá-los corretamente (Bao; Chetvernin; Tatusova, 2014).

No entanto, a análise convencional de sequências requer o alinhamento das mesmas, o que pode ser um desafio em se tratando de sequências de genomas completos, que consomem tempo e memória computacional (Zielezinski *et al*., 2017). Os métodos atuais fundamentados em alinhamento usam principalmente programação dinâmica, técnicas progressivas e algoritmos iterativos. Entretanto, o tempo computacional de alinhar simultaneamente um grande grupo de sequências de DNA ou proteínas ainda é enorme, mesmo em supercomputadores (Yu, 2018).

Nesse sentido, nas últimas décadas vem crescendo o número de análises empregando métodos livres de alinhamento que apresentam como principal vantagem a economia de tempo. Os autores Vinga e Almeida (2003) presumiram que o uso desses métodos se tornaria uma ferramenta importante para estudos de classificação e filogenia.

Como os métodos livres de alinhamento não dependem de programação dinâmica, eles são computacionalmente mais baratos e, portanto, adequados para comparações de genomas completos, tornam-se cada vez mais populares com a crescente quantidade de dados (He *et al*., 2021; Guan; Zhao; Yau, 2022).

Neste contexto, dado o surgimento recente do SARS-CoV-2, a fundamentação desta tese considera a relevância em efetuar a classificação ágil de sequências, de modo a auxiliar a elaboração de medidas de contenção do espalhamento de diferentes cepas virais. Tal abordagem implica na identificação de métodos que não comprometam a precisão e reduzam significativamente o tempo, em contraste com os métodos que envolvem alinhamento, cuja operação demanda extenso período para a realização da tarefa.

O objetivo geral deste trabalho é avaliar dois métodos livres de alinhamento, um método baseado em k-mer e o outro, o *Natural Vector*, em comparação com o método tradicional de

alinhamento, considerando tanto a precisão quanto o tempo necessário para gerar o agrupamento. O objetivo específico, que é o principal objetivo da tese, é desenvolver um aplicativo para análise de genomas virais, especialmente para os tipos de coronavírus, utilizando um método baseado em k-mer. O propósito é obter agrupamentos que diferenciem os tipos de coronavírus e também as variantes dentro de cada grupo, mantendo a mesma precisão do método com alinhamento tradicional, porém com um tempo de análise consideravelmente menor.

A estrutura dessa tese foi dividida em duas partes. A primeira parte corresponde aos tópicos que fundamentam a base teórica, e a segunda é constituída por dois artigos científicos. No primeiro artigo foi realizada uma análise comparando os métodos com alinhamento e livre de alinhamento (baseado em k-mer) quanto a classificação usando genomas de SARS-CoV-2. Já no segundo artigo é realizada uma análise de classificação, comparando os métodos com alinhamento e livre de alinhamento (*Natural Vector*) usando genomas de três tipos de coronavírus, SARS-CoV, MERS-CoV e SARS-CoV-2.

2 REFERENCIAL TEÓRICO

O objetivo desta seção é apresentar alguns conceitos voltados para a estruturação desta tese, abordando as etapas que foram realizadas até gerar o agrupamento das sequências, mencionando os genomas dos diferentes tipos de coronavírus, os métodos de alinhamento e os métodos livres de alinhamento, as medidas de distâncias utilizadas e as técnicas de agrupamento empregadas.

2.1 Genomas virais

Os genomas virais representam uma área fundamental no estudo da biologia e da medicina. Compreender a estrutura, função e evolução desses genomas é essencial para uma variedade de aplicações, desde o desenvolvimento de vacinas até o controle de surtos de doenças virais.

Os genomas virais podem ser de fita simples ou dupla, linear ou circular, e em configuração segmentada ou não segmentada. Por exemplo, os coronavírus possuem um genoma de RNA de fita simples e não segmentado.

Todos os vírus pertencentes à mesma família devem ter a mesma classificação de Baltimore, baseada na síntese viral de RNA mensageiro. Além disso, utiliza-se como unidade de medida de comprimento de uma molécula de DNA o número de pares de bases que a formam, denotado por bp (*base pair*) (Okura, 2002).

Os dados no que diz respeito às características estruturais e funcionais do genoma de vários organismos estão sendo acumulados e analisados em laboratórios do mundo todo. O volume de dados genômicos tem crescido a uma taxa contínua e significativa, enquanto suas propriedades e relacionamentos fundamentais ainda não são totalmente conhecidos e estão sujeitos a revisões constantes (Dougherty *et al*., 2005).

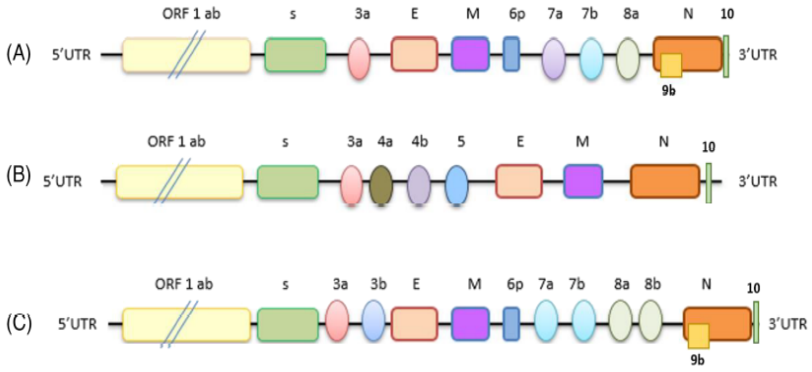
2.2 Tipos de coronavírus

Os coronavírus, uma extensa família de vírus de RNA, têm sido objeto de crescente interesse e preocupação global devido ao seu impacto significativo na saúde humana. Os coronavírus são vírus zoonóticos, hospedados naturalmente por morcegos (Lau *et al*., 2020), com habilidade para infectar uma diversidade de animais (domésticos ou não) e humanos (Khalil; da Silva Khalil, 2020).

Até o presente momento, existem sete coronavírus patogênico para o ser humano, sendo eles: SARS-CoV-2, MERS-CoV, SARS-CoV, HCoV-229E, HCoV-HKU1, HCoV-NL63 e HCoV-OC43 (Pei; Yau, 2021), mas os três principais avaliados neste estudo foram MERS-CoV, SARS-CoV e SARS-CoV-2, tendo este último ganhado maior destaque na pandemia da COVID-19.

Na Figura 2.1 é apresentada a estrutura genômica dos três coronavírus que resultaram em pandemias ou epidemias.

Figura 2.1 – Estrutura genômica do coronavírus. (a) SARS-CoV-2. (b) MERS-CoV. (c) SARS-CoV.



Fonte: Mohamadian *et al*. (2021).

Os coronavírus pertencem à família *Coronaviridae* e são classificados em quatro gêneros: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* e *Deltacoronavirus* (Li, 2016). De acordo com o Comitê International de Taxonomia de Vírus, a espécie SARS coronavírus está dentro do gênero de *Betacoronavirus*, da subfamília de *Coronavirinae*, da família de *Coronaviridae* e da ordem de *Nidovirales* (Yu, 2018).

2.2.1 Síndrome respiratória aguda grave (SARS) e a síndrome respiratória do Oriente Médio (MERS)

Os coronavírus da SARS e MERS representam duas linhagens significativas dentro da vasta família de coronavírus. Eles compartilham a capacidade de infectar o trato respiratório humano, resultando em síndromes respiratórias graves.

A disseminação da síndrome respiratória aguda grave (SARS) em 2002, causada pelo coronavírus conhecido como SARS-CoV, surpreendeu o mundo ao se propagar rapidamente por diferentes continentes. Isso resultou em mais de 8.000 casos de infecção, com uma taxa de mortalidade em torno de 10%, provocando um impacto devastador nas economias de diversas nações e configurando-se como a primeira epidemia do século XXI (Leduc; Barry, 2004).

Por outro lado, a síndrome respiratória do Oriente Médio (MERS) foi a segunda doença epidêmica a ser detectada, surgiu em 2012, com o primeiro caso relatado em março na Arábia Saudita (Who *et al*., 2022).

Nas epidemias de coronavírus provocadas pelo SARS-CoV e MERS-CoV, os vírus provavelmente se originaram de morcegos-ferradura (*Rhinolophus sinicus*) e depois avançaram para outro mamífero hospedeiro de amplificação, como a civeta da palma (*Paguma larvata*) na SARS e o dromedário (*Camelus dromedarius*) na MERS, antes de cruzar a barreira de espécies para infectar seres humanos (Khalil; da Silva Khalil, 2020).

2.2.2 Síndrome respiratória aguda grave 2 (SARS 2)

O coronavírus da SARS 2, identificado como o agente causador da doença conhecida como COVID-19, emergiu como um fenômeno global de extrema relevância para a saúde pública, sendo que a COVID-19 foi a terceira doença altamente epidêmica do século XXI a ser detectada (Mohamadian *et al*., 2021).

O primeiro caso de SARS-CoV-2 foi relatado no final de 2019 na China. Desde então, diversos casos foram notificados em praticamente todos os países e a doença foi declarada Emergência de Saúde Pública de Preocupação Internacional pela Organização Mundial da Saúde (OMS) em 30 de janeiro de 2020. Posteriormente, foi descrita como uma pandemia em março de 2020 (Yuce; Filiztekin, 2021).

A origem da doença no país está relacionada ao vasto território chinês, caracterizado por uma diversidade climática que contribui para uma ampla biodiversidade, incluindo morcegos e vírus. A maioria dos morcegos hospedeiros de coronavírus vive em proximidade com humanos, aumentando o potencial de transmissão para humanos e animais selvagens. Além disso, o hábito de consumir animais selvagens é comum no país devido à crença de que esses animais são mais nutritivos, o que pode contribuir para o aumento da transmissão viral (Khalil; da Silva Khalil, 2020).

A estrutura morfológica do vírus da SARS-CoV-2 é composta por várias partes essenciais para seu funcionamento e reprodução: o envelope viral (E), proteínas de espícula (S), membrana viral (M), proteína do nucleocapsídeo (N), RNA genômico, proteínas não estruturais e enzimas.

O SARS-CoV-2 é um vírus envelopado com um genoma de RNA de fita simples de sentido positivo e pertence ao gênero *Betacoronavírus*, juntamente com o SARS-CoV e o MERS-

CoV, sendo que os CoVs apresentam os maiores genomas (26-32 kb) entre todas as famílias de vírus RNA (Kim *et al*., 2020).

Após a identificação do vírus SARS-CoV-2, diversas variantes e linhagens surgiram e se espalharam globalmente. No Brasil, a epidemia teve início com duas linhagens distintas, B.1.1.28 e B.1.1.33, detectadas no país em fevereiro de 2020 (Candido *et al*., 2020). Adicionalmente, as linhagens P1 e P2 foram inicialmente identificadas em Manaus e no Rio de Janeiro, respectivamente (Rambaut *et al*., 2020).

É importante ressaltar que algumas mutações específicas definem os grupos genéticos virais, também denominados linhagens. O surgimento de mutações adicionais, gerando diferenças dentro de cada grupo genético, dá origem às denominadas variantes (Opas, 2021).

Diferentes nomenclaturas foram designadas às variantes para fins científicos. Um grupo de especialistas convocado pela OMS propôs a utilização de letras do alfabeto grego; por exemplo, a variante P1 também é conhecida como Gamma, enquanto a P2 recebe a denominação de variante Zeta (Mohammadi; Shayestehpour; Mirzael, 2021). O comitê da OMS responsável por monitorar a evolução do SARS-CoV-2, considerando aspectos como transmissibilidade, virulência, modificações fenotípicas e disseminação, classificou as variantes em circulação global como variantes de preocupação e variantes de interesse em saúde pública (Michelon, 2021).

O vírus se espalhou pelo mundo, causando uma pandemia que trouxe consequências catastróficas para a saúde e a economia global. Isso se refletiu em diversos problemas de saúde, como sequelas permanentes que afetaram vários sistemas do corpo, incluindo o respiratório, nervoso, neurocognitivo, metabólico, cardiovascular e gastrointestinal (Maia; Dias, 2020; Rocha *et al*., 2021; Al-Aly; Xie; Bowe, 2021). Além disso, houve o colapso dos sistemas de saúde e a superlotação de hospitais, devido à alta transmissibilidade da doença e à necessidade de internação e suporte de UTI para parte dos infectados (Schuchmann *et al*., 2020). Essa crise também intensificou as dificuldades econômicas, especialmente em países em desenvolvimento, como o Brasil, onde foi observado um impacto particularmente severo (Campos *et al*., 2020; Porsse *et al*., 2020).

Atualmente, após vários estudos, diversas vacinas foram autorizadas e administradas mundialmente, incluindo no Brasil, como a Astrazeneca, Pfizer, Coronavac e Janssen. Cada uma dessas vacinas requer diferentes quantidades de doses, de acordo com seu protocolo específico.

2.3 Métodos com alinhamento

Os dados de sequências de coronavírus, obtidos por meio do sequenciamento genômico, desempenham um papel crucial na compreensão da variabilidade do vírus e na orientação de estratégias de controle. O método de alinhamento de sequências é uma ferramenta fundamental nesse processo, permitindo a comparação sistemática das sequências genéticas dos coronavírus. A combinação de dados de sequenciamento e técnicas de alinhamento representou uma abordagem integrada e poderosa para compreender e enfrentar os desafios dinâmicos apresentados pelos coronavírus SARS-CoV, MERS-CoV e SARS-CoV-2.

O alinhamento consiste em comparar sítios homólogos em diferentes sequências (nucleotídicas ou proteicas) de modo a detectar diferenças entre elas e classificá-las em função de similaridades.

Segue um exemplo de alinhamento para as respectivas sequências: *GCGCATGGATTGAGCGA* e *TGCGCCATTGATGACCA*.

–GCGC–ATGGATTGAGCGA
TGCGCCATTGAT–GACC–A

O objetivo do alinhamento é posicionar na mesma coluna os caracteres idênticos (*match*, cor verde), inserindo inserções (*gaps*, cor azul) em uma delas de forma a obter o máximo possível de *matches*. Na cor vermelha tem-se as diferenças (*mismatch*).

Além disso, as sequências não precisam ter o mesmo número de nucleotídeos para que se realize o alinhamento. O alinhamento de múltiplas sequências é uma ferramenta importante para estrutura da proteína e predição da função, além da inferência filogenética na análise de sequências (Edgar; Batzoglou, 2006).

Os métodos baseados em alinhamento de sequências biológicas de última geração, CLUSTAL, OMEGA e MUSCLE (*Multiple Sequence Comparison by Log-Expectation*), são métodos de referência nessa área (He *et al*., 2020), essenciais na análise comparativa de dados genômicos.

A quantidade necessária de tempo para alinhar as sequências aumenta proporcionalmente ao número e tamanho das sequências, o que pode tornar-se inviável em análises que envolvem um grande número de sequências genômicas.

2.4 Métodos livres de alinhamento

Após considerarmos os métodos tradicionais que envolvem o alinhamento de sequências, é importante explorar abordagens alternativas que se destacam por sua flexibilidade e eficiência: os métodos livres de alinhamento. Os métodos livres de alinhamento buscam identificar padrões globais e estruturais sem a necessidade de alinhamentos prévios.

Abordagens livres de alinhamento incluem qualquer método para avaliar similaridade ou dissimilaridade de sequência que não aplica nem produz alinhamento de sequência em nenhuma etapa do algoritmo; ao contrário disso, eles usam extração de recursos para extrair as informações necessárias das sequências (He *et al*., 2020).

Segundo Han e Cho (2019), os métodos livres de alinhamento podem ser divididos em quatro categorias: k-mer, string comum, teoria da informação e métodos baseados em representação gráfica. Na literatura, existem muitos métodos livres de alinhamento; entretanto, neste trabalho o interesse é avaliar sequências genômicas do coronavírus por meio dos métodos baseado em k-mer e o *Natural Vector*.

As metodologias livres de alinhamento superam com êxito as limitações sérias das formas baseadas em alinhamento, especialmente para o tempo de computação e espaço de armazenamento. Os métodos para análise de similaridade de sequência comumente dependem de um alinhamento de múltiplas sequências, o que geralmente requer um tempo longo para obter resultados, desse modo, métodos livres de alinhamento foram propostos para superar esta ineficácia (Guan; Zhao; Yau, 2022).

As abordagens livres de alinhamento são matematicamente bem fundamentadas nas áreas de álgebra linear, teoria da informação e estatística, sendo que a lógica por trás desses métodos é simples: sequências semelhantes compartilham palavras semelhantes/k-mers (subsequências de comprimento k) e operações matemáticas com as ocorrências de palavras fornecem uma boa medida relativa de dissimilaridade da sequência. Entretanto, ainda é difícil afirmar qual método livre de alinhamento pode ser particularmente adequado para uma determinada tarefa (Zielezinski *et al*., 2017).

A seguir são apresentados dois métodos livres de alinhamento, o primeiro baseado em k-mer, e o segundo denominado *Natural Vector*.

2.4.1 Métodos baseados em k-mer

De acordo com Huang (2016), um k-mer é um segmento de k nucleotídeos consecutivos da sequência do genoma. Uma sequência de DNA usando o k-mer é composta por todas as possíveis subsequências contíguas de comprimento k de quatro nucleotídeos: Adenina (A), Citosina (C), Guanina (G) e Timina (T) (Yin; Yau, 2015), sendo que o k-mer é sempre extraído no sentido 5'-3' (uma referência à direção que ocorre a síntese ou leitura de ácidos nucleicos como o DNA ou RNA) da cadeia de DNA original.

Dada uma determinada sequência $CGTAGTAC$, é possível avaliar diferentes valores de k, conforme o exemplo abaixo:

$$\begin{cases} k=1:\{C,G,T,A,G,T,A,C\}; \\ k=2:\{CG,GT,TA,AG,GT,TA,AC\}; \\ k=3:\{CGT,GTA,TAG,AGT,GTA,TAC\}; \\ k=4:\{CGTA,GTAG,TAGT,AGTA,GTAC\}; \\ k=5:\{CGTAG,GTAGT,TAGTA,AGTAC\}; \\ k=6:\{CGTAGT,GTAGTA,TAGTAC\}; \\ k=7:\{CGTAGTA,GTAGTAC\}; \\ k=8:\{CGTAGTAC\}. \end{cases}$$

Considerando então a matriz de contagens k-mer a partir dessa sequência $CGTAGTAC$, quando $k=2$ teremos 16 (4^2) combinações que são: $AA,AC,AG,AT,CA,CC,CG,CT,GA,GC$, GG,GT,TA,TC,TG,TT. A matriz de contagens k-mer é descrita da seguinte forma:

$$\begin{bmatrix} AA & AC & AG & AT & CA & CC & CG & CT & GA & GC & GG & GT & TA & TC & TG & TT \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 2 & 2 & 0 & 0 & 0 \end{bmatrix}.$$

A ideia geral do método baseado em k-mer é gerar todas as possíveis subsequências de comprimento k a partir das sequências biológicas, onde k representa o tamanho do k-mer. Isso é realizado por meio de uma janela deslizante que percorre as sequências, identificando e contando cada k-mer encontrado. Posteriormente, esses valores de contagem podem ser comparados entre os genomas analisados para identificar padrões ou similaridades.

Alguns trabalhos em que o método k-mer foi usado: para montagem de genomas e metagenomas (é um processo fundamental na biologia computacional que envolve a reconstrução de sequências de DNA ou RNA a partir de fragmentos curtos de sequenciamento), em que os

autores desenvolveram um novo conjunto de algoritmos, chamados de “Velvet”, para manipular gráficos de Bruijn (Zerbino; Birney, 2008), para alinhamento de sequências, no qual também foi desenvolvido o programa “BLASTN” (Chen *et al*., 2015), para classificação e caracterização de amostras de metagenoma (Ainsworth *et al*., 2017). Os métodos baseados em k-mer também apresentaram alguns resultados promissores para comparação de sequências de DNA e proteínas (Yu, 2018).

De acordo com Chikhi e Medvedev (2014), um grande valor de k é desejado, mas por outro lado, quanto maior o valor de k, maiores são as chances de um k-mer apresentar erros; portanto, tornar k muito grande diminui o número correto de k-mers presentes nos dados. O mesmo autor defende que na prática, k é frequentemente escolhido com base em experiência com conjuntos de dados semelhantes.

Ademais, a mineração de dados é definida por Linoffy e Berry (2011) como sendo a exploração e a análise por meio automático ou semiautomático de grandes quantidades de dados, a fim de descobrir padrões e regras significativas. Neste sentido, o objetivo do cálculo de frequência de k-mers é reconhecer a repetição de padrões específicos, os k-mers, dentro de uma grande quantidade de dados. Portanto, o cálculo da frequência de repetição de k-mers está contido no conjunto de problema de mineração de dados.

2.4.2 Método *Natural Vector*

Entre as técnicas livres de alinhamento, o método *Natural Vector* (*NV*) em sequências de DNA/proteína é utilizado para caracterizar genomas e proteínas (Yu, 2018). O *NV* de uma sequência é baseado nas bases nitrogenadas da sequência de DNA: Adenina (A), Citosina (C), Guanina (G) e Timina (T).

Para definir um *NV*, primeiro considera-se $S = (s_1, s_2, \ldots, s_n)$ uma sequência de DNA de tamanho n, ou seja, cada $s_i \in \{\text{A, C, G, T}\}$, para $i = 1, \ldots, n$. Para $k = A, C, G, T$, define-se a função $w_k(\cdot) : \{\text{A, C, G, T}\} \to \{0, 1\}$ de forma que, para cada elemento s_i da sequência, tem-se:

$$w_k(s_i) = \begin{cases} 1, & \text{se } s_i = k \\ 0, & \text{caso contrário.} \end{cases}$$

Com isso, as estatísticas utilizadas para compor o *Natural Vector* da sequência S são (Yu *et al*., 2013):

a) O número de letras k na sequência S:

$$n_k = \sum_{i=1}^{n} w_k(s_i). \tag{2.1}$$

b) A posição média da letra k na sequência S:

$$\mu_k = \sum_{i=1}^{n} i \frac{w_k(s_i)}{n_k}. \tag{2.2}$$

c) O j-ésimo momento central normalizado da letra k na sequência S:

$$D_j^k = \sum_{i=1}^{n} \frac{(i-\mu_k)^j w_k(s_i)}{n_k^{j-1} n^{j-1}}, j = 2,3,\ldots,n_k. \tag{2.3}$$

O *Natural Vector* é obtido concatenando o primeiro grupo de parâmetros (Equação 2.1) e o segundo grupo de parâmetros (Equação 2.2) para os momentos centrais normalizados (Equação 2.3), sendo o NV de uma sequência S de DNA definido por um vetor de dimensão $n+4$:

$$NV(S)=(n_A,\mu_A,D_2^A,\ldots,D_{n_A}^A,n_C,\mu_C,D_2^C,\ldots,D_{n_C}^C,n_G,\mu_G,D_2^G,\ldots,D_{n_G}^G,n_T,\mu_T,D_2^T,\ldots,D_{n_T}^T). \tag{2.4}$$

Quanto maiores forem os momentos centrais normalizados incluídos no NV (Equação 2.4), tais como 16 ($j=3$) ou 20 ($j=4$) dimensões, as relações próximas e distantes entre esses vírus permanecem inalteradas no contexto da distância Euclidiana, sendo que 12 ($j=2$) e 16 ($j=3$) dimensões ou mesmo vetores naturais de dimensões superiores produzem os mesmos resultados da classificação (Yu, 2018).

Segundo Yu *et al*. (2013), a abordagem foi nomeada como *Natural Vector* pois os três grupos de parâmetros usados no NV são naturais, baseados nas quantidades e distribuições de quatro nucleotídeos (A, C, G, T) da sequência original do DNA. Além disso, o método *Natural Vector* ajuda a determinar as relações evolutivas e de classificação entre os vírus em qualquer nível, como classe, família, subfamília, gênero e espécie de Baltimore (Yu, 2018).

A ideia central do NV é que se os genes ou proteínas são biologicamente próximos uns dos outros, seus vetores naturais estão próximos e assim, com base em características estatísticas, as sequências podem ser representadas como pontos em um espaço de alta dimensão (Lv *et al*., 2020). Isto é, a caracterização do *Natural Vector* constrói uma correspondência um para um entre as sequências de genoma e os vetores numéricos (Deng *et al*., 2011).

2.5 Distância euclidiana

Após utilizar o método com alinhamento ou livre de alinhamento, o passo seguinte pode envolver a aplicação de medidas para calcular a distância entre as sequências, buscando quantificar identidades ou similaridades entre elas. Para este propósito, diversas medidas de distância podem ser empregadas, tais como as distâncias Euclidiana, Mahalanobis, Manhattan, entre outras.

A semelhança entre os objetos é quantificada por meio de uma medida de proximidade, que engloba tanto as medidas de similaridades quanto as de dissimilaridades (Ferreira, 2018). Na avaliação da similaridade, um valor mais elevado indica maior semelhança entre os objetos. Em contraste, nas medidas de dissimilaridade, um valor mais alto implica em uma menor semelhança entre os objetos.

Para a análise de dissimilaridades das sequências dos genomas, pode-se utilizar, por exemplo, a distância Euclidiana. De acordo com Yu *et al*. (2013), a distância Euclidiana entre dois pontos representa a distância biológica dos vírus correspondentes. A distância Euclidiana em $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ é definida pela função d (Yin; Yau, 2015):

$$d\left((x_1,\ldots,x_n),(y_1,\ldots,y_n)\right) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}, \tag{2.5}$$

em que $x_1,\ldots,x_n$ e $y_1,\ldots,y_n$ são armazenados em termos de vetores, sendo que um valor menor da distância Euclidiana significa maior grau de similaridade entre x e y (Phannachitta, 2017).

Em síntese, a distância Euclidiana é mais apropriada para grupos de variáveis que possuem escalas similares, pois caso contrário, variáveis com maior variabilidade dominarão a classificação das distâncias (Ferreira, 2018).

2.6 Distância p

Entre os modelos de distância, outro exemplo é a distância p, definida como a razão entre o número de nucleotídeos diferentes (nd) e o número total de nucleotídeos comparados (nt):

$$p = \frac{nd}{nt}.$$

Se o valor de $p < 0,15$, isso indica que o modelo de distância é adequado; caso contrário, outro modelo deve ser usado.

2.7 Análise de agrupamento

Após considerar as medidas de distância como ferramentas fundamentais na avaliação da similaridade de sequências, torna-se pertinente explorar a análise de agrupamento como uma estratégia poderosa na interpretação dessas relações. Ao utilizar as medidas de distância, a análise de agrupamento torna-se um componente essencial, permitindo a formação de grupos correlacionados com base na proximidade entre essas sequências.

Desse modo, a técnica de agrupamentos é uma técnica estatística usada para classificar elementos em grupos, de forma que elementos dentro de um mesmo grupo sejam muito parecidos, e os elementos em diferentes *clusters* sejam distintos entre si. A classificação de sequências é outro campo que se beneficia da combinação de diversas abordagens livres de alinhamento (Zielezinski *et al*., 2017).

A comparação de sequências biológicas é um passo fundamental para inferir o parentesco entre vários organismos e a semelhança funcional de seus componentes (Kimothi *et al*., 2016).

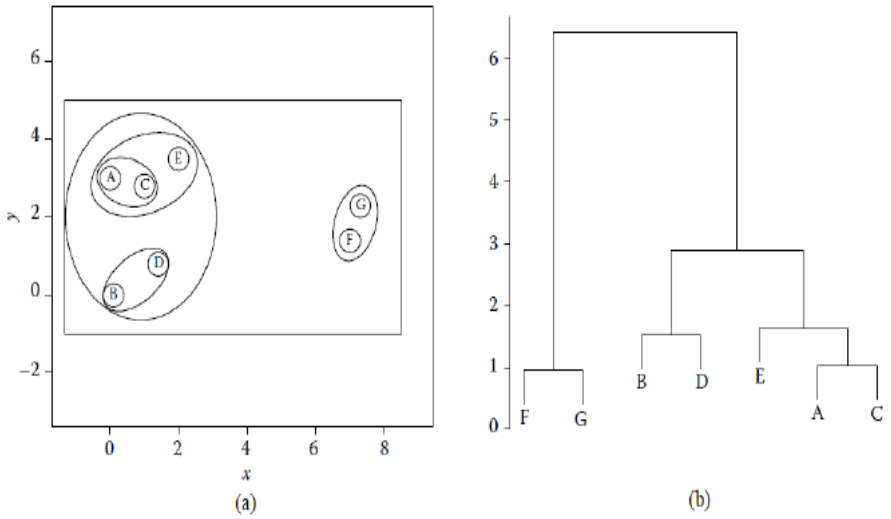
As técnicas de agrupamento são muito úteis, pois revelam padrões existentes no conjunto de dados que não poderiam ser detectados com inspeções visuais simples da amostra e de suas estatísticas descritivas (Ferreira, 2018).

A seguir, são apresentados dois métodos distintos de agrupamento: o hierárquico e o não hierárquico. Além disso, é importante observar que diversos algoritmos têm a capacidade de gerar tanto árvores enraizadas quanto não enraizadas com base em matrizes de distância. Por exemplo, o algoritmo de junção de vizinhos (*neighbor-joining*) (Saitou; Nei, 1987) resulta em árvores não enraizadas, enquanto o algoritmo UPGMA (Sokal; Michener, 1958) gera árvores enraizadas.

2.7.1 Agrupamento hierárquico

Segundo Abu-Jamous, Fa e Nandi (2015), um método de agrupamento hierárquico é um procedimento para transformar uma matriz de proximidade em uma partição aninhada, que pode ser representada graficamente por uma árvore (Figura 2.2 (b)).

Figura 2.2 – Agrupamento hierárquico. (a) hierarquia de conjuntos. (b) dendrograma.



Fonte: D'urso *et al*. (2016).

Quando duas amostras estão próximas (por exemplo, os conjuntos F e G na Figura 2.2), elas devem ter valores semelhantes para o valor medido. Portanto, quanto maior a proximidade entre as medidas relacionadas às amostras, maior a similaridade entre elas. O dendrograma, que possui uma estrutura em árvore, hierarquiza essa semelhança para que se tenha uma visão bidimensional da semelhança de todo o conjunto de amostras utilizadas no estudo. Em outras palavras, o dendrograma organiza determinados fatores e variáveis (Abonyi; Feil, 2007).

Existem dois tipos de métodos de agrupamento hierárquico: aglomerativo e divisivo. O agrupamento aglomerativo inicia com elementos únicos e os agrega em *clusters* maiores. Por outro lado, o agrupamento divisivo começa com o conjunto completo e o divide em partições menores. Para o agrupamento aglomerativo, temos os métodos de ligação simples (mínima distância ou vizinho mais próximo), ligação completa (máxima distância ou vizinho mais distante) e ligação média (distância média).

2.7.1.1 Agrupamento hierárquico da ligação média - UPGMA

De acordo com Hua *et al*. (2017), o algoritmo do Método de Grupo de Pares Não Ponderado com Média Aritmética (UPGMA) para a construção da árvore filogenética utiliza uma matriz de semelhança de pares, também conhecida como matriz de distância, que descreve as semelhanças entre todos os pares possíveis de dados para criar a árvore de classificação.

No UPGMA, o critério de cálculo da distância entre um grupo recém-formado e outros grupos pré-existentes é a média de todas as distâncias entre os objetos dos dois grupos. Desta

forma, se fundirmos o grupo r com n_r objetos ao grupo s com n_s objetos, a distância entre esse grupo e um grupo pré-existente t com n_t objetos é dada por (Ferreira, 2018):

$$d_{(rs),t} = \frac{1}{n_r n_t + n_s n_t} \left[\sum_{i \in r} \sum_{k \in t} d_{ik} + \sum_{j \in s} \sum_{k \in t} d_{jk} \right] \tag{2.6}$$

$$= \frac{n_r}{n_r + n_s} \cdot \frac{1}{n_r n_t} \sum_{i \in r} \sum_{k \in t} d_{ik} + \frac{n_s}{n_r + n_s} \cdot \frac{1}{n_s n_t} \sum_{j \in s} \sum_{k \in t} d_{jk}, \tag{2.7}$$

ou seja, representa a média de todas as distâncias entre os objetos dos grupos (rs) e os objetos do grupo t. Analisando a segunda parte da equação 2.7, quando $n_r = n_s$, percebe-se que a distância entre os grupos (rs) e t é igual à média aritmética das médias aritméticas das distâncias entre os objetos dos grupos r e t, e entre os objetos dos grupos s e t.

O algoritmo UPGMA pode ser aplicado a partir dos seguintes passos:

a) Nesta etapa, a matriz de distância (D) é inicializada preenchendo suas entradas com valores das distâncias entre os pares de grupos $(S_1, \ldots, S_n)$ correspondentes. Um valor menor indica um relacionamento mais próximo e, portanto, uma maior semelhança. Assim, cada elemento na matriz registra o valor da distância entre os elementos da linha e da coluna, conforme descrito a seguir.

$$D_{\mathrm{S}} = \begin{bmatrix} D(S_1,S_1) & D(S_1,S_2) & \ldots & D(S_1,S_N) \\ D(S_2,S_1) & D(S_2,S_2) & \ldots & D(S_2,S_N) \\ \vdots & \vdots & \vdots & \vdots \\ D(S_N,S_1) & D(S_N,S_2) & \ldots & D(S_N,S_N) \end{bmatrix}. \tag{2.8}$$

b) Na segunda etapa, deve-se identificar na matriz D o par de grupos mais similar, ou seja, o par com menor distância entre todos os pares. Se houver mais de um elemento que registra o valor mínimo, será selecionado aleatoriamente um deles. Nesse exemplo, vamos supor que o valor mínimo corresponda a $D(S_N, S_2)$, que representam os grupos N e 2.

c) Após o par de grupos com a menor distância ser agrupado em um novo ramo, a matriz de distância deve ser atualizada. As linhas e colunas referentes aos grupos N e 2 são removidas da matriz, e uma nova linha e uma nova coluna são adicionadas para registrar as novas distâncias entre o grupo recém-formado N2 e os grupos restantes.

d) Os passos b) e c) são repetidos de forma consistente até que todos os grupos sejam mesclados em um único grupo, resultando na construção completa da árvore. Consequentemente, a matriz de distância reduz suas dimensões de linha e coluna em uma unidade a cada iteração.

Todos os métodos hierárquicos buscam o par de objetos com a menor distância para formar um grupo, identificando os dois objetos mais próximos para que possam ser agrupados (Ferreira, 2018). Uma das vantagens em utilizar o algoritmo hierárquico é não ser necessário informar *a prior* o número de grupos nos quais os dados devem ser divididos.

Segundo Rencher (2002), muitos estudos mostraram que o método de ligação média possui um bom desempenho de forma geral. No entanto, esse desempenho pode variar dependendo dos dado. Uma boa estratégia é testar vários métodos e considerar aquele que melhor se alinha com algum tipo de agrupamento natural.

2.7.2 Agrupamento não hierárquico

Existem muitos métodos não hierárquicos baseados em misturas de distribuição, estimação de densidades e partições, sendo esse último o mais usual. Além disso, o método das k-médias é o mais popular entre os métodos de partição (Ferreira, 2018).

Nos métodos não hierárquicos, o processo é aplicado à matriz de dados inicial e não à matriz de dissimilaridades, como no método hierárquico. Além disso, os agrupamentos não hierárquicos buscam a partição de n objetos em k grupos.

O método k-médias (ou *k-means*) é um algoritmo de agrupamento que organiza dados em grupos ou *clusters* com base em características semelhantes. De forma simplificada, o algoritmo k-médias pode ser aplicado usando os seguintes passos:

a) Primeiro, alocar aleatoriamente os n objetos aos k grupos e calcular os seus centroides a partir do conjunto de dados.

b) Para cada dado (ou por exemplo um genoma) específico, atribuir ao *cluster* cujo centroide é o mais próximo.

c) Determinar os centroides reais para os grupos. É necessário recalcular os novos centroides para cada *cluster*, geralmente tomando a média dos pontos atribuídos a esse *cluster*.

d) Repetir os passos b) e c), até que não ocorram mais mudanças de objetos de um grupo para outro.

e) Os centroides finais representam os centros dos *clusters* otimizados. As observações são atribuídas aos *clusters* finais. Quando o centroide não precisa mais ser reposicionado, isso significa que o algoritmo convergiu.

Uma desvantagem em usar o algoritmo k-médias é a necessidade de informar *a prior* o número de grupos (k) em que os dados se encontram divididos. Além disso, esse método é sensível à escolha inicial dos grupos ou de seus centroides (Ferreira, 2018). Entretanto, as vantagens em usar o método k-médias são a simplicidade e a facilidade de implementar, além de ser computacionalmente rápido e eficiente.

É importante enfatizar que todas as análises associadas ao método com alinhamento foram realizadas no *software* MEGA (Kumar *et al*., 2018), empregando a distância-p e o método UPGMA. Com relação aos métodos livres de alinhamento, as análises foram realizadas no *software* `R` (R Core Team, 2023), sendo que para o método *Natural Vector* foi usado a distância Euclidiana e o método UPGMA. Para o método baseado em k-mer foi usado a matriz de contagens k-mer e o algoritmo k-médias.

3 CONCLUSÃO

As metodologias aplicadas para a análise de agrupamento de genomas apresentaram resultados satisfatórios, semelhantes aos obtidos com o alinhamento convencional das sequências genômicas.

O primeiro método, baseado em k-mer, conseguiu classificar as sequências nas diferentes linhagens e variantes do SARS-Cov-2 em um intervalo de tempo significativamente menor do que o método de alinhamento.

O segundo método, *Natural Vector*, conseguiu separar corretamente as sequências quanto aos tipos de coronavírus (SARS-CoV-2, SARS-CoV e MERS-CoV), em um tempo menor em comparação ao alinhamento convencional. No entanto, ao tentar classificar as sequências nas diferentes linhagens e variantes, o método NV não agrupou corretamente, misturando os grupos correspondentes.

Ambos os métodos mostraram-se úteis para aplicações na análise de agrupamentos, mantendo a precisão e obtendo resultados rapidamente. Entretanto, o método NV não é recomendado quando o objetivo é agrupar e/ou separar as variantes e linhagens do SARS-CoV-2. Portanto, as vantagens em utilizar os métodos livres de alinhamento incluem a precisão na classificação, além da redução do tempo de análise, que se torna relativamente menor em comparação ao processo de alinhamento.

Além disso, o aplicativo desenvolvido em *Shiny* tem como finalidade acelerar as análises relacionadas à classificação utilizando genomas do coronavírus, especialmente para futuras epidemias, contribuindo para o desenvolvimento de estratégias de diagnóstico mais eficazes. O aplicativo pode ser acessado por meio do link: <https://labinufopa.shinyapps.io/Kvid19/>.

Como trabalho futuro, seria interessante realizar um estudo usando métodos livres de alinhamento para classificar outros tipos de vírus que afetaram e continuam afetando a humanidade até os dias atuais.

Por fim, ressalta-se que no Apêndice A tem-se exemplos em `R` dos códigos utilizados em cada um dos artigos.

REFERÊNCIAS

ABONYI, J.; FEIL, B. **Cluster Analysis for Data Mining and System Identification**. Germany: Springer Science & Business Media, 2007.

ABU-JAMOUS, B.; FA, R.; NANDI, A. K. **Integrative Cluster Analysis in Bioinformatics**. United Kingdom: John Wiley & Sons, 2015.

AINSWORTH, D. *et al*. k-SLAM: accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. **Nucleic Acids Research**, v. 45, n. 4, p. 1649-1656, 2017.

AL-ALY, Z.; XIE, Y.; BOWE, B. High-dimensional characterization of post-acute sequelae of covid-19. **Nature**, v. 594, n. 7862, p. 259-264, 2021.

BAO, Y.; CHETVERNIN, V.; TATUSOVA, T. Improvements to pairwise sequence comparison (pasc): a genome-based web tool for virus classification. **Archives of Virology**, v. 159, p. 3293-3304, 2014.

CAMPOS, G. W. D. S. *et al*. O pesadelo macabro da Covid-19 no Brasil: entre negacionismos e desvarios. **Trabalho, Educação e Saúde**, v. 18, n. 3, p. 1-5, 2020.

CANDIDO, D. S. *et al*. Evolution and epidemic spread of SARS-CoV-2 in Brazil. **Science**, v. 369, n. 6508, p. 1255-1260, 2020.

CHEN, Y. *et al*. High speed BLASTN: an accelerated megaBLAST search tool. **Nucleic Acids Research**, v. 43, n. 16, p. 7762-7768, 2015.

CHIKHI, R.; MEDVEDEV, P. Informed and automated k-mer size selection for genome assembly. **Bioinformatics**, v. 30, n. 1, p. 31-37, 2014.

DENG, M. *et al*. A novel method of characterizing genetic sequences: genome space with biological distance and applications. **PloS One**, v. 6, n. 3, p. 1-9, 2011.

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

D'URSO, P. *et al*. **Handbook of Cluster Analysis**. New York: Chapman and Hall, 2016.

EDGAR, R. C.; BATZOGLOU, S. Multiple sequence alignment. **Current Opinion in Structural Biology**, v. 16, n. 3, p. 368-373, 2006.

FERREIRA, D. F. **Estatística Multivariada.** 3 ed. Lavras-MG: UFLA, 2018. ISBN: 9788581270630.

FOSTER, J. E. Viruses as pathogens: animal viruses affecting wild and domesticated species. **Viruses: Molecular Biology, Host Interactions, and Applications to Biotechnology**. United Kingdom: Elsevier, p. 189-216, 2018.

GAO, G. F. From "A" IV to "Z" IKV: Attacks from Emerging and Re-emerging Pathogens. **Cell**, v. 172, n. 6, p. 1157-1159, 2018.

GUAN, M.; ZHAO, L.; YAU, S. S.-T. Classification of protein sequences by a novel alignment-free method on bacterial and virus families. **Genes**, v. 13, n. 10, p. 1-12, 2022.

HAN, G.-B.; CHO, D.-H. Genome classification improvements based on k-mer intervals in sequences.**Genomics**, v. 111, n. 6, p. 1574-1582, 2019.

HE, L. *et al*. A novel alignment-free method for HIV-1 subtype classification. **Infection, Genetics and Evolution**, v. 77, p. 104080, 2020.

HE, L. *et al*. Alignment-free sequence comparison for virus genomes based on location correlation coefficient. **Infection, Genetics and Evolution**, v. 96, p. 1-13, 2021.

HUA, G.-J. *et al*. MGUPGMA: A Fast UPGMA Algorithm With Multiple Graphics Processing Units Using NCCL. **Evolutionary Bioinformatics**, v. 13, p. 1-7, 2017.

HUANG, H.-H. An ensemble distance measure of k-mer and Natural Vector for the phylogenetic analysis of multiple-segmented viruses. **Journal of Theoretical Biology**, v. 398, p. 136-144, 2016.

KHALIL O. A. K.; DA SILVA KHALIL S. SARS-CoV-2: taxonomia, origem e constituição. **Revista de Medicina**, v. 99, n. 5, p. 473-479, 2020.

KIM, D. *et al*. The Architecture of SARS-CoV-2 Transcriptome. **Cell**, v. 181, n. 4, p. 914-921, 2020.

KIMOTHI, D. *et al*. Distributed representations for biological sequence analysis. **arXiv**, p. 1-7, 2016.

KUMAR, S. *et al*. Mega x: Molecular evolutionary genetics analysis across computing platforms. **Molecular Biology and Evolution**, v. 35, n. 6, p. 1547-1549, 2018.

LAU, S. K. P. *et al*. Possible Bat Origin of Severe Acute Respiratory Syndrome Coronavirus 2. **Emerging Infectious Diseases**, v. 26, n. 7, p. 1542-1547, 2020.

LEDUC, J. W.; BARRY, M. A. SARS, the first pandemic of the 21st century. **Emerging Infectious Diseases**, v. 10, n. 11, p. e26, 2004.

LI, F. Structure, function, and evolution of coronavirus spike proteins. **Annual Review of Virology**, v. 3, n. 1, p. 237-261, 2016.

LINOFF, G. S.; BERRY, M. J. **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management**. 3 ed. Indiana: John Wiley & Sons, 2011.

LV, H. *et al*. Evaluation of Different Computational Methods on 5-methylcytosine Sites Identification. **Briefings in Bioinformatics**, v. 21, n. 3, p. 982-995, 2020.

MAIA, B. R.; DIAS, P. C. Ansiedade, depressão e estresse em estudantes universitários: o impacto da COVID-19. **Estudos de Psicologia (Campinas)**, v. 37, p. 1-8, 2020.

MICHELON, C. M. Principais variantes do SARS-CoV-2 notificadas no Brasil. **Revista Brasileira de Análises Clínicas**, v. 53, n. 2, p. 109-116, 2021.

MOHAMADIAN, M. *et al*. Covid-19: Virology, biology and novel laboratory diagnosis. **The Journal of Gene Medicine**, v. 23, n. 2, p. e3303, 2021.

MOHAMMADI, M.; SHAYESTEHPOUR, M.; MIRZAEI, H. The impact of spike mutated variants of SARS-CoV2 [Alpha, Beta, Gamma, Delta, and Lambda] on the efficacy of subunit recombinant vaccines. **Brazilian Journal of Infectious Diseases**, v. 25, p. 1-9, 2021.

OPAS: Organização Pan-Americana da Saúde. 2021. Disponível em: https://iris.paho.org/bitstream/handle/10665.2/53376/EpiUpdate24March2021_por.pdf?sequence=1&isAllowed=y. Acesso em: 20 dez. 2021.

OKURA, V. K. **Bioinformática de Projetos Genoma de Bactérias**. 2002. 115 p. Dissertação (Mestrado em Ciência da Computação) — Universidade Estadual de Campinas, Campinas, 2002.

PEI, S.; YAU, S. S.-T. Analysis of the genomic distance between bat coronavirus ratg13 and SARSs-CoV-2 reveals multiple origins of covid-19. **Acta Mathematica Scientia**, v. 41, n. 3, p. 1017-1022, 2021.

PHANNACHITTA, P. Robust comparison of similarity measures in analogy based software effort estimation. **2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)**, p. 1-7, 2017.

PORSSE, A. A. *et al*. Impactos Econômicos da COVID-19 no Brasil. **Nota Técnica NEDUR-UFPR No 01-2020**, Núcleo de Estudos em Desenvolvimento Urbano e Regional (NEDUR) da Universidade Federal do Paraná, v. 1, p. 1-21, 2020.

R CORE TEAM. **R: a language and environment for statistical computing**. 2023. Disponível em: https://www.R-project.org/. 2023. Acesso em: 25 jan. 2023.

RAMBAUT, A. *et al*. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. **Nature Microbiology**, v. 5, n. 11, p. 1403-1407, 2020.

RENCHER, A. C. **Methods of Multivariate Analysis 2ed.** New York: John Wiley, 2002.

ROCHA, M. S. *et al*. Ansiedade, depressão e estresse em estudantes universitários durante a pandemia do COVID-19. **Brazilian Journal of Development**, v. 7, n. 8, p. 80959-80970, 2021.

SAITOU, N.; NEI, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. **Molecular Biology and Evolution**, v. 4, n. 4, p. 406-425, 1987.

SCHUCHMANN, A. Z. *et al*. Isolamento social vertical x Isolamento social horizontal: os dilemas sanitários e sociais no enfrentamento da pandemia de COVID-19. **Brazilian Journal of Health Review**, v. 3, n. 2, p. 3556-3576, 2020.

SOKAL, R. R.; MICHENER, C. A statistical method for evaluating systematic relationships. **University Kans Sci Bull**, v. 38, p. 1409-1438, 1958.

VINGA, S.; ALMEIDA, J. Alignment-free sequence comparison - a review. **Bioinformatics**, v. 19, n. 4, p. 513-523, 2003.

WHO *et al*. COVID-19 Weekly Epidemiological Update. **World Health Organization**, p. 1-9, 2022.

YIN, C.; YAU, S. S.-T. An improved model for whole genome phylogenetic analysis by Fourier transform. **Journal of Theoretical Biology**, v. 382, p. 99-110, 2015.

YU, C. *et al*. Real Time Classification of Viruses in 12 Dimensions. **PloS One**, v. 8, n. 5, p. 1-10, 2013.

YU, C. Natural Vector Method for Virus Phylogenetic Classification: A Mini-Review. **Current Bioinformatics**, v. 13, n. 4, p. 332-336, 2018.

YÜCE, M.; FILIZTEKIN, E.; ÖZKAYA, K. G. Covid-19 diagnosis - a review of current methods. **Biosensors and Bioelectronics**, v. 172, p. 1-15, 2021.

ZERBINO, D. R.; BIRNEY, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. **Genome Research**, v. 18, n. 5, p. 821-829, 2008.

ZIELEZINSKI, A. *et al*. Alignment-free sequence comparison: Benefits, applications, and tools. **Genome Biology**, v. 18, p. 1-17, 2017.

SEGUNDA PARTE - ARTIGOS

ARTIGO 1 - *Advancing Viral Genome Classification: Assessing the Efficiency and Accuracy of the Alignment-Free K-mer Method in Emerging Pandemics*

Redigido conforme as normas da revista *Anais da Academia Brasileira de Ciências* (versão em processo de revisão).

Advancing Viral Genome Classification: Assessing the Efficiency and Accuracy of the Alignment-Free K-mer Method in Emerging Pandemics

Abstract

In January 2020, the World Health Organization officially recognized the novel coronavirus outbreak as a worldwide public health crisis, marking the onset of what would become the COVID-19 pandemic. Since then, extensive research efforts have been initiated to describe the virus, understand mutation patterns, transmission dynamics, and develop vaccines. Many of these studies require the classification of various virus strains, which is crucial for accurately characterizing the variants that emerged during the pandemic. However, classifying these strains requires methods for comparing genomic sequences, typically involving sequence alignment, a time-consuming process. In our study, we focused on assessing the accuracy and time efficiency of the k-mer method, which does not rely on sequence alignment but can enhance genomic comparisons. Using data from the National Center for Biotechnology Information's SARS-CoV-2 website, we classified 17 complete genomes from different groups detected or emerging in Brazil, employing both alignment-based and k-mer approaches. Both methods yielded identical classifications, but the k-mer method outperformed significantly, being 97% faster. Therefore, we advocate for the use of the k-mer method in viral genome analysis, particularly during emerging pandemics. Its combination of speed and accuracy can greatly expedite responses to new viral threats.

Introduction

Toward the end of 2019, in Wuhan, China, the global media reported the first case of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the coronavirus disease (COVID-19). Instances of the illness have been extensively documented in nearly every nation, prompting the World Health Organization (WHO) to declare COVID-19 a matter of international public health concern on January 30, 2020. Following this, in March 2020, the ailment was formally classified as a pandemic (Silva & Lima 2021). Currently, COVID-19 has emerged as the most notable health crisis of the 21st century, causing around 6.9 million fatalities globally (WHO 2023).

Based on molecular and phylogenetic characteristics, the International Committee on Taxonomy of Viruses (ICTV) classified SARS-CoV-2 into the order *Norividaes*, family *Coronaviridae*, subfamily *Coronavirinae*, genus *Betacoronavirus* and subgenus *Sarbecovirus* (Boni et al. 2020). The SARS-CoV-2 is an enveloped virus with a positive-sense single-stranded RNA genome of approximately 30 kb (Kim et al. 2020). Previously, two other types of coronaviruses had already emerged: SARS-CoV in 2002 and Middle East respiratory syndrome (MERS)-CoV in 2012 in the Middle East, derived from genetic elements between coronaviruses in camels and bats (Mohamadian et al. 2021). Genomic sequence comparison studies unveiled a 96% similarity between the SARS-CoV-2 virus and a bat CoV RaTG13, raising suspicions that bats could potentially serve as the natural hosts of the virus (Zhou et al. 2020).

COVID-19 has a high ability to spread and causes an extended range of symptoms, including fever, fatigue, dry cough, and shortness of breath (Sousa et al. 2020). The interplay between its distinct transmissibility and relatively moderate impact on many people facilitated its swift worldwide spread, leading to the emergence of diverse variants and lineages worldwide (Chen et al. 2020). These variants can be distinguished based on

their distinct symptom patterns and transmission rates, with genome sequence analysis proving to be the most effective method for their classification.

Accurate classification of virus variants has empowered authorities to make informed decisions during each phase of the pandemic (Flores-Vega et al. 2022). Therefore, precise and rapid sequence analyses, particularly during health emergencies, are crucial for effectiveness. Identifying variants through whole genome sequencing often involves sequence alignment to identify differences between sequences and group them accordingly. However, this process is both time-consuming and computationally intensive.

The issue of time consumption associated with alignment-based methods has prompted the development of a genome assembly tool to pinpoint phylogenetic clusters among coronaviruses (Cleemput et al. 2020), the employment of spatial analysis to identify high-risk COVID-19 areas (Moreira 2021), the application of wavelet transform for genomic classification (Kar et al. 2023), and the utilization of a k-mer-based technique for predicting SARS-CoV-2 variants of concern (Arslan 2023).

The objective of the current study was to evaluate the k-mer-based technique, an alignment-free method, in terms of its accuracy and time efficiency. The assessment of accuracy involved comparing the topologies (classifications) generated by the k-mer approach and a traditional alignment-based method, assuming that the latter produces the correct topology. The assessment of time efficiency involved comparing the time consumed by each method, justifying the use of the k-mer approach when it significantly accelerates task completion compared to the alignment method.

Materials and Methods

A total of 16 complete genomes of SARS-CoV-2, encompassing variants B.1, AY.99.2, P.1, and P.2, identified and sequenced in Brazil, were subject to analysis

alongside a genome isolated in China (B) (Table I). These sequences are accessible through the National Center for Biotechnology Information (NCBI 2023) under the categories "Genomes & Maps" and subtopics "Genome," "Viruses," and "MERS coronavirus". We included in the analysis only sequences without uncertain bases.

Table I: Genome Information in Cluster Analyses. Sequence Id, GenBank accession numbers, variant identifications, genome sizes (in terms of number of base pairs – bp), location id, and institution that generated each genome included in the cluster analyses (Figures 1 and 2).

Sequence id	GenBank #	Lineage/ Variant	bp	Location id	Submitted by
1	NC_045512.2	B	29903	China	FUDAN
2	OQ430686.1	B.1	29835	Rio de Janeiro	UFRJ
3	MZ419858.1	B.1	29903	Brazil	FIOCRUZ RJ
4	MZ419859.1	B.1	29903	Brazil	FIOCRUZ RJ
5	OQ521691.1	AY.99.2	29818	Brazil	FIOCRUZ AM
6	OQ521707.1	AY.99.2	29767	Brazil	FIOCRUZ AM
7	OQ521823.1	AY.99.2	29813	Brazil	FIOCRUZ AM
8	MZ477746.1	P.1	29741	Paraná	FIOCRUZ
9	MZ477753.1	P.1	29735	Paraná	FIOCRUZ
10	MZ477748.1	P.1	29738	Paraná	FIOCRUZ
11	MZ477756.1	P.1	29731	Paraná	FIOCRUZ
12	MZ264787.1	P.1	29866	Manaus	USP
13	MZ477745.1	P.2	29744	Paraná	FIOCRUZ
14	OL442154.1	P.2	29903	Mato Grosso do Sul	UFMS
15	MZ397170.1	P.2	29842	Brazil	UFLA
16	MZ477802.1	P.2	29680	Paraná	FIOCRUZ

17	MZ169912.1	P.2	29903	Brazil	UFLA

FUDAN: Fudan University (Xangai, China); **UFRJ:** Federal University of Rio de Janeiro; **FIOCRUZ:** Oswaldo Cruz Foundation (RJ: Rio de Janeiro; AM: Amazonas); **USP:** University of São Paulo; **UFMS:** Federal University of Mato Grosso do Sul; **UFLA:** Federal University of Lavras.

The world reference sequence of SARS-CoV-2, known as variant B, was initially sequenced in China (Id. 1 in our analysis – Table I). This same lineage, also referred to as B.1, was subsequently identified in Europe. In the Brazilian context, the epidemic started with two lineages, namely B.1.1.28 and B.1.1.33. Our analysis encompasses three representatives of the B.1 lineage, designated as Ids. 2, 3, and 4 (Table I). Moreover, the variant AY (Ids. 5, 6, and 7 – Table I), also identified in India, made its way to Brazil. This variant was later named Delta, following the Greek letter-based nomenclature proposed by a panel of WHO experts. Variants P.1 (later termed Gamma – Ids. 8 to 12 – Table I) and P.2 (later termed Zeta – Ids. 13 to 17 – Table I) are believed to have originated within Brazilian territory, specifically in Manaus and Rio de Janeiro, respectively.

Method with alignment

The alignment process is one of the ways to quantify the differences between DNA or protein sequences. It involves searching for similar regions in two or more sequences to minimize discrepancies, which might be overestimated due to positional, directional, or starting point factors. The process is dynamic, with various alignment positions being experimented with, assessed, and compared. When using distance-based grouping methods, the optimal alignment should be considered to determine the differences between the sequences. The process can be carried out manually or with the assistance of specialized software.

Specifically, the 17 sequences included in our study were aligned using the MUSCLE (Multiple Sequence Comparison by Log-Expectation) algorithm (Edgar 2004a, Edgar 2004b). Following alignment, a pairwise matrix was constructed, which included the pairwise distances between the sequences. These distances were computed using an uncorrected distance model known as p-distance, which is the ratio of differing nucleotides to the total number of nucleotides compared. Subsequently, the sequences were clustered based on their distances using the UPGMA (Unweighted Pair Group with Arithmetic Mean) method (Sokal & Michener 1958). All these processes were performed using MEGA 11 software (Kumar et al. 2018).

Alignment-free method (k-mer)

A k-mer was defined as a contiguous segment of k nucleotides within a genomic sequence (Huang 2016). In simpler terms, a DNA sequence using k-mers encompasses all the possible permutations of length *k* of the four nucleotides: adenine, cytosine, guanine, and thymine (Yin & Yau 2015). The algorithm receives the user-defined value of k to be tested and then generates all potential combinations of k nucleotides. Subsequently, it scans for each of these combinations across all sequences. The occurrence frequency and position of each combination in every sequence contribute to calculating the similarity between the sequences. This calculation considers all possible combinations.

Specifically, the 17 sequences included in our study were clustered using the packages *Biostrings* (Pagès et al. 2020) and *kmer* (Wilkinson 2018), both available in the R software (R Core Team 2020). Various values of 'k' were assessed to identify the minimum threshold beyond which the final clustering remained unchanged. The script in the R software is available in Supplementary Material.

Results

When comparing the clusters (classification), Figures 1 and 2 displayed identical groups, with some few differences within the groups, which did not compromise the lineage classification. We also observed discrepancies in the distances calculated using the different methods, which were consistently smaller in the UPGMA method compared to the k-mer method.

In terms of time taken, the alignment process took about one hour and 14 minutes to align the 17 sequences using the alignment method on a computer with a Core i5 processor, 8 GB of RAM, and the Microsoft Windows 10.0.19045 operating system. Following that, the generation of the pairwise distance matrix and the construction of the UPGMA dendrogram took an additional 3 minutes, resulting in one hour and 17 minutes to complete the task. The outcome of this process is depicted in Figure 1.

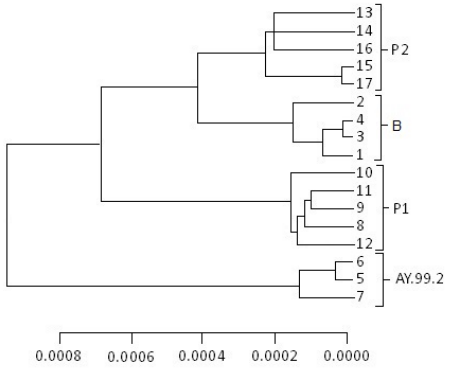


Figure 1: Clustering of 17 SARS-CoV-2 Complete Genomes Using Sequence Alignment and UPGMA Method. The genomes are numbered and identified in Table I. The keys indicate lineage/variant nomenclature according to the World Health Organization.

In the realm of the alignment-free method, generating the cluster using 'k = 6' required a mere two minutes of analysis time on the same computer. The outcomes of the clustering procedure achieved through the alignment-free method are depicted in Figure

2. The choice of 'k = 6' was grounded in the observation that, beyond this point, the clusters consistently maintained an identical pattern.

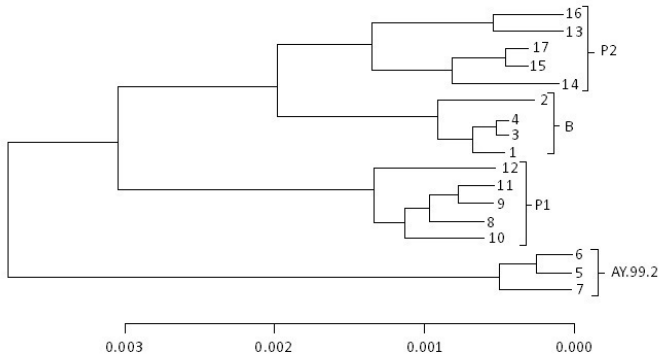


Figure 2: Clustering of 17 SARS-CoV-2 Complete Genomes Using the k-mer Method (k = 6). The genomes are numbered and identified in Table I. The keys indicate lineage/variant nomenclature according to the World Health Organization.

Discussion

The comparison of the two examined methods revealed that the k-mer approach is accurate, as it generates the same clustering pattern as the alignment-based method. Despite the topologies not being identical and the k-mer resulting in systematically greater distances than those calculated by the alignment-based method, which was expected as the distances were calculated using different models, our main objective was to obtain the clusters to properly classify the sequences and thus the virus lineages. It's worth noting that neither of the two methods aims to reconstruct the history of infections or the phylogenetic relationship between groups. Both of these objectives should be pursued with more efficient methods such as network-based approaches (Bandelt et al. 1999) or probabilistic methods like maximum likelihood (Huelsenbeck & Crandall 1997),

or Bayesian inference-based methods (Yang & Rannala 1997), all of which are beyond the scope of this study.

Regarding processing time, aligning the 17 genomes alone took one hour and 14 minutes, in addition to an extra 3 minutes required for calculating the distance matrix and UPGMA. On the other hand, the k-mer approach only took 2 minutes to generate the same cluster pattern, representing a 97% reduction in time compared to the traditional alignment-based method. In times of a pandemic, where new variants emerge rapidly and can be transported across countries within hours, it's crucial to swiftly identify each obtained sequence so that health authorities can take timely action. With the k-mer methodology, it becomes feasible to appropriately classify each sequence within a genome set more rapidly without compromising accuracy.

During the COVID-19 pandemic, hundreds or even thousands of sequences were generated daily. More importantly, these variants exhibited significant differences in terms of virulence and transmissibility (Khalil & Khalil 2020, Sanches et al. 2021), making it even more vital to classify sequences as quickly as possible to prevent hospital overcrowding during the early phases of the pandemic and to select the most suitable variants for vaccine production and distribution in each region.

If an epidemic involved a bacterial infection instead of a viral one, the challenge of producing alignments would likely escalate due to the larger size and intricacy of bacterial genomes. In such cases, the k-mer approach might similarly expedite the classification of diverse bacterial lineages. It's pivotal to subject the k-mer approach and other alignment-free methods to testing across varied pathogenic types well in advance of the emergence of a potent new microorganism that might imperil human health.

Considering the high probability of new pandemics arising from distinct microorganisms (Neumann & Kawaoka 2023), the importance of testing and propagating existing sequencing and identification methodologies cannot be overstated. These more

efficient techniques will facilitate the swifter identification of potential pathogens that could pose threats to humanity, especially considering the vast number of sequences accessible in existing databases. Moreover, rapid methods will enable the implementation of more effective border control measures by swiftly detecting and preempting the spread of new strains — a pivotal stride in the management of pandemics.

Conclusion

The k-mer method has proven to be accurate and offers a crucial time-saving advantage in determining COVID-19 variants compared to traditional alignment-based sequence methods. Thus, when the goal is rapid classification based on complete genomes, the k-mer method should be chosen.

Acnowledgements

We would like to thank Leila Maria Ferreira for helping with the analysis of the article and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES – Financing Code 001) for providing the financial support in the form of scholarship.

Contributors

DAP was responsible for the preparation of the article, running the analysis, building the table, and generating the figures; KSCY helped in the bibliographic research, analysis and discussion of the results; TS was the supervisor of the article and approved the final version of the manuscript.

References

1. ARSLAN H. 2023. A k-mer based metaheuristic approach for detecting COVID-19 variants. Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi 14: 17-26.
2. BANDELT H-J, FORSTER P & RÖHL A. 1999. Median-Joining Networks for Inferring Intraspecific Phylogenies. Mol Biol Evol 16: 37-48.
3. BONI MF, LEMEY P, JIANG X, LAM TTY, PERRY BW, CASTOE TA, RAMBAUT A & ROBERTSON DL. 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nature Microbiology 5: 1408-1417.
4. CHEN Q ET AL. 2020. Mental health care for medical staff in China during the COVID- 19 outbreak. The Lancet Psychiatry 7: e15-e16.
5. CLEEMPUT S, DUMON W, FONSECA V, KARIM WA, GIOVANETTI M, ALCANTARA LC, DEFORCHE K & DE OLIVEIRA T. 2020. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. Bioinformatics 36: 3552-3555.
6. EDGAR RC. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 1-19.
7. EDGAR RC. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792-1797.
8. FLORES-VEGA VR, MONROY-MOLINA JV, JIMÉNEZ-HERNÁNDEZ LE, TORRES AG, SANTOS-PRECIADO JI & ROSALES-REYES R. 2022. SARS-CoV-2: Evolution and Emergence of New Viral Variants. Viruses 14: 1-14.
9. HUANG HH. 2016. An ensemble distance measure of k-mer and Natural Vector for the phylogenetic analysis of multiple-segmented viruses. Journal of Theoretical Biology 398: 136-144.

10. HUELSENBECK JP & CRANDALL KA. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annual Review of Ecology and Systematics 28: 437-466.
11. KAR S, GANGULY M & SEN S. 2023. Lifting scheme-based wavelet transform method for improved genomic classification and sequence analysis of Coronavirus. Innovation and Emerging Technologies 10: 1-17.
12. KHALIL OAK & DA SILVA KHALIL S. 2020. SARS-CoV-2: taxonomia, origem e constituição. Revista de Medicina 99: 473-479.
13. KIM D, LEE JY, YANG JS, KIM JW, KIM VN & CHANG H. 2020. The Architecture of SARS-CoV-2 Transcriptome. Cell 181: 914-921.
14. KUMAR S, STECHER G, LI M, KNYAZ C & TAMURA K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35: 1547-1549.
15. MOHAMADIAN M, CHITI H, SHOGHLI A, BIGLARI S, PARSAMANESH N & ESMAEILZADEH A. 2021. COVID-19: Virology, biology and novel laboratory diagnosis. The Journal of Gene Medicine 23: e3303.
16. MOREIRA RDS. 2021. Análises de classes latentes dos sintomas relacionados à COVID-19 no Brasil: resultados da PNAD-COVID19. Cad Saúde Pública 37: 1-14.
17. NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI) [Internet]. SARS-CoV-2 Data Hub. Available from: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?VirusLineage_ss=Severe%20acute%20respiratory%20syndrome%20coronavirus%202,%20taxid:2697049&SeqType_s=Nucleotide (accessed on 15/March/2023).
18. NEUMANN G & KAWAOKA Y. 2023. Which Virus Will Cause the Next Pandemic?. Viruses 15: 1-8.

19. PAGÈS H, ABOYOUN P, GENTLEMAN R & DEBROY S. 2020. Biostrings: Efficient manipulation of biological strings [Software]. R package version 2.58.0. Available from: https://bioconductor.org/packages/Biostrings.
20. R CORE TEAM. 2020. R: A language and environment for statistical computing [Software]. Version 4.0.3. R Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/.
21. SANCHES PRS, CHARLIE-SILVA I, BRAZ HLB, BITTAR C, CALMON MF, RAHAL P & CILLI EM. 2021. Recent advances in SARS-CoV-2 Spike protein and RBD mutations comparison between new variants Alpha (B.1.1.7, United Kingdom), Beta (B.1.351, South Africa), Gamma (P.1, Brazil) and Delta (B.1.617.2, India). Journal of Virus Eradication 7: 100054.
22. SILVA HPD & LIMA LDD. 2020. Política, economia e saúde: lições da COVID-19. Cad Saúde Pública 37: e00200221.
23. SOKAL RR & MICHENER CD. 1958. A statistical method for evaluating systematic relationships. University Kans Sci Bull 38: 1409-1438.
24. SOUSA FCB, SILVA LCNT, SOUSA MAA, SANTANA MAA & SILVA RB. 2020. Protocolos utilizados para diagnóstico de COVID-19. Revista da FAESF 4: 35-39.
25. WILKINSON SP. 2018. kmer: an R package for fast alignment-free clustering of biological sequences [Software]. R package version 1.0.0. Available from: https://cran.r-project.org/package=kmer.
26. WORLD HEALTH ORGANIZATION [Internet]. WHO Coronavirus (COVID-19) Dashboard. Geneva: World Health Organization. Available from: https://covid19.who.int (accessed on 06/May/2023).
27. YANG Z & RANNALA B. 1997. Bayesian Phylogenetic Inference Using DNA Sequences: A Markov Chain Monte Carlo Method. Mol Biol Evol 14: 717-724.

28. YIN C & YAU SST. 2015. An improved model for whole genome phylogenetic analysis by Fourier transform. Journal of Theoretical Biology 382: 99-110.
29. ZHOU P ET AL. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579: 270-273.

ARTIGO 2 - *An Alignment-Free Method for Classification Coronavirus Types*

Redigido conforme as normas da revista *Austrian Journal of Statistics* (versão em processo de revisão).

Austrian Journal of Statistics
MMMMMM YYYY, Volume VV, Issue II, 1–xx.
http://www.ajs.or.at/

An Alignment-Free Method for Classification Coronavirus Types

Denise de Assis Paiva
Federal University of Lavras

Ana Claúdia Festucci de Herval
Federal University of Lavras

Karla Suemy Clemente Yotoko
Federal University of Viçosa

Thelma Sáfadi
Federal University of Lavras

Abstract

A few years ago, in 2003, the first coronavirus appeared, the most recent being the SARS-CoV-2 outbreak, which triggered a global public health crisis in January 2020, marking the COVID-19 pandemic. The characteristics of the main coronaviruses, SARS-CoV-2, SARS-CoV and MERS-CoV, are distinct, with different mutation patterns, transmission dynamics and prevention strategies. Studies involving viruses require genomic sequences, which are essential for determining the viral strain under investigation. However, the classification of these sequences requires comparison with sequences of strains that have already been classified, which generally involves alignment, the slowest stage of the virus identification protocol. In this study, we focused on evaluating the accuracy and time efficiency of the Natural Vector method, which, unlike traditional methods, does not depend on sequence alignment. Using data from the National Center for Information and Biotechnology related to the SARS-CoV-2, SARS-CoV and MERS-CoV viruses, we classified 11 complete genomes obtained from different countries. We used approaches based on alignment and the Natural Vector method. Both methods resulted in similar classifications; however, the Natural Vector method showed significantly better performance in terms of time efficiency. Therefore, the application of the Natural Vector method in the analysis viral genome, especially during pandemics or epidemics, is extremely important. The combination of speed and precision provided by this method can substantially speed up responses to new viral threats.

Keywords: alignment, clustering, covid-19, natural vector.

1. Introduction

Zoonoses are infections naturally transmitted between vertebrate animals and humans, as observed in the coronaviruses of Severe Acute Respiratory Syndrome (SARS-CoV), Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome 2 (SARS-CoV-2) (Khalil and da Silva Khalil 2020). According to these authors, the spatial and physical proximity between wild or domesticated species creates conditions for viruses to break the barrier between host species, leading to the emergence of new diseases.

SARS-CoV-2, which is responsible for COVID-19, represents the seventh coronavirus identified as pathogenic to humans and is classified within the genus Betacoronavirus. Its predecessors include MERS-CoV, SARS-CoV, HCoV-229E, HCoV-HKU1, HCoV-NL63, and HCoV-OC43 (Pei and Yau 2021). Of these, three led to significant outbreaks. The first was SARS-CoV, detected in Asia in 2002; the second was MERS-CoV, identified in 2012 in Saudi Arabia (Khalil and da Silva Khalil 2020); and the third was SARS-CoV-2, responsible for COVID-19, detected in Wuhan, China, at the end of 2019 (Mohamadian, Chiti, Shoghli, Biglari, Parsamanesh, and Esmaeilzadeh 2021). As of 2023, the virus causing Severe Acute Respiratory Syndrome 2 (SARS-CoV-2) has infected more than 772 million people worldwide (WHO 2023), with Brazil being one of the hardest hit countries, accounting for 38 million infections and 708 thousand deaths (MS 2023). Although these three viruses belong to the same coronavirus family and cause respiratory illnesses, they exhibit differences in terms of transmission, symptoms, and disease severity.

RNA viruses account for a significant portion of infectious diseases affecting humans and animals, such as SARS, hepatitis, influenza, and avian infectious bronchitis (IB) (Mohamadian *et al.* 2021), and also include viruses like AIDS, dengue, and chikungunya. Similar to other RNA viruses, Coronaviruses (CoVs) experience frequent recombination, enabling rapid evolution. This evolution can lead to changes in host/tissue specificity and affect the effectiveness of drugs designed to control them (Kim, Lee, Yang, Kim, Kim, and Chang 2020).

Viruses are classified based on their properties, such as morphology, serology, host range, genome organization, and genetic material sequences. With the advancement of better sequencing technologies, more virus sequences have become available in public databases, making these databases crucial for the classification and identification of viruses (Bao, Chetvernin, and Tatusova 2014).

The classification of viruses based on genomic sequences provides valuable information for bioinformatics (Meier-Kolthoff and Göker 2017). Most existing methods are based on sequence alignment algorithms, a method that becomes prohibitively slow and expensive in terms of computational resources as the size of the database increases (Zielezinski, Vinga, Almeida, and Karlowski (2017); Guan, Zhao, and Yau (2022)).

Considering the growing number of complete genomes available, alignment-free methods have become increasingly popular (He, Dong, He, and Yau 2020). Disease outbreaks demand quick and effective identification of causative agents to inform strategic planning, containment, and treatment (Randhawa, Soltysiak, El Roz, de Souza, Hill, and Kari 2020). Since they do not rely on dynamic programming, are computationally less expensive, and provide accurate results much faster than alignments, alignment-free methods are proving to be suitable tools for comparing complete genomes (He, Sun, Zhang, Bao, and Li 2021).

The current study aimed to evaluate the Natural Vector (NV) method, an alignment-free approach, in terms of its accuracy and temporal efficiency. The accuracy evaluation involved comparing the topologies (classifications) generated by the NV approach with those of a traditional alignment-based method, assuming that the latter produces the correct classification of viruses. Regarding time efficiency, we compared the time consumed by each method, justifying the preference for the NV approach, which significantly speeds up the completion of the task compared to the alignment method.

2. Materials and methods

2.1. Genome data

A total of 11 complete genomes were analyzed, including five from SARS-CoV-2, three from SARS-CoV, and three from MERS-CoV, sequenced in Brazil, Saudi Arabia, and China, respectively (Table 1). These sequences are available at the National Center for Biotechnology Information (NCBI 2023). We included in the analysis only sequences without uncertain bases

(N's or letters other than A, C, T, and G) to ensure the accuracy of the analysis (Zhang, Wen, Li, and Li (2021); Pei and Yau (2021)).

Table 1: Genome Information in Cluster Analyses. Sequence Id, GenBank accession numbers, variant identifications, genome sizes (in terms of number of base pairs – bp), location id, and institution that generated each genome included in the cluster analyses (Figures 1(a) and 1(b)).

Sequence id	GenBank #	Virus	bp	Location id	Submitted by[1]
1	KF600627.1	MERS-CoV	30076	Saudi Arabia	KSAMH
2	KJ156866.1	MERS-CoV	30054	Saudi Arabia	KSAMH
3	KF186567.1	MERS-CoV	30117	Saudi Arabia	KSAMH
4	AY545916.1	SARS-CoV	29721	China	HKU
5	AY545918.1	SARS-CoV	29737	China	HKU
6	AY274119.3	SARS-CoV	29751	Canada	BCCRC
7	MZ477746.1	SARS-CoV-2	29741	Paraná	FIOCRUZ PR
8	MZ419858.1	SARS-CoV-2	29903	Brazil	FIOCRUZ RJ
9	OQ521691.1	SARS-CoV-2	29818	Brazil	FIOCRUZ AM
10	MZ397170.1	SARS-CoV-2	29842	Brazil	UFLA
11	NC_045512.2	SARS-CoV-2	29903	China	FUDAN

[1]**KSAMH:** Kingdom of Saudi Arabia Ministry of Health; **HKU:** The University of Hong Kong; **BCCRC:** British Columbia Cancer Research Centre; **FIOCRUZ:** Oswaldo Cruz Foundation (PR: Paraná; RJ: Rio de Janeiro; AM: Amazonas); **UFLA:** Federal University of Lavras; **FUDAN:** Fudan University (Xangai, China).

2.2. Method with alignment

Sequence alignment is a fundamental technique in bioinformatics. This process involves comparing sequences, whether they are of nucleotides or proteins, to identify the degree of similarity between them.

In this study, the 11 sequences were automatically aligned using the MUSCLE algorithm (Multiple Sequence Comparison by Log-Expectation; Edgar (2004a); Edgar (2004b)). After alignment, a matrix of pairwise distances between the sequences was constructed using an uncorrected method known as p-distance, which is the ratio of different nucleotides to the total number of nucleotides compared. Subsequently, the sequences were grouped based on their distances using the UPGMA method (Unweighted Pair Group with Arithmetic Mean) (Sokal and Michener 1958). All these processes were performed using the MEGA 11 software (Kumar, Stecher, Li, Knyaz, and Tamura 2018).

2.3. Alignment-free method - natural vector

According toYu, Deng, Cheng, Yau, He, and Yau (2013a), among alignment-free techniques, the Natural Vector (NV) method in DNA/protein sequences is used for characterizing genomes and proteins. The NV of a DNA sequence is based on the nitrogenous bases of the DNA sequence: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T).

To define NV, first consider $S = (s_1, s_2, \ldots, s_n)$ a DNA sequence of size n, i.e., each $s_i \in$ {A, C, G, T}, for $i = 1, \ldots, n$. For $k = A, C, G, T$, define the function $w_k(\cdot)$: {A, C, G, T} $\rightarrow \{0, 1\}$ such that, for each element s_i of the sequence, it is as follows (Equation 1):

$$w_k(s_i) = \begin{cases} 1, & \text{se } s_i = k \\ 0, & \text{caso contrário.} \end{cases} \tag{1}$$

With this, the statistics used to compose the Natural Vector of sequence S are (Yu *et al.*

2013a):

- The number of occurrences of letter k in sequence S (Equation 2):

$$n_k = \sum_{i=1}^{n} w_k\left(s_i\right). \tag{2}$$

- The mean position of letter k in sequence S (Equation 3):

$$\mu_k = \sum_{i=1}^{n} i \frac{w_k\left(s_i\right)}{n_k}. \tag{3}$$

- The j-th normalized central moment of letter k in sequence S (Equation 4):

$$D_j^k = \sum_{i=1}^{n} \frac{\left(i-\mu_k\right)^j w_k\left(s_i\right)}{n_k^{j-1} n^{j-1}}, j = 2, 3, \ldots, n_k. \tag{4}$$

The Natural Vector of a DNA sequence S is defined as a vector of dimension $n+4$ (Equation 5):

$$NV(S)=(n_A,\mu_A,D_2^A,\ldots,D_{n_A}^A,n_C,\mu_C,D_2^C,\ldots,D_{n_C}^C,n_G,\mu_G,D_2^G,\ldots,D_{n_G}^G,n_T,\mu_T,D_2^T,\ldots,D_{n_T}^T). \tag{5}$$

According to Yu, Hernandez, Zheng, Yau, Huang, He, Yang, and Yau (2013b), the approach was named Natural Vector because the three groups of parameters used in NV are natural, based on the quantities and distributions of the four nucleotides (A, C, G, T) in the original DNA sequence.

The 11 sequences included in our study were grouped using the seqinr package (Charif and Lobry 2007) and stats (R Core Team 2023), for reading the fasta file, calculating the distance matrix, and performing cluster analysis, respectively, both available in the R software (R Core Team 2023).

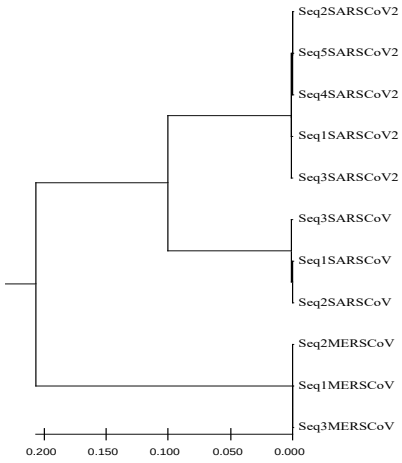
2.4. Cluster analysis

Cluster analysis is a statistical technique used to classify elements into groups or clusters, such that elements within the same group are very similar, and elements in different groups are distinct from each other.

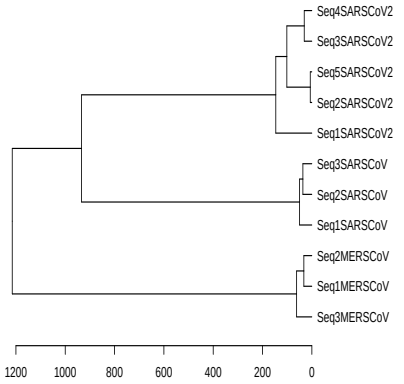
Clustering methods are generally divided into hierarchical and non-hierarchical methods (Ferreira 2018). For this analysis, the UPGMA method was used to generate the clusters for both methods, with and without alignment. For the Natural Vector method, the clustering was carried out using the Euclidean distance, which is the square root of the sum of the squares of the differences.

3. Results

In the comparison of the clusters, Figures 1(a) (with alignment) and 1(b) (Natural Vector) revealed identical groups, albeit with some minor divergences within these clusters that did not affect the overall classification. Generally, the distances obtained by the alignment method were smaller than those calculated using the NV method. The alignment method consumed 57 minutes to align the 11 sequences, calculate the distance matrix and generate the cluster on a computer with a Core i5 processor, 8 GB of RAM, and Microsoft Windows 10.0.19045 operating system. The NV consumed 5 minutes to construct the vector, calculate the matrix and generate the cluster using the same computer.



(a)



(b)

Figure 1: Clustering of 11 complete genomes of SARS-CoV-2, SARS-CoV and MERS-CoV. (a) using sequence alignment and UPGMA method. (b) using the Natural Vector method and UPGMA method. The genomes are numbered and identified in Table 1.

4. Discussions

The comparison between the two methods revealed the precision of the Natural Vector approach. Taking only 5 minutes, as opposed to the 57 minutes required by the alignment method, the NV generated the same clustering pattern as the alignment-based method, though the topologies differed in minor details. The variations observed in the values of the distance matrices were expected and can be attributed to the use of distinct models for calculating distances.

The exploration of increasingly more habitats on the planet, coupled with recent population growth and connectivity between continents, puts us in contact for the first time with viruses that were until now restricted to unexplored localities or to previously unknown animals that can be harmful to human health due to a lack of antibodies. In these cases, the rapid identification of the pathogen is essential, and in the event of an epidemic, the detection of new strains is crucial to support policies aimed at mitigating the spread of pandemics, so that health authorities can take timely measures.

Using alignment-based methods, it is almost impossible, in terms of the required computational time, to classify a new virus by comparing its genome sequence with the thousands of complete genomes already available in databases (Randhawa *et al.* 2020). On the other hand, the NV method allows for the determination of phylogenetic relationships for all viruses at any level (for example, Baltimore class, family, subfamily, genus, and species) in real time. This approach has been successfully used to predict and correct viral classification information, as well as to identify viral origins, for example, a recent threat to public health (Yu *et al.* 2013b).

5. Final remarks

The Natural Vector method correctly classified the sequences, segregating them into groups according to the corresponding types of coronaviruses (MERS-CoV-2, SARS-CoV, and SARS-CoV-2), demonstrating its efficiency as a tool for classification into broad categories. Collectively, our results suggest using NV for the initial classification, allowing the identification of a new virus based on its genome compared to various other viruses.

6. Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Bao Y, Chetvernin V, Tatusova T (2014). "Improvements to Pairwise Sequence Comparison (PASC): A Genome-Based Web Tool for Virus Classification." *Archives of Virology*, **159**, 3293–3304. doi:10.1007/s00705-014-2197-x.

Charif D, Lobry JR (2007). "SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis." In U Bastolla, M Porto, H Roman, M Vendruscolo (eds.), *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*, Biological and Medical Physics, Biomedical Engineering, pp. 207–232. Springer Verlag, New York. ISBN : 978-3-540-35305-8.

Edgar RC (2004a). "MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity." *BMC Bioinformatics*, **5**, 1–19. doi:https://doi.org/10.1186/1471-2105-5-113.

Edgar RC (2004b). "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research*, **32**(5), 1792–1797. doi:https://doi.org/10.1093/nar/gkh340.

Ferreira DF (2018). "Estatística Multivariada 3 ed." p. 624. UFLA, Lavras - MG. ISBN : 9788581270630.

Guan M, Zhao L, Yau SST (2022). "Classification of Protein Sequences by a Novel Alignment-Free Method on Bacterial and Virus Families." *Genes*, **13**(10), 1744. doi:https://doi.org/10.3390/genes13101744.

He L, Dong R, He RL, Yau SST (2020). "A Novel Alignment-Free Method for HIV-1 Subtype Classification." *Infection, Genetics and Evolution*, **77**, 104080. doi:https://doi.org/10.1016/j.meegid.2019.104080.

He L, Sun S, Zhang Q, Bao X, Li PK (2021). "Alignment-Free Sequence Comparison for Virus Genomes Based on Location Correlation Coefficient." *Infection, Genetics and Evolution*, **96**, 1–13. doi:https://doi.org/10.1016/j.meegid.2021.105106.

Khalil OAK, da Silva Khalil S (2020). "SARS-CoV-2: Taxonomia, Origem e Constituição." *Revista de Medicina*, **99**(5), 473–479. doi:https://doi.org/10.11606/issn.1679-9836.v99i5p473-479.

Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H (2020). "The Architecture of SARS-CoV-2 Transcriptome." *Cell*, **181**(4), 914–921. doi:https://doi.org/10.1016/j.cell.2020.04.011.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018). "MEGA X: Molecular Evolutionary Genetics Analysis Across Computing Platforms." *Molecular Biology and Evolution*, **35**(6), 1547–1549. doi:10.1093/molbev/msy096.

Meier-Kolthoff JP, Göker M (2017). "VICTOR: Genome-Based Phylogeny and Classification of Prokaryotic Viruses." *Bioinformatics*, **33**(21), 3396–3404. doi:10.1093/bioinformatics/btx440.

Mohamadian M, Chiti H, Shoghli A, Biglari S, Parsamanesh N, Esmaeilzadeh A (2021). "COVID-19: Virology, Biology and Novel Laboratory Diagnosis." *The Journal of Gene Medicine*, **23**(2), e3303. doi:https://doi.org/10.1002/jgm.3303.

MS (2023). *Painel Coronavírus*. Ministério da Saúde. Secretaria de Vigilância à Saúde (SVS): Guia de vigilância Epidemiológica. URL https://covid.saude.gov.br/.

NCBI (2023). *SARS-CoV-2 Data Hub*. National Center for Biotechnology Information. URL https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?VirusLineage_ss=Severe%20acute%20respiratory%20syndrome%20coronavirus%202,%20taxid:2697049&SeqType_s=Nucleotide/.

Pei S, Yau SST (2021). "Analysis of the Genomic Distance Between Bat Coronavirus RaTG13 and SARS-CoV-2 Reveals Multiple Origins of COVID-19." *Acta Mathematica Scientia*, **41**(3), 1017–1022. doi:https://doi.org/10.1007/s10473-021-0323-x.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Randhawa GS, Soltysiak MP, El Roz H, de Souza CP, Hill KA, Kari L (2020). "Machine Learning Using Intrinsic Genomic Signatures for Rapid Classification of Novel Pathogens: COVID-19 Case Study." *Plos One*, **15**(4), e0232391. doi:https://doi.org/10.1371/journal.pone.0232391.

Sokal RR, Michener C (1958). "A Statistical Method for Evaluating Systematic Relationships." *University Kans Sci Bull*, **38**, 1409–1438.

WHO (2023). *WHO COVID-19 Dashboard*. World Health Organization. URL https://covid19.who.int/.

Yu C, Deng M, Cheng SY, Yau SC, He RL, Yau SST (2013a). "Protein Space: A Natural Method for Realizing the Nature of Protein Universe." *Journal of Theoretical Biology*, **318**, 197–204. doi:https://doi.org/10.1016/j.jtbi.2012.11.005.

Yu C, Hernandez T, Zheng H, Yau SC, Huang HH, He RL, Yang J, Yau SST (2013b). "Real Time Classification of Viruses in 12 Dimensions." *PloS One*, **8**(5), e64328. doi:https://doi.org/10.1371/journal.pone.0064328.

Zhang Y, Wen J, Li X, Li G (2021). "Exploration of Hosts and Transmission Traits for SARS-CoV-2 Based on the K-mer Natural Vector." *Infection, Genetics and Evolution*, **93**, 104933. doi:https://doi.org/10.1016/j.meegid.2021.104933.

Zielezinski A, Vinga S, Almeida J, Karlowski WM (2017). "Alignment-Free Sequence Comparison: Benefits, Applications, and Tools." *Genome Biology*, **18**, 1–17. doi:10.1186/s13059-017-1319-7.

Affiliation:

Denise de Assis Paiva
Departament of Statistics
Federal University of Lavras
Lavras, MG, Brazil
E-mail: denisepaiva1310@gmail.com
URL: http://lattes.cnpq.br/6114238098801542

APÊNDICE A - Códigos

A seguir são apresentados os códigos em R utilizados neste trabalho. As sequências, assim como os valores dos parâmetros podem ser alterados.

Artigo 1

```
#=========================================================
## ARTIGO 1
## Autores: Denise Paiva, Karla Yotoko e Thelma Sáfadi
## 03 de agosto de 2023
#=========================================================
# ADVANCING VIRAL GENOME CLASSIFICATION: ASSESSING THE EFFICIENCY AND
# ACCURACY OF THE ALIGNMENT-FREE K-MER METHOD IN EMERGING PANDEMICS

# Install Packages
install.packages('Biostrings')
install.packages('kmer')

# If you encounter issues while installing 'Biostrings', please
# ensure that you have at least
# R version 4.3.1 and try running this script:

sessionInfo()
if (!require("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install("Biostrings")

# Load Libraries:
library(Biostrings)
library(kmer)

# It may be helpful to consider shortening the sequence identifiers
# in the .fasta files to make them more concise and informative.
# Otherwise, the names could become excessively long.
```

```
29
30 # Load the current directory - getwd(c://documents/Fasta_files)
31 getwd()
32
33 # Load sequences (sequences already downloaded into the R project)
34
35 # List all files in the current directory
36 list.files()
37
38 # Get a list of FASTA files in the directory
39 fasta_files <- list.files(pattern = '\\.fasta$')
40
41 # Read the nucleotide sequences in FASTA format.
42 sequences <- readAAStringSet(fasta_files)
43
44 # Now, you need to ensure that all sequences have the same length
45 # by determining the number of base pairs in each sequence and
46 # trimming all sequences to have the same number of bases as the
47 # shortest sequence.
48
49 # Count the number of base pairs in each sequence
50 lengths <- lengths(sequences)
51
52 # Print the length of the shortest sequence
53 shortest_length <- min(lengths)
54
55 # Trim sequences to the shortest length
56 trimmed_sequences <- subseq(sequences, end = shortest_length)
57
58 # Convert to a matrix for grouping
59 seqs <- as.matrix(trimmed_sequences)
60
61 # Generate clustering. Here, k=6, but you can experiment with other
62 # values of k.
```

```
63 # Remember that for k = n, you need to set nstart = n.
64 w <- cluster(seqs, k = 6, nstart = 6, residues = NULL, gap = "-")
65 plot(w, main = "", horiz = TRUE)
```

Artigo 2

```
#=====================================================================
# ARTIGO 2
## Autores: Denise Paiva, Ana Festucci, Karla Yotoko e Thelma Sáfadi
## 10 de novembro de 2023
#=====================================================================
### AN ALIGNMENT-FREE METHOD FOR CLASSICATION CORONAVIRUS TYPES

# Install Packages
install.packages("ape")
install.packages("seqinr")

# Load Libraries:
library(ape)
library(seqinr)

# It may be helpful to consider shortening the sequence identifiers
# in the .fasta files to make them more concise and informative.
# Otherwise, the names could become excessively long.

# Load the current directory - getwd(c://documents/Fasta_files)
getwd()

#============== Sequence 1 ==============#

# Running the first sequence in FASTA file:
dados <- read.fasta(file = "Seq1.fasta")
seq <- as.character(dados[[1]])

# Transforming into a DNAbin file:
X <- as.DNAbin(seq)

# Calculating the number of bases in the sequence:
```

```
N <- length(X)

# Defining names for the bases A, C, G and T present in sequence 1:
Afac <- seq=="a"
Cfac <- seq=="c"
Gfac <- seq=="g"
Tfac <- seq=="t"

# Note that Tfac as true and false is already a binary display:
cbind(Tfac[1:6],seq[1:6])

# Statistics to compose the Natural Vector.
# 1) Number of letter k in S:
na <- sum(Afac)
nc <- sum(Cfac)
ng <- sum(Gfac)
nt <- sum(Tfac)

n <- sum(na,nc,ng,nt)

# 2) Mean position of letter k in S:
i <- c(1:n)

mua <- sum(i*(Afac/na))
muc <- sum(i*(Cfac/nc))
mug <- sum(i*(Gfac/ng))
mut <- sum(i*(Tfac/nt))

# 3) Normalized central moment of letter k in S:
j=2

D2a <- sum((((i-mua)^(j)) * Afac)/((na^(j-1))*(n^(j-1))))
D2c <- sum((((i-muc)^(j)) * Cfac)/((nc^(j-1))*(n^(j-1))))
D2g <- sum((((i-mug)^(j)) * Gfac)/((ng^(j-1))*(n^(j-1))))
```

```
68  D2t <- sum((((i-mut)^(j)) * Tfac)/((nt^(j-1))*(n^(j-1))))
69
70
71  # Natural Vector n-dimensional of sequence 1:
72  Seq1 <- c(na,mua,D2a,nc,muc,D2c,ng,mug,D2g,nt,mut,D2t)
73
74  #============== Sequence 2 ==============#
75
76  # Running the second sequence in FASTA file:
77  dados <- read.fasta(file = "Seq2.fasta")
78  seq <- as.character(dados[[1]])
79
80  # Defining names for the bases A, C, G and T present in sequence 2:
81  Afac <- seq=="a"
82  Cfac <- seq=="c"
83  Gfac <- seq=="g"
84  Tfac <- seq=="t"
85
86  # Statistics to compose the Natural Vector.
87  # 1) Number of letter k in S
88  na <- sum(Afac)
89  nc <- sum(Cfac)
90  ng <- sum(Gfac)
91  nt <- sum(Tfac)
92  n <- sum(na,nc,ng,nt)
93
94  # 2) Mean position of letter k in S
95  i <- c(1:n)
96
97  mua <- sum(i*(Afac/na))
98  muc <- sum(i*(Cfac/nc))
99  mug <- sum(i*(Gfac/ng))
100 mut <- sum(i*(Tfac/nt))
101
```

```
# 3) Normalized central moment of letter k in S
j=2

D2a <- sum((((i-mua)^(j)) * Afac)/((na^(j-1))*(n^(j-1))))
D2c <- sum((((i-muc)^(j)) * Cfac)/((nc^(j-1))*(n^(j-1))))
D2g <- sum((((i-mug)^(j)) * Gfac)/((ng^(j-1))*(n^(j-1))))
D2t <- sum((((i-mut)^(j)) * Tfac)/((nt^(j-1))*(n^(j-1))))

# Natural Vector 12-dimensional of sequence 2:
Seq2 <- c(na,mua,D2a,nc,muc,D2c,ng,mug,D2g,nt,mut,D2t)

#========== Euclidean Distance ==========#

# Calculating the Euclidean distance:
X <- dist(rbind(Seq1, Seq2))

#============== Dendrogram ==============#

# Generating the clustering from Euclidean distance:
hc <- as.dendrogram(hclust(X, method="average",members=NULL))
plot(hc,xlab="",ylab="",sub="",main="", horiz = T)
```