



UNIVERSIDADE DO ESTADO DA BAHIA
DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

MAURICIO SOUZA MENEZES

FERRAMENTA COMPUTACIONAL PARA ESTUDO DA EVOLUÇÃO DE ESPÉCIES
VIRAIS BASEADO NO USO DE CÓDONS

SALVADOR

2023

MAURICIO SOUZA MENEZES

FERRAMENTA COMPUTACIONAL PARA ESTUDO DA EVOLUÇÃO DE ESPÉCIES
VIRAIS BASEADO NO USO DE CÓDONS

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia - UNEB, como requisito parcial à obtenção do grau de bacharel em Sistemas de Informação. Área de Concentração: Ciência da Computação

Orientador: PhD Diego Gervasio Frías Suárez

Coorientador: PhD Vagner Fonseca

SALVADOR

2023

Termo de Anuência do Orientador

Declaro para os devidos fins que li e revisei este trabalho e atesto sua qualidade como resultado final desta monografia. Confirmo que o referencial teórico apresentado é completo e suficiente para fundamentar os objetivos propostos e que a metodologia científica utilizada e os resultados finais são consistentes e com qualidade suficiente para submissão à banca examinadora final do Trabalho de Conclusão de Curso II do curso de Bacharelado em Sistemas de Informação.

PhD Diego Gervasio Frías Suárez

MAURICIO SOUZA MENEZES

FERRAMENTA COMPUTACIONAL PARA ESTUDO DA EVOLUÇÃO DE ESPÉCIES
VIRAIS BASEADO NO USO DE CÓDONS

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia - UNEB, como requisito parcial à obtenção do grau de bacharel em Sistemas de Informação. Área de Concentração: Ciência da Computação

Aprovada em:

BANCA EXAMINADORA

PhD Diego Gervasio Frías Suárez (Orientador)
Universidade do Estado da Bahia – UNEB

PhD Vagner Fonseca (Coorientador)
Universidade do Estado da Bahia – UNEB

Dedico este trabalho, com muito amor, a minha
rainha, Miriam Souza Menezes

AGRADECIMENTOS

Agradeço a Deus pela vida e por me guiar nos caminhos certos; Agradeço aos meus pais, Mauricio Porto e Miriam Souza, pela criação e por todo o apoio que me deram; Agradeço também ao meu irmão, Maurílio Souza (mesmo sem merecer. . .) por torrar paciência; Agradeço a minha namorada, Yasmim Arrais, por todo o apoio, conversas e momentos em que me tranquilizou; Agradeço ao meu orientador, Diego Frias, pela amizade, paciência e atenção dada. Agradeço a todos os colegas de curso, em especial aos amigos Joílson Argolo e Marcelo Henrique, que estiveram sempre próximos durante toda essa caminhada. Agradeço ao meu amigo, Alexandre Aquiles, por me ensinar ainda mais, que ajudar ao próximo é essencial em todos os momentos da nossa vida.

RESUMO

Este trabalho tem como objetivo principal o desenvolvimento de um modelo para a análise de genomas virais, baseado no uso de códons. Essa ferramenta se propõe a ser uma importante ferramenta para a análise da evolução de espécies, utilizando sequências genômicas do SARS-COV-2 como base de estudo. A implementação desse modelo visa proporcionar maior eficiência computacional e alcançar resultados mais precisos. Adicionalmente, a ferramenta será capaz de apresentar visualizações gráficas dos resultados obtidos, facilitando a interpretação dos dados e auxiliando na tomada de decisões científicas. Espera-se que essa abordagem proporcione insights valiosos sobre a evolução de espécies virais, contribuindo para o avanço da virologia e da genômica comparativa. Os resultados obtidos serão analisados com o objetivo de demonstrar a eficácia dessa ferramenta na compreensão dos padrões evolutivos em espécies virais, tornando-a uma promissora aliada para pesquisadores e profissionais da área.

Palavras-chave: Bioinformática; Códons; Filogenia; Viral.

ABSTRACT

This work aims to develop a model for the analysis of viral genomes based on the use of codons, which will serve as a tool for studying the evolution of species using genomic sequences of SARS-CoV-2. It is expected that this approach will enable a more efficient computational process and yield improved results. Additionally, the tool will provide graphical visualization of the results, facilitating data interpretation and supporting scientific decision-making. Through the application of this tool, valuable insights into the evolution of viral species are anticipated, contributing to advancements in virology and comparative genomics. The obtained results are expected to demonstrate the effectiveness of the tool in analyzing and understanding evolutionary patterns in viral species, making it a promising resource for researchers and professionals in the field.

Keywords: Bioinformatics; Codons; Phylogeny; Viral.

LISTA DE FIGURAS

Figura 1 – Estrutura do DNA.	19
Figura 2 – Tabela de Códon.	21
Figura 3 – Estrutura do coronavírus.	23
Figura 4 – Processo iterativo da metodologia Design Science Research (DSR).	32
Figura 5 – Pipeline de Download das Sequências Genômicas.	35
Figura 6 – Dataset de sequências genômicas.	36
Figura 7 – Pipeline de Filtragem de Sequências Genômicas Duplicadas.	37
Figura 8 – Pipelines descontinuado.	38
Figura 9 – Pipeline de Alinhamento de Sequências Genômicas Duplicadas.	38
Figura 10 – Pipeline de Extração do Gene Spike das Sequências Genômicas Alinhadas. .	39
Figura 11 – Pipeline de Filtragem de Sequências Genicas Duplicadas.	39
Figura 12 – Pipeline de Filtragem de Sequências Genicas de Má Qualidade.	39
Figura 13 – Arquivos de Entrada do AGUA.	40
Figura 14 – Arquivos de entrada do AGUA e do IQ-TREE.	43

LISTA DE TABELAS

Tabela 1 – <i>IUPAC Nucleotide Code</i>	20
Tabela 2 – Variantes VOC e suas Nomenclaturas <i>PANGO e WHO</i>	25
Tabela 3 – Informações do dataset de sequências genômicas.	36
Tabela 4 – Informações do dataset de teste de sequências genômicas.	37
Tabela 5 – Quantidade de sequências por treinamento utilizados para testes de performance nos dois modelos.	43
Tabela 6 – Tempo de processamento, em segundos, das sequências gênicas com o AGUA.	44
Tabela 7 – Tempo de processamento, em segundos, das sequências gênicas com o IQ-TREE.	44

LISTA DE ABREVIATURAS E SIGLAS

A	adenina
AGUA	<i>Ad hoc Genotyping with Unsupervised Algorithm</i>
BV-BRC	Bacterial and Viral Bioinformatics Resource Center
C	citossina
CBUC	<i>Codon Based Unsupervised Classification</i>
COVID-19	<i>Coronavirus Disease 2019</i>
CPU	<i>Central Processing Unit</i>
DNA	Ácido Desoxirribonucleico (<i>Deoxyribonucleic Acid</i>)
DSR	Design Science Research
G	guanina
G2BC	Grupo de Pesquisa de Bioinformática e Biologia Computacional
GB	Gigabyte
GVF	<i>Gene Variant File</i>
IUPAC	União Internacional de Química Pura e Aplicada (<i>International Union of Pure and Applied Chemistry</i>)
LTS	<i>Long-term support</i>
MB	Megabyte
ML	Máxima Verossimilhança
mRNA	Ácido Ribonucleico Mensageiro (<i>Messenger Ribonucleic Acid</i>)
NCBI	<i>National Center for Biotechnology Information</i>
NGS	Sequenciamento de Nova Geração (<i>Next-generation Sequence</i>)
PANGO	<i>Phylogenetic Assignment of Named Global Outbreak Lineages</i>
RNA	Ácido Ribonucleico (<i>Ribonucleic Acid</i>)
SARS-CoV-2	Coronavírus da Síndrome Respiratória Aguda Grave 2 (<i>Severe Acute Respiratory Syndrome Coronavirus 2</i>)
T	timina
tRNA	Ácido Ribonucleico Transportador (<i>Transporter Ribonucleic Acid</i>)
U	uracila
VOC	Variante de Preocupação (<i>Variant of Concern</i>)

VOI	Variante de Interesse (<i>Variant of Interest</i>)
VUM	Variante Sob Monitoramento (<i>Variant Under Monitoring</i>)
WHO	Organização Mundial da Saúde (<i>World Health Organization</i>)

SUMÁRIO

1	INTRODUÇÃO	13
2	FUNDAMENTAÇÃO TEÓRICA	18
2.1	Biologia Molecular	18
2.2	Vírus	22
2.2.1	SARS-CoV-2	22
2.2.1.1	Nomenclaturas de Linhagens do SARS-CoV-2	24
2.3	Bioinformática	25
2.4	Filogenia	26
2.5	Machine Learning	28
2.6	Trabalhos Correlatos	29
3	DESCRIÇÃO DO PROJETO	31
3.1	Metodologia	31
3.2	Ambiente Experimental e Ciclo de Experimentos	32
3.3	Materiais e Métodos	33
3.4	Plano de Implementação	34
3.4.1	Montagem e Preparação do Dataset	34
3.4.2	AGUA	40
3.4.2.1	Pré-processamento	41
3.4.3	Análise comparativa entre o método proposto (AGUA) e outro método existente	42
4	CONSIDERAÇÕES FINAIS	46
5	TRABALHOS FUTUROS	47
	REFERÊNCIAS	48

1 INTRODUÇÃO

Os problemas impostos pela pandemia do novo coronavírus (do inglês *Coronavirus Disease 2019* (COVID-19)) incluíram a falta de conhecimento suficiente para a compreensão da importância das ameaças biológicas e para a preparação médica, apesar dos avanços científicos e tecnológicos já alcançados na área em questão(1). Em vista disso, o conhecimento prévio sobre os agentes biológicos com potencial para causar pandemias, tem o poder de melhorar substancialmente uma preparação pré-pandemia (2).

Dentre esses desafios, o Coronavírus da Síndrome Respiratória Aguda Grave 2 (*Severe Acute Respiratory Syndrome Coronavirus 2*) (SARS-CoV-2) emergiu como o patógeno de maior prioridade entre 2020 e 2023, devido à sua alta transmissibilidade e a ausência de medidas eficazes. A limitação do conhecimento sobre a epidemiologia, imunidade e patogênese do vírus destacou a necessidade de aprimorar a biossegurança e expandir a capacidade dos laboratórios e profissionais de saúde (3, 4, 5, 6). Com mais de 771 milhões de casos confirmados e quase 7 milhões de mortes globalmente, sendo mais de 193 milhões de casos e quase 3 milhões de mortes concentrados nas Américas(7), ações intergovernamentais foram estabelecidas para responder à pandemia e controlar a transmissão do vírus.

Diante desse cenário, a bioinformática, que é a junção de métodos computacionais e técnicas estatísticas com o objetivo de extrair informações de dados biológicos brutos, desempenha um papel fundamental na interpretação de dados genômicos e na compreensão de processos evolutivos(8). Juntamente com a análise genética, é uma especialidade crucial na compreensão da diversidade e evolução de vírus que podem afetar organismos, incluindo seres humanos, animais e plantas. Essa análise não apenas fornece *insights* sobre a classificação e identificação de vírus, mas também é fundamental para a pesquisa em saúde pública, agronegócio e ecologia. No entanto, a complexidade inerente às sequências genéticas dos vírus e a crescente disponibilidade de dados genômicos desafiam a capacidade de análise humana (9, 10).

A reconstrução filogenética é uma das abordagens amplamente utilizadas na análise da evolução de espécies, que permite investigar as relações evolutivas entre diferentes linhagens de vírus (11, 12, 13). Essas observações são realizadas com base em dados como sequências de

Ácido Desoxirribonucleico (*Deoxyribonucleic Acid*) (DNA) e Ácido Ribonucleico (*Ribonucleic Acid*) (RNA). Essas sequências são formadas por blocos fundamentais chamados de nucleotídeos, que são compostos por uma base nitrogenada, um açúcar e um grupo fosfato. As bases presentes nos nucleotídeos do DNA são adenina (A), timina (T), citosina (C) e guanina (G), enquanto no RNA a base T é substituída pela uracila (U) (14, 15, 16). Segundo Hall e Barlow os métodos filogenéticos podem ser usados para analisar os dados da sequência de nucleotídeos de forma que a ordem de descendência de cepas relacionadas possa ser determinada. Quando associada à análise filogenética apropriada, a epidemiologia molecular tem o potencial de elucidar os mecanismos que levam a surtos microbianos e epidemias.

Uma das principais formas de análise filogenética é realizada através da árvore filogenética, onde são representadas as relações evolutivas entre um conjunto de espécies. De acordo com Morrison elas tem função importante porque apresentam de forma sucinta e particular a evolução dos descendentes partindo de ancestrais em comum.

Seguindo a linha dos métodos até então desenvolvidos, este trabalho busca tentar desenvolver um método de construção de árvores com base nas distâncias obtidas a partir da diferença do uso de códons, e assim poder contribuir com a tarefa de classificação de cepas para entes responsáveis por controles voltados a áreas da saúde, especialmente na descoberta de novas cepas emergentes com potenciais pandêmicos. Ademais, é também importante dispor de alternativas à filogenia molecular atualmente utilizada, para gerar informações de outro ponto de vista e-ou para servir de referência aos métodos filogenéticos (19).

Os métodos atuais ainda demandam de um alto custo computacional, devido principalmente a quantidade de dados a serem tratados, sendo assim, existe a necessidade de desenvolver outros mais baratos e que possam suportar o volume crescente de dados (sequências). Sendo assim, o projeto visa apresentar um método que seja capaz de realizar classificações, com um custo computacional baixo, em relação a outros métodos, e que possa apresentar, do ponto de vista científico, alternativas de comparação com outras técnicas já existentes (20, 21). A hipótese referente à menor complexidade computacional do novo método deverá ser testada no trabalho.

A semelhança genética entre vários vírus infecciosos e mortais fornece uma visão do fato de que o RNA é a chave para discernir e marcar os possíveis patógenos que podem causar uma pandemia. Embora um padrão geral e motivos conservados possam ser observados em ancestrais imediatos, as regiões não conservadas das sequências são o resultado da acumulação

de mutações, seja por inserção ou deleção de um ou vários nucleotídeos ou por substituição pontual de um nucleotídeo por outro. A fonte principal de mutações em vírus são percalços na replicação e a recombinação de RNA (2).

Apesar da utilidade da filogenética e dos softwares comerciais e públicos disponíveis para análises filogenéticas, os métodos filogenéticos são muitas vezes aplicados de forma inadequada. A aplicação inadequada de métodos filogenéticos pode levar a resultados imprecisos e interpretações incorretas. Algumas das práticas inadequadas podem incluir a escolha inadequada de modelos evolutivos, a falta de avaliação estatística adequada, a utilização de dados de baixa qualidade ou a interpretação incorreta dos resultados filogenéticos (22, 23). Mesmo quando aplicados adequadamente, são mal explicados e, portanto, mal compreendidos. (17, p. 1) Além disso, por trabalhar com grandes quantidades de dados, os métodos utilizados devem ser avaliados também em relação ao seu custo computacional.

Até o momento, as soluções desenvolvidas têm sido direcionadas pela reconstrução de árvores filogenéticas, as quais são construídas a partir de diferentes informações genéticas. Uma abordagem comum utiliza as mutações de nucleotídeos em sequências de DNA. Além disso, há métodos que exploram a evolução de sequências de aminoácidos em proteínas, constituindo a filogenia baseada em aminoácidos (24). Outra perspectiva relevante é a filogenia baseada em distâncias, que calcula as distâncias evolutivas entre sequências, podendo utilizar métricas como a distância de Jaccard ou a distância de Hamming (25). Essas diferentes estratégias contribuem para uma compreensão mais abrangente das relações evolutivas entre organismos. Neste aspecto, as ferramentas disponíveis não oferecem uma aplicação no contexto de árvores reconstruídas com distâncias obtidas a partir da diferença do uso de códons. Estes são sequências de três nucleotídeos responsáveis pela codificação dos aminoácidos nas proteínas. Os códons desempenham um papel crucial na determinação da função e estrutura das proteínas, e alterações nos códons podem resultar em mudanças significativas nas características fenotípicas dos vírus. Portanto, é necessário a realização de pesquisas e desenvolvimento de ferramentas que capazes de classificar sequências genéticas com base no uso de códons.

Com base no problema de pesquisa proposto, foram construídos os objetivos que deveriam ser atingidos, os mesmos são apresentados a seguir:

- **Objetivo Geral**

- (i) Desenvolver um novo método de análise da evolução molecular viral.

- **Objetivos Específicos**

- (i) Montagem de dataset do projeto
- (ii) Definir um modelo para validação do método proposto
- (iii) Desenvolver uma de ferramenta para caracterizar/validar o método
- (iv) Coletar os dados necessários para validar o método
- (v) Realizar a comparação da performance computacional do novo método com algum dos métodos do estado da arte.

Esta monografia seguirá uma estrutura cuidadosamente elaborada para abordar de forma abrangente o projeto de desenvolvimento da ferramenta de análise de genes virais baseada no uso de códons. A seguir, descrevemos cada seção, delineando seu conteúdo e importância na apresentação do trabalho:

- **Capítulo 1: Introdução**

O capítulo introdutório contextualiza o problema abordado, destacando sua relevância, importância e necessidade na área de análise de genes virais. Além disso, apresenta a estrutura da monografia, fornecendo uma visão geral das seções subsequentes.

- **Capítulo 2: Fundamentação Teórica**

Este capítulo estabelece as bases teóricas para o projeto. Explora conceitos essenciais relacionados a genes virais, códons, filogenética, metodologia DSR e outras áreas relevantes. É fundamental para a compreensão dos métodos e resultados apresentados posteriormente.

- **Capítulo 3: Descrição do Projeto**

Neste capítulo, detalhamos o projeto em sua totalidade. Isso inclui a metodologia utilizada, materiais e métodos empregados, bem como o plano de implementação. Abordaremos a montagem e preparação do dataset, desenvolvimento do modelo e a análise comparativa do modelo proposto e outro existente.

- **Capítulo 4: Montagem e Preparação do Dataset**

Este capítulo concentra-se nas etapas iniciais do projeto, destacando o processo de montagem do dataset. Descreveremos o procedimento de recuperação de sequências, filtragem, alinhamento, extração de genes de interesse e a remoção de sequências duplicadas. Essas etapas são fundamentais para obter um dataset de alta qualidade.

- **Capítulo 5: Desenvolvimento do Modelo**

Aqui, detalharemos a implementação do modelo de classificação não supervisionada com

base em códons. Isso envolverá a tradução de sequências de DNA, a extração de códigos únicos e o processo de agrupamento. O capítulo também abordará a associação de clusters com classes de sequências.

- **Capítulo 6: Análise Comparativa do Modelo Proposto e Outro Existente**

Nesta seção, realizaremos uma análise comparativa entre o método desenvolvido neste projeto e as técnicas clássicas de filogenética. Avaliaremos o desempenho, precisão e eficiência do nosso modelo em relação aos métodos tradicionais.

- **Capítulo 7: Considerações Finais**

A última seção da monografia destacará as principais conclusões, contribuições do projeto e recomendações para pesquisas futuras. Também enfocará as implicações práticas da ferramenta desenvolvida e seu impacto na análise de genes virais.

2 FUNDAMENTAÇÃO TEÓRICA

A análise de genes virais está localizada em um campo interdisciplinar que exige uma sólida compreensão de diversos conceitos e técnicas. Neste capítulo, exploraremos a fundamentação teórica necessária para a compreensão completa do projeto. Começaremos por abordar os princípios fundamentais da genética viral, discutindo o que é um genoma viral e o papel dos genes em vírus. Em seguida, examinaremos detalhadamente os códigos de codificação genética fornecida pela União Internacional de Química Pura e Aplicada (*International Union of Pure and Applied Chemistry*) (IUPAC), que é essencial para traduzir sequências de nucleotídeos em sequências de aminoácidos.

Este capítulo também destacará a importância da filogenética na classificação de genes virais e como as árvores filogenéticas são construídas com base em informações genéticas.

Para entender completamente os métodos e resultados apresentados nos capítulos subsequentes, é crucial absorver os conceitos apresentados aqui. O conhecimento teórico sólido proporcionará a base necessária para a análise crítica do desenvolvimento da nossa ferramenta de análise de genes virais baseada em codons.

Vamos começar esta jornada pela fundamentação teórica que sustenta o projeto, garantindo uma compreensão sólida e fundamentada de cada passo subsequente.

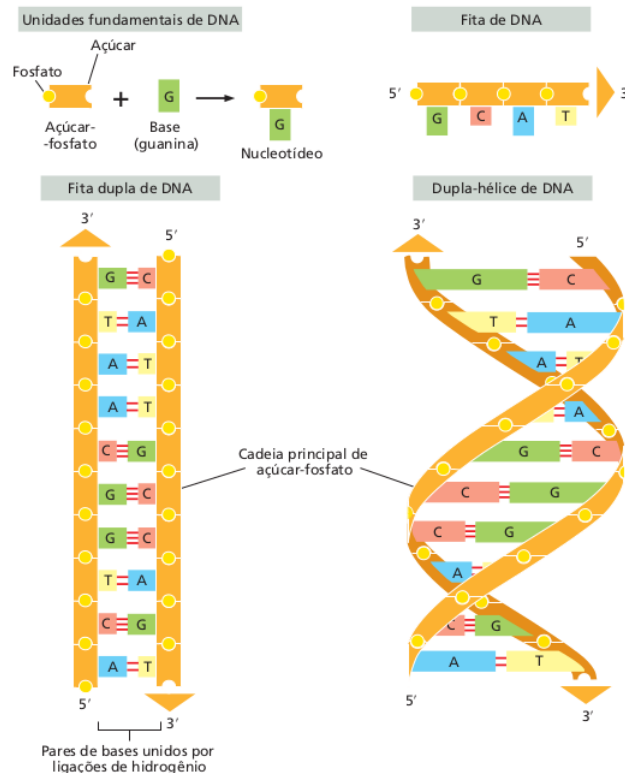
2.1 BIOLOGIA MOLECULAR

A Biologia Molecular é um ramo da biologia que lida e investiga os processos e mecanismos moleculares relacionados à estrutura, função e interações das biomoléculas presentes nos organismos vivos (26). Consiste principalmente em estudar as interações entre os vários sistemas da célula, partindo da relação entre o DNA, RNA e a síntese de proteínas, e o modo como essas interações são reguladas.

É fundamental entender a estrutura do DNA apresentada na Figura 1. Está é uma molécula em forma de dupla hélice que carrega a informação genética em organismos vivos. Ela é composta por duas cadeias polinucleotídicas complementares enroladas em torno de um eixo

central. Cada cadeia é composta por uma sequência de nucleotídeos, que consistem em uma pentose (a desoxirribose), um grupo fosfato e uma base nitrogenada que pode ser A, T, C ou G. A maior parte dos organismos carrega suas informações genéticas no DNA, mas alguns vírus carregam essa informação no RNA, que possui as mesmas bases do DNA só substituindo a T pela U (16).

Figura 1 – Estrutura do DNA.



Fonte: Retirada de *Alberts et al.*(26)

Além das bases nitrogenadas principais já apresentadas (A, T, C e G), a IUPAC que é uma organização não governamental internacional dedicada ao avanço da química, desenvolveu também outras codificações, conhecida como *IUPAC Nucleotide Code*, para representar de maneira padronizada as bases nitrogenadas encontradas nas moléculas de ácido nucleico (27, 28). Também são apresentadas outras letras que representam pares de bases ou misturas específicas, a letra “N” que é usada para representar uma base desconhecida ou não especificada e os símbolos de “.” ou “-”, conhecidos como “GAP”, ou seja, representa onde há uma base identificável ou onde a informação é faltante(27). As mesmas são apresentadas na tabela 1 a seguir.

O conjunto completo de material genético contido em um organismo, seja ele um vírus, uma bactéria, uma planta ou um animal é conhecido como genoma. Ele abrange todas

Tabela 1 – IUPAC Nucleotide Code.

Base	IUPAC Nucleotide Code
Adenina	A
Citosina	C
Guanina	G
Timina	T
Uracila	U
A ou G	R
C ou T	Y
A ou C	M
G ou T	K
G ou C	S
A ou T ou G	W
C ou G ou T	B
A ou C ou T	D
A ou G ou T	H
C ou G ou A	V
A ou C ou G ou T	N
GAP	. ou -

Fonte: Adaptada de Cornish-Bowden(27)

as informações genéticas necessárias para o desenvolvimento, funcionamento e reprodução do organismo. O genoma é composto por sequências de DNA que carregam as instruções para a síntese de proteínas e regulam várias funções celulares (26, 29). A análise do genoma desempenha um papel fundamental na genética, na biologia molecular e na compreensão da hereditariedade e da evolução (26).

As informações contidas no DNA são copiadas em uma molécula de RNA, esse processo é conhecido como transcrição. A transcrição ocorre no núcleo das células e envolve a separação das duas fitas do DNA e o pareamento de nucleotídeos complementares para sintetizar uma molécula de Ácido Ribonucleico Mensageiro (*Messenger Ribonucleic Acid*) (mRNA). O mRNA é uma cópia do DNA que carrega a sequência de bases nitrogenadas correspondente a um gene específico (26). Após isso, ocorre o processo de tradução onde a sequência de bases nitrogenadas do mRNA é utilizada para sintetizar proteínas. A tradução ocorre nos ribossomos, presentes no citoplasma celular. Durante a tradução, o mRNA é lido em grupos de três bases nitrogenadas, chamados de códons (26). Os códons são sequências de três nucleotídeos consecutivos no RNA que correspondem a um aminoácido específico. Uma vez que o número de diferentes combinações de tricas formadas com quatro nucleotídeos é 4^3 , há 64 códons possíveis, entretanto existem apenas 20 aminoácidos correspondentes, como apresentado na Figura 2

conhecida como ‘código genético’, além de sinais de início e parada da tradução, em muitos casos códons correspondem ao mesmo aminoácido (26). A tradução é o processo pelo qual a sequência de códons no RNA é utilizada para sintetizar proteínas. Durante a tradução, os códons são reconhecidos por moléculas de RNA transportador (tRNA) que trazem os aminoácidos correspondentes.

Figura 2 – Tabela de Cótons.

		Segunda letra					
		U	C	A	G		
Primeira letra	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } Ser UCA } UCG }	UAU } Tyr UAC } UAA Parada UAG Parada	UGU } Cys UGC } UGA Parada UGG Trp	U C A G	Terceira letra
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G	
	A	AUU } AUC } Ile AUA } AUG Met	ACU } ACC } Thr ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G	
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G	

Fonte: Adaptada de OpenStax(30)

A relação entre os códons, o DNA e o RNA é crucial para a síntese de proteínas e a expressão genética. O sequenciamento do DNA e a identificação dos códons correspondentes permitem a inferência das sequências de aminoácidos nas proteínas codificadas por um determinado gene (26). Cada códon especifica um aminoácido distinto. Os aminoácidos são transportados para o ribossomo por moléculas de Ácido Ribonucleico Transportador (*Transporter Ribonucleic Acid*) (tRNA), que possuem um anticódon complementar ao códon do mRNA. À medida que o ribossomo percorre o mRNA, os aminoácidos são ligados em uma sequência específica, formando uma cadeia polipeptídica que será dobrada e modificada para se tornar uma proteína funcional (26).

Para determinar a ordem exata dos nucleotídeos em uma molécula de DNA ou RNA é realizada uma técnica conhecida como sequenciamento genético. Existem várias técnicas de sequenciamento, cada uma com suas vantagens e desvantagens, sendo as mais notáveis o sequenciamento de Sanger, o Sequenciamento de Nova Geração (*Next-generation Sequence*)

(NGS) e o sequenciamento de terceira geração. (31, 32, 33)

2.2 VÍRUS

Os vírus são agentes infecciosos que possuem uma estrutura viral que varia entre os seus diferentes tipos, mas que de modo geral é composta por uma cápsula proteica chamada capsídeo, que envolve o material genético viral, que pode ser DNA ou RNA. O capsídeo pode apresentar diferentes formas, como hélices, icosaedros ou formas complexas. Além do capsídeo, alguns vírus possuem uma camada lipídica chamada envelope viral, que é derivada da membrana da célula hospedeira e contém glicoproteínas virais que são importantes para a entrada do vírus nas células hospedeiras (34). O ciclo e vida viral é conjunto de etapas que um vírus passa para se reproduzir e infectar novas células. Esse ciclo pode variar entre diferentes tipos de vírus, mas geralmente envolve as seguintes etapas (14):

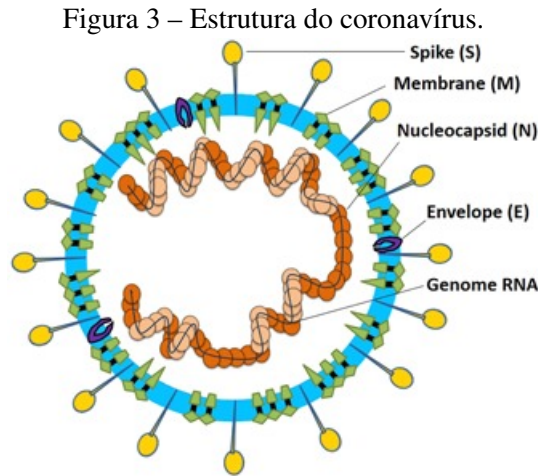
1. **Adsorção:** o vírus se liga especificamente a receptores na superfície da célula hospedeira.
2. **Penetração:** o vírus é internalizado na célula hospedeira, liberando seu material genético.
3. **Replicação e síntese de proteínas virais:** o material genético viral é transportado para os ribossomos da célula hospedeira, replicado e transcrito em moléculas de mRNA, que são utilizadas para a síntese de proteínas virais.
4. **Montagem:** as proteínas virais se unem para formar novas partículas virais.
5. **Liberação:** as novas partículas virais são liberadas da célula hospedeira, para a montagem de novos vírus e para a modificação do ambiente celular para garantir a sua replicação.

2.2.1 SARS-CoV-2

O SARS-CoV-2 é um vírus da família Coronaviridae, que causa a doença chamada COVID-19. Ele foi identificado pela primeira vez em dezembro de 2019 na cidade de Wuhan, na província de Hubei, na China, e desde então se espalhou para todo o mundo, resultando em uma pandemia global (35, 36).

O SARS-CoV-2 possui uma estrutura viral apresentada na Figura 3, característica dos coronavírus. Ele é composto por uma partícula viral esférica, com um envelope lipídico que envolve seu material genético. A estrutura do vírus inclui proteínas de espículas na sua superfície, conhecidas como proteína spike (S). Além disso, o SARS-CoV-2 possui proteínas de membrana (M), envelope (E) e nucleocapsídeo (N), que desempenham papéis importantes na

estrutura e na replicação viral.



Fonte: Retirada de *Li et al.*(37)

A proteína Spike (S) do vírus SARS-CoV-2 é uma das principais proteínas de superfície do vírus e desempenha um papel crucial na infecção das células hospedeiras. Ela é uma glicoproteína que forma estruturas semelhantes a espículas na superfície do vírus, dando-lhe uma aparência coroada. A proteína Spike é o alvo principal das respostas imunes do hospedeiro e é fundamental para a ligação do vírus às células humanas e sua subsequente entrada.

A proteína Spike é composta por três domínios principais: o domínio de ligação ao receptor (RBD - Receptor-Binding Domain), o domínio de fusão (FD - Fusion Domain) e o domínio N-terminal (NTD - N-Terminal Domain). O RBD é particularmente importante, pois é responsável pela interação com o receptor da enzima conversora de angiotensina 2 (ACE2) nas células hospedeiras humanas. Essa interação é crucial para a entrada do vírus nas células.

A estrutura da proteína Spike é altamente dinâmica e pode mudar de conformação para facilitar a fusão da membrana viral com a membrana da célula hospedeira, permitindo assim a entrada do vírus. Essa capacidade de mudança conformacional torna a proteína Spike um alvo promissor para o desenvolvimento de vacinas e terapias antivirais.

Estudos detalhados da proteína Spike são essenciais para compreender a patogenicidade do vírus SARS-CoV-2 e para o desenvolvimento de estratégias terapêuticas eficazes. Além disso, mutações na proteína Spike têm sido identificadas como uma das principais causas de variantes do vírus, o que destaca ainda mais a importância de sua investigação contínua.

2.2.1.1 Nomenclaturas de Linhagens do SARS-CoV-2

A nomenclatura das linhagens do SARS-CoV-2 é uma parte crucial na classificação e rastreamento das diferentes variantes do vírus. Duas das principais nomenclaturas usadas para descrever essas linhagens são a nomenclatura *Phylogenetic Assignment of Named Global Outbreak Lineages* (PANGO) e a nomenclatura da Organização Mundial da Saúde (*World Health Organization*) (WHO). Essas nomenclaturas são usadas para descrever as diferentes variantes do vírus com base em suas características genéticas e filogenéticas.

A nomenclatura PANGO é uma abordagem baseada na filogenia para nomear e rastrear as linhagens do SARS-CoV-2. Ela atribui um nome único a cada linhagem com base em sua posição na árvore filogenética do vírus. Isso permite uma identificação clara das diferentes linhagens e ajuda na compreensão de como o vírus está evoluindo ao longo do tempo. (38)

Já a WHO também desenvolveu sua própria nomenclatura para classificar as variantes do SARS-CoV-2. Essa nomenclatura envolve a utilização de letras gregas em ordem alfabética (Alpha, Beta, Gamma, Delta, etc.) e foi aplicada para evitar a estigmatização de locais geográficos ou populações. (39) Além disso, a WHO também dividiu as variantes em 3 (três) grupos distintos: Variante Sob Monitoramento (*Variant Under Monitoring*) (VUM), Variante de Interesse (*Variant of Interest*) (VOI) e Variante de Preocupação (*Variant of Concern*) (VOC).

A VUM é um termo usado para sinalizar às autoridades de saúde pública que uma variante do SARS-CoV-2 pode exigir atenção e monitoramento priorizados. O principal objetivo desta categoria é investigar se esta variante (e outras intimamente relacionadas com ela) pode representar uma ameaça adicional à saúde pública global em comparação com outras variantes em circulação. Já a VOI, é um termo usado para descrever uma variante do SARS-CoV-2 com alterações que afetam o comportamento do vírus ou seu impacto potencial na saúde humana. Isto pode incluir, por exemplo, a sua capacidade de propagação, a sua capacidade de causar doenças graves ou a facilidade com que pode ser detectada ou tratada. Um VOI também pode ser identificado porque tem uma maior capacidade de propagação quando comparado com outras variantes em circulação, sugerindo um potencial risco emergente para a saúde pública global. Por fim, a VOC é um termo que descreve uma variante do SARS-CoV-2 que atende à definição de VOI, mas também atende a pelo menos um dos seguintes critérios quando comparado com outras variantes(39):

- pode causar uma mudança prejudicial na gravidade da doença.
- pode ter um impacto substancial na capacidade dos sistemas de saúde de prestar cuidados a pacientes com COVID-19 ou outras doenças e, portanto, exigir grandes intervenções de saúde pública.
- há uma diminuição significativa na eficácia das vacinas disponíveis na proteção contra doenças graves.

A tabela 2 apresentada a seguir, contem as VOC *Alpha*, *Beta*, *Gamma*, *Delta* e *Omicron*¹, com as suas respectivas nomenclaturas PANGO e WHO

Tabela 2 – Variantes VOC e suas Nomenclaturas *PANGO* e *WHO*.

Nomenclatura WHO	Nomenclatura PANGO
Alpha	B.1.1.7
Omicron	B.1.1.529
Beta	B.1.351
Delta	B.1.617.2
Gamma	P.1

Fonte: O Autor

2.3 BIOINFORMÁTICA

A bioinformática é um campo interdisciplinar que aplica técnicas de ciência da computação e estatística para entender e interpretar dados biológicos. Essa área surgiu com o advento das tecnologias de sequenciamento de DNA e tem se expandido para abranger diversos aspectos da biologia molecular e genômica. Vamos explorar os principais domínios da bioinformática:

Alinhamento de Sequências: O alinhamento de sequências é uma tarefa central na bioinformática, permitindo comparar e identificar similaridades entre sequências biológicas. Algoritmos como Needleman-Wunsch e Smith-Waterman são usados para alinhamentos globais e locais, respectivamente. Ferramentas populares incluem o BLAST (Basic Local Alignment Search Tool) e o ClustalW.

Análise de Genomas: A análise de genomas visa entender a organização, função e evolução dos genes em um organismo. Ferramentas como Artemis e o Ensembl Genome

¹ Essa variante possui sublinhagens

Browsers facilitam a navegação e interpretação de genomas completos.

Análise de Expressão Gênica: Métodos como RNA-Seq e microarrays são usados para medir a expressão de genes em diferentes condições. Softwares como DESeq2 e edgeR realizam análise diferencial de expressão, identificando genes com níveis significativamente diferentes de expressão.

Filogenética: A filogenética estuda as relações evolutivas entre organismos. Métodos como Máxima Verossimilhança (RAxML), Métodos Bayesianos (MrBayes) e Métodos de Distância (Neighbor-Joining) são utilizados para reconstruir árvores filogenéticas.

Estrutura de Proteínas: Ferramentas como o Phyre2 e o SWISS-MODEL permitem a modelagem e predição da estrutura tridimensional de proteínas, proporcionando insights sobre sua função.

Metagenômica: A metagenômica analisa comunidades microbianas em amostras ambientais. Ferramentas como QIIME e MEGAN auxiliam na análise de dados de sequenciamento de amplicons e metagenomas.

2.4 FILOGENIA

A filogenia é uma disciplina da biologia que estuda as relações evolutivas entre organismos, buscando reconstruir a história evolutiva e a ancestralidade comum. A filogenética molecular é uma abordagem utilizada para inferir a filogenia com base em informações moleculares, como sequências de DNA, RNA e proteínas(22).

A construção de árvores filogenéticas é um aspecto fundamental da filogenética molecular. Existem vários métodos utilizados para construir árvores filogenéticas, que podem ser classificados em dois grupos principais: métodos baseados em distância e métodos baseados em caracteres. Os métodos baseados em distância medem a similaridade ou a dissimilaridade entre sequências moleculares e constroem árvores filogenéticas com base nessas medidas. Alguns exemplos de métodos baseados em distância incluem o método de Neighbor Joining (NJ) e o método de Mínima Evolução (ME). Por outro lado, os métodos baseados em caracteres analisam as mudanças nos caracteres moleculares ao longo do tempo para inferir as relações filogenéticas. Exemplos de métodos baseados em caracteres são o método de Máxima Parcimônia (MP) e o método de Inferência Bayesiana(40).

Ao longo dos anos, vários métodos utilizados para análise filogenética foram desenvolvidos, logo após, será apresentado alguns dos principais e amplamente utilizados:

1. **Método de reconstrução de árvore filogenética de distância (1957):** Esse método é baseado na construção de árvores filogenéticas a partir de uma matriz de distâncias que quantifica a diferença evolutiva entre diferentes sequências. A árvore é construída de modo que as sequências mais semelhantes estejam mais próximas umas das outras. Esse método é amplamente utilizado em análises filogenéticas e é uma das técnicas mais antigas (25).
2. **Método de máxima parsimônia (1966):** A máxima parsimônia busca a árvore filogenética mais simples, ou seja, aquela que requer o menor número de mudanças evolutivas para explicar as sequências observadas. Esse método é baseado no princípio de que a evolução segue o caminho mais econômico, evitando mudanças desnecessárias (41).
3. **Método de máxima verossimilhança (1981):** O método de máxima verossimilhança estima a árvore filogenética que maximiza a probabilidade de observar as sequências dadas, dadas as hipóteses filogenéticas. Ele é baseado na modelagem estatística da evolução molecular e é amplamente considerado um dos métodos mais precisos para a reconstrução de árvores filogenéticas (42).
4. **Método de junção de vizinhos (1987):** O método Neighbor-Joining é uma técnica de construção de árvore filogenética que se baseia em uma abordagem de aglomeração hierárquica. Ele é amplamente utilizado para criar árvores filogenéticas a partir de matrizes de distância, representando a proximidade evolutiva entre sequências ou espécies. Esse método é especialmente útil para análises de grandes conjuntos de dados e é conhecido por sua eficiência computacional (43).
5. **Método de inferência bayesiana (2001):** A inferência bayesiana combina informações a priori com dados observados para estimar a árvore filogenética mais provável. Ela se baseia no Teorema de Bayes e permite incorporar informações prévias sobre as relações filogenéticas. Esse método é particularmente útil quando se dispõe de conhecimento prévio sobre as relações entre as espécies (44).
6. **Método de coalescência (2004):** O método de coalescência, também conhecido como filogenia de coalescência, aborda a filogenia a partir do ponto de vista do ancestral comum mais recente. Ele modela a história da população ancestral e como as sequências evoluíram a partir dessa população. Esse método é especialmente útil para analisar sequências de genes individuais (45).

7. **Método de redes filogenéticas (2005):** As redes filogenéticas são uma extensão das árvores filogenéticas que permitem representar relacionamentos mais complexos, como reticulações ou eventos de hibridização. Elas são úteis quando as relações entre as espécies não podem ser adequadamente representadas por uma árvore simples (46).
8. **Método de filogenia de genoma inteiro (2010):** Esse método se concentra na análise comparativa de genomas completos para inferir relações filogenéticas. Ele utiliza informações genômicas de alta resolução, como sequências de genes e elementos regulatórios, para construir árvores filogenéticas que refletem a evolução das espécies (47).

2.5 MACHINE LEARNING

Machine learning é uma subárea da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos que permitem a um sistema aprender a partir de dados e realizar tarefas específicas sem ser explicitamente programado (48). Dentro do campo do machine learning, existem duas categorias principais de aprendizado: supervisionado e não supervisionado. Neste contexto, abordaremos a segunda categoria, com ênfase nos modelos não supervisionados.

O aprendizado não supervisionado é uma abordagem de machine learning na qual o algoritmo é treinado em dados não rotulados, ou seja, dados que não têm rótulos ou categorias previamente atribuídos. O objetivo do aprendizado não supervisionado é explorar a estrutura e os padrões subjacentes aos dados sem orientação externa. Isso torna o aprendizado não supervisionado útil para tarefas em que a natureza dos dados é desconhecida, e os padrões emergentes devem ser identificados (49).

Um dos principais tipos de tarefa no aprendizado não supervisionado é o agrupamento (clustering). Nessa tarefa, o algoritmo identifica grupos ou clusters de dados que compartilham características semelhantes. O objetivo é agrupar dados de acordo com suas propriedades intrínsecas, sem conhecimento prévio das categorias. Algoritmos de clustering, como o K-Means e o Hierarchical Clustering, são amplamente utilizados em campos como biologia, processamento de imagem, análise de dados e muito mais (49).

O aprendizado não supervisionado é fundamental em diversas aplicações. Na biologia, por exemplo, algoritmos de clustering podem ser usados para identificar grupos de genes que são coexpressos, revelando padrões de regulação genética. Em finanças, a redução de

dimensionalidade pode ser aplicada para entender a relação entre diferentes ativos financeiros. Na área de processamento de linguagem natural, o aprendizado não supervisionado é usado para detectar tópicos em grandes volumes de texto (50).

Apesar de sua versatilidade, o aprendizado não supervisionado também apresenta desafios. A interpretação dos resultados pode ser complexa, pois não há rótulos de classe para validar as descobertas. Além disso, a escolha de hiperparâmetros e a avaliação da qualidade do agrupamento ou da redução de dimensionalidade podem ser complicadas. Em resumo, o aprendizado não supervisionado desempenha um papel fundamental no campo do machine learning, permitindo a extração de informações valiosas de dados não rotulados. Sua capacidade de encontrar estrutura oculta nos dados é crucial em uma variedade de domínios, tornando-o uma ferramenta poderosa na análise e interpretação de informações complexas(51).

2.6 TRABALHOS CORRELATOS

Na busca de trabalhos relacionados, vários métodos foram encontrados, e a seguir são apresentados.

O método de Máxima Verossimilhança (ML) (ou *Maximum Likelihood*), não é exclusivo da filogenia, mas sim uma abordagem estatística. A sua aplicação em filogenia consiste em avaliar a probabilidade de que o modelo de evolução escolhido gere os dados observados, que são por exemplo, características de um organismo. Essa proposta foi utilizada nos seguintes trabalhos:

- *Behl et al.*(2)
- *Fall et al.*(52)
- *Shabbir et al.*(53)
- *Hudu et al.*(54)
- *Sallard et al.*(55)
- *Paez-Espino et al.*(56)
- *Tang et al.*(57)
- *Cho et al.*(58)

Já em *Yin et al.* e *Bedoya-Pilozo et al.*, foi usada a inferência bayesiana, que é fundamentada no teorema de Bayes, que permite a atualização das probabilidades a priori para

probabilidades a posteriori à medida que novas evidências são incorporadas.

Além desses, *Potdar et al.*(61) utilizou a junção de vizinhos (ou *Neighbor-Joining*), que é baseado em uma abordagem heurística que visa construir uma árvore filogenética a partir de uma matriz de distância entre as sequências estudadas. O trabalho de Lichtblau(62) expõe o Frequency Chaos Game Representation e *Kim et al.*(63) a floresta aleatória. Por fim, *Dimitrov et al.*(64) comparou três modelos para reconstrução de árvores filogenéticas: junção de vizinhos; ML e inferência bayesiana.

3 DESCRIÇÃO DO PROJETO

A seguir, serão apresentadas a metodologia e os softwares utilizados neste estudo, bem como as etapas detalhadas de sua implementação.

3.1 METODOLOGIA

Um ponto importante para a obtenção dos objetivos deste trabalho está relacionada a definição da metodologia que servirá como alicerce. Com a proposta de desenvolver e validar um método de análise da evolução molecular de vírus com base no uso de códons, a metodologia escolhida para isso é o DSR. Essa metodologia, proporciona um framework teórico e prático para a criação de artefatos inovadores, como métodos, modelos ou frameworks, visando resolver problemas específicos (65). Neste projeto, a ferramenta de análise de genes virais baseada em códons é o artefato que será desenvolvido e avaliado. Além disso, o DSR enfatiza a validação e a avaliação da utilidade e eficácia do artefato em relação aos seus objetivos práticos. No caso deste projeto, a validação será realizada através da comparação dos resultados obtidos com a ferramenta proposta em relação às técnicas clássicas filogenéticas, que são amplamente utilizadas para a análise de genes virais. Essa comparação permitirá avaliar a eficácia e o valor agregado da abordagem baseada em códons.

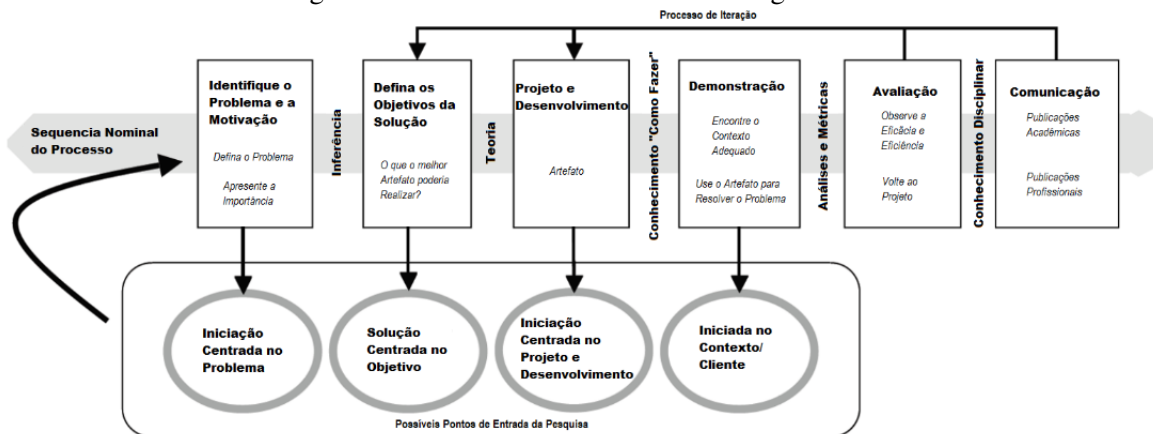
Para a obtenção de sucesso ao utilizar o DSR os seguintes passos serão seguidos:

1. **Identificação do problema e definição dos objetivos:** Nesta fase, o problema a ser resolvido é identificado e compreendido em detalhes. No contexto deste projeto, isso envolveria a compreensão das limitações das abordagens existentes para a classificação de genes virais.
2. **Concepção e planejamento:** Aqui, são definidos os objetivos do artefato a ser criado, suas características e funcionalidades. No projeto em questão, isso envolveria a definição da funcionalidade da ferramenta de análise de genes virais com base em códons.
3. **Desenvolvimento dos artefatos:** Nesta fase, o artefato é desenvolvido. Para o projeto, isso incluiria a criação dos scripts de download, processamento, tradução de sequências, algoritmos de agrupamento e outras partes da ferramenta.

4. **Avaliação do artefato:** O artefato é testado e avaliado quanto à sua eficácia na resolução do problema. Isso pode incluir testes de desempenho, experimentos e comparações com métodos existentes.
5. **Apresentar contribuições científicas:** Os resultados e contribuições do artefato são comunicados, geralmente por meio de artigos científicos e relatórios técnicos.
6. **Iteração:** O processo é iterado conforme necessário. À medida que novos problemas ou insights surgem, o artefato é aprimorado.

Sendo assim, como apresentado na figura4, os passos citados anteriormente são realizados em uma iteração constante, até a obtenção do objetivo final.

Figura 4 – Processo iterativo da metodologia DSR.



Fonte: Adaptada de *Peppers et al.*(65)

Também será utilizada análises quantitativas, ou seja, medidas estatísticas para mensurar e comparar os resultados obtidos.

A pesquisa quantitativa só tem sentido quando há um problema muito bem definido e há informação e teoria a respeito do objeto de conhecimento, entendido aqui como o foco da pesquisa e/ou aquilo que se quer estudar. Esclarecendo mais, só se faz pesquisa de natureza quantitativa quando se conhece as qualidades e se tem controle do que se vai pesquisar (66).

3.2 AMBIENTE EXPERIMENTAL E CICLO DE EXPERIMENTOS

Esta seção descreve o ambiente experimental utilizado para o desenvolvimento do trabalho e a realização dos experimentos. É essencial compreender o contexto no qual os testes foram conduzidos, incluindo as especificações da máquina utilizada.

- **Configurações do Ambiente:** O desenvolvimento do presente trabalho e a execução

dos experimentos foram realizados em um sistema operacional Linux Ubuntu na versão 22.04.3 *Long-term support* (LTS) x86_64. As características fundamentais do computador são detalhadas abaixo:

- Kernel: 6.2.0-37-generic.
- *Central Processing Unit* (CPU): AMD Ryzen 9 7900 (24 núcleos) 3.700GHz
- Memória 32Gigabyte (GB)

Essa máquina está integrada ao ambiente do Grupo de Pesquisa de Bioinformática e Biologia Computacional (G2BC), fornece de uma infraestrutura computacional robusta para proporcionar um ambiente robusto e eficiente para o desenvolvimento do código, bem como para a execução dos experimentos necessários.

- **Ciclo de Experimentos:** Os experimentos foram conduzidos em um ciclo iterativo para garantir a consistência e a validade dos resultados obtidos. Cada iteração envolveu a execução do código no ambiente descrito acima, seguida pela análise cuidadosa dos resultados e possíveis ajustes nos parâmetros experimentais.

Este ciclo permitiu o aperfeiçoamento contínua da abordagem experimental, assegurando a confiabilidade e relevância dos resultados apresentados neste trabalho.

3.3 MATERIAIS E MÉTODOS

Nesta sessão, será apresentada as ferramentas utilizadas para a construção e desenvolvimento de todo o trabalho.

O Python é uma linguagem de programação de alto nível, interpretada, iterativa e de código aberto. Foi criada por Guido van Rossum e lançada em 1991. A linguagem é conhecida por ter uma sintaxe simples, tornando-a popular para o desenvolvimento de software, automação, análise de dados, aprendizado de máquina entre outras aplicações. A mesma apresenta suporte a vários paradigmas de programação, como a orientada a objetos, imperativa, procedural e funcional. Além disso, o Python é portátil, podendo ser executado em diversos sistemas operacionais como Linux, Mac e Windows (67).

Para a construção dos pipelines do projeto, utilizamos Python em conjunto com o Jupyter Notebook. O Jupyter Notebook é uma aplicação de código aberto que permite criar documentos interativos que integram código, texto narrativo e visualizações. É uma ferramenta amplamente adotada por cientistas de dados, pesquisadores e desenvolvedores para explorar

dados, prototipar código, documentar projetos e facilitar a colaboração. Além disso, o Jupyter Notebook oferece suporte a diversas linguagens de programação, incluindo Python (68).

O python possui uma gama de bibliotecas que facilitam a implementação de soluções complexas. A seguir serão apresentadas as bibliotecas utilizadas:

- **Biopython:** Coleção de bibliotecas e ferramentas em Python, disponíveis gratuitamente para biologia molecular computacional. Ele fornece uma ampla gama de funcionalidades, desde a leitura e análise de arquivos de sequência biológica até a execução de algoritmos sofisticados de bioinformática. Desenvolvida e mantida pelo Projeto Biopython, que é uma associação internacional de desenvolvedores de ferramentas python (69).
- **Selenium:** Biblioteca de código aberto que fornece uma interface programática para automatizar interações com navegadores da web. É amplamente utilizado por desenvolvedores e testadores de software para realizar testes automatizados, raspagem de dados na web e outras tarefas que envolvem interações com páginas da web. O Selenium para Python permite a automação de ações como clicar em botões, preencher formulários, navegar em sites e extrair informações da web, tornando-o uma ferramenta valiosa para desenvolvimento e automação de tarefas na web (70).

3.4 PLANO DE IMPLEMENTAÇÃO

Durante o desenvolvimento do projeto foi necessário dividir o projeto em fases com base nas atividades que deveriam ser realizadas de forma a atender todos os passos descritos na seção 3.1. As principais fases identificadas foram: Montagem e preparação do dataset a ser utilizado pelo modelo; Desenvolvimento completo do modelo, com todas as definições, implementações, testes e correções necessárias; e a análise comparativa que será realizada com um outro método existente e já tradicional. Esses pontos são apresentados de forma minuciosa a seguir.

3.4.1 Montagem e Preparação do Dataset

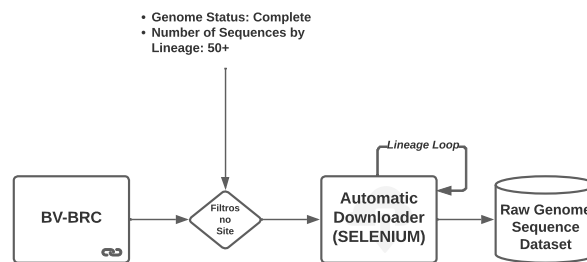
Para realizar o treinamento do modelo a ser construído, eram necessárias sequências únicas e alinhadas do gene Spike. Em vista disso, é importante salientar que o site Bacterial and Viral Bioinformatics Resource Center (BV-BRC) disponibiliza sequências genômicas, e sendo assim, foi preciso construir um pipeline para, após o download das sequências, transformar as

mesmas para a criação de um dataset com as sequências que atendessem os requisitos esperados.

Inicialmente, foi realizada uma análise do BV-BRC, para entender a sua estrutura e verificar também se era possível realizar o download de todas as sequências queridas de forma manual. Foi verificado que o site possuía uma área de seleção de filtros, e foi definido que só seriam selecionadas sequências completas no campo *Genome Status* e no campo *Lineage*, onde é possível filtrar as sequências pelo seu tipo *Pango* e também verificar a quantidade, só os que tivessem mais de 50 sequências do mesmo tipo.

Após a análise, foi constatado que realizar o download manualmente era infactível, e que seria preciso automatizar esse processo de iteração com a página, como vistos na Figura 5. Isto posto, foi realizada uma sequência de passos conhecidos como *Web Scrapping* utilizando o Python juntamente com o Selenium, apresentados em seguida:

Figura 5 – Pipeline de Download das Sequências Genômicas.



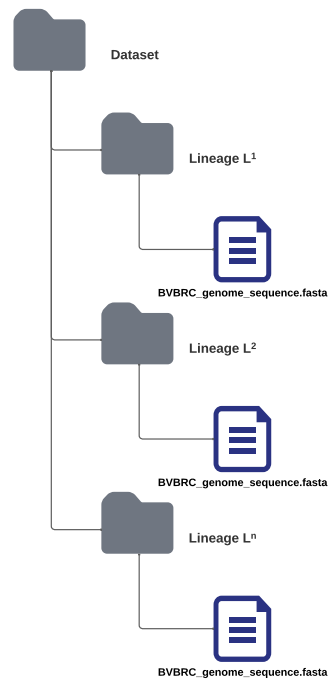
Fonte: O Autor

1. Criar uma lista com todas as linhagens disponíveis e a quantidade de sequências de cada.
2. Desenvolver script Python para remover as linhagens com menos de 50 sequências da lista.
3. Desenvolver script Python para gerar uma url personalizada do BV-BRC, já com os filtros, para cada linhagem.
4. Desenvolver script Python juntamente com o Selenium para abrir as urls de forma automática e realizar o download das sequências.

Ao final do processo de montagem do dataset, realizado no dia 02 junho de 2023, com sequências genômicas completas, foi gerado um diretório raiz (dataset), e dentro deste, um diretório para cada linhagem ($Lineage L^1$, $Lineage L^2$, ..., $Lineage L^n$), contendo um arquivo nomeado BV-BRC_genome_sequence.fasta, como apresentado na figura 6.

Ao finalizar o processo de download, o *dataset* completo ficou com as seguintes

Figura 6 – Dataset de sequências genômicas.



Fonte: O Autor

informações apresentadas na tabela 3.

Tabela 3 – Informações do dataset de sequências genômicas.

Campo	Valor
Quantidade de Linhagens	1086
Quantidade de Sequências Genômicas	1.494.650
Tamanho em GB	47.5

Fonte: O Autor

A seguir, era necessário construir um dataset de sequências do gene Spike a partir do existente. Para isso, com o objetivo de diminuir o tempo de execução dos pipelines durante o desenvolvimento, foi construído também, um dataset de testes, exposto na tabela 4, que serviria como base para execução das atividades, e após as verificações, seria realizado o processo com o dataset completo. No dataset de testes, foram escolhidas 5 (cinco) linhagens que são amplamente conhecidas (gamma, delta, alpha, beta e omicron), o que tornaria mais preciso o processo de validação futuramente.

Depois, foi verificado a existência de sequências genômicas idênticas, melhor dizendo, sequências com exatamente a mesma quantidade e ordem dos nucleotídeos. Por isso, foi

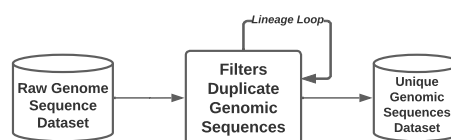
Tabela 4 – Informações do dataset de teste de sequências genômicas.

Pango	WHO	Quantidade de Sequências	Tamanho em Megabyte (MB)
B.1.1.7	Alpha	9982	307
B.1.1.529	Omicron	3694	112.7
B.1.351	Beta	5256	160.6
B.1.617.2	Delta	9996	305.3
P.1	Gamma	10000	305.9
Total		38928	1191,5

Fonte: O Autor

necessário desenvolver um pipeline que filtrasse as sequências repetidas, e mantivesse apenas uma, como exibido na figura 7, gerando assim um novo dataset de sequências genômicas únicas.

Figura 7 – Pipeline de Filtragem de Sequências Genômicas Duplicadas.

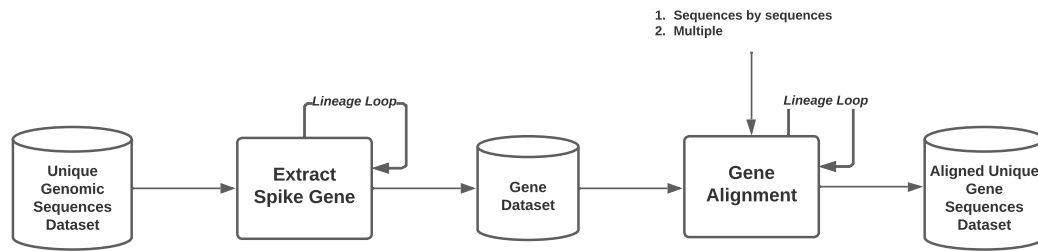


Fonte: O Autor

Com o dataset de sequências únicas montado, foi decidido os passos a seguir, que seriam executados com o papel de atingir o objetivo de obter o dataset de sequências gênicas únicas e alinhadas. Primeiro, como visto na figura 8, foi realizado o processo de extração do gene Spike, utilizando uma sequência de referência do gene obtida do dataset da *National Center for Biotechnology Information* (NCBI)¹, juntamente com o software Blast, para encontrar e extrair o gene Spike de cada uma das sequências genômicas de todo o dataset. Após isso, com o dataset de sequências genicas já montado, era necessário realizar o processo de alinhamento das sequências. Foram verificada duas formas de realizar esse procedimento utilizando o software Clustalo Omega, uma passando uma sequência por vez (*sequence by sequence*) com a sequência do Spike de referência e outra passando todas as sequências com a referência (*multiple sequence*). Mesmo com o dataset de teste, que possuía um tamanho e quantidade de sequências muito inferior ao principal, o processo de alinhamento apresentou uma demora excessiva (até 2 dias) para finalizar. Em decorrência disso, o processo foi descontinuado, e um novo foi repensado e reconstruído e será apresentado a seguir.

¹ Url para download: <https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=genbank&from=21563&to=25384>

Figura 8 – Pipelines descontinuado.

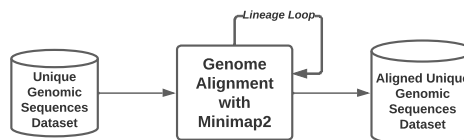


Fonte: O Autor

Após um processo de análise, o desenvolvimento continuou a partir das sequências genômicas únicas, foi feito seguindo os quatro (4) passos apresentados a seguir:

1. **Alinhamento das sequências genômicas:** A primeira etapa, como apresentado na figura 9, consistiu em alinhar as sequências genômicas únicas já disponíveis, utilizando a ferramenta de alinhamento Minimap2. O Minimap2 é uma ferramenta eficiente para mapear sequências genômicas em um genoma de referência. Esse processo permitiu a identificação de regiões específicas relacionadas ao gene spike nas sequências genômicas.

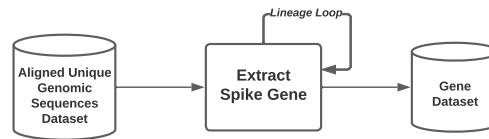
Figura 9 – Pipeline de Alinhamento de Sequências Genômicas Duplicadas.



Fonte: O Autor

2. **Extração do Gene Spike com Uso de Isca:** Após o alinhamento, as sequências de referência do gene spike foram usadas como “iscas” para identificar e extrair as sequências correspondentes nas sequências genômicas alinhadas, como visto na figura 10. O gene spike é de importância crítica, pois desempenha um papel fundamental na interação do vírus com as células hospedeiras. A utilização de iscas garantiu que as sequências genéticas fossem extraídas corretamente do genoma completo.
3. **Filtragem de Sequências Genéticas Duplicadas:** Uma das preocupações na criação do dataset foi a presença de sequências duplicadas, que podem enviesar os resultados da análise. Portanto, as sequências genômicas duplicadas foram identificadas e removidas do conjunto de dados. Esse processo garantiu que cada sequência fosse única, evitando

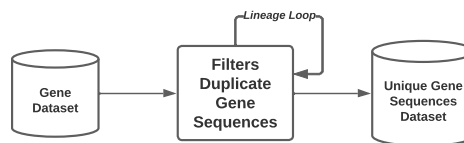
Figura 10 – Pipeline de Extração do Gene Spike das Sequências Genômicas Alinhadas.



Fonte: O Autor

redundâncias.

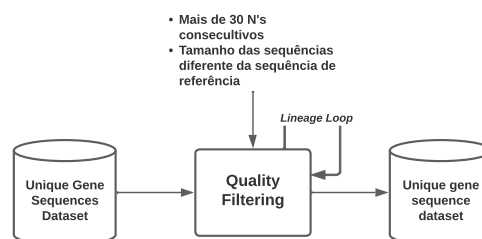
Figura 11 – Pipeline de Filtragem de Sequências Genicas Duplicadas.



Fonte: O Autor

4. **Filtragem de Sequências de Má Qualidade:** Para garantir a qualidade do dataset, as sequências genômicas que continham características indesejáveis foram filtradas. Isso incluiu a remoção de sequências que continham mais de 30 bases nitrogenadas (N) consecutivas e aquelas que não correspondiam ao tamanho esperado das sequências de referência do gene spike como visto na imagem 12. Essa filtragem ajudou a garantir que as sequências incluídas no dataset fossem de alta qualidade e relevantes para a análise subsequente.

Figura 12 – Pipeline de Filtragem de Sequências Genicas de Má Qualidade.

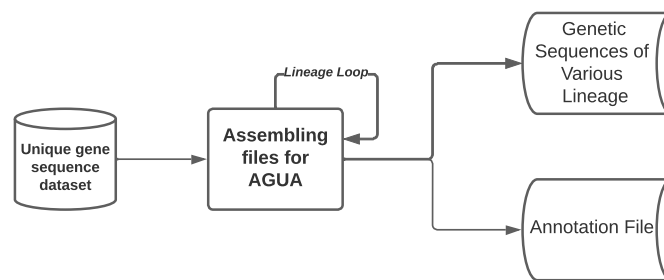


Fonte: O Autor

Ao final do processo, foi gerado um *dataset* de alta qualidade, contendo sequências gênicas únicas para as variantes alpha, beta, delta, gamma e omicron. Esse dataset será a base para a análise de genes virais com base no uso de códons e a aplicação de técnicas de classificação não supervisionada.

A partir deste dataset, foi elaborado 2 (dois) arquivos, conforme a figura 13 apresenta, que serviria de entrada, tanto para o modelo desenvolvido como para a geração de árvores filogenéticas utilizando o modelo convencional, a fim de se realizar análises futuras. Um arquivo compreendia a mescla de sequências genicas de cada uma das linhagem, e um arquivo de anotações que serviria como base no treinamento, contendo o cabeçalho da sequência, na mesma ordem em que estava no arquivo mesclado, juntamente com a linhagem da sequência.

Figura 13 – Arquivos de Entrada do AGUA.



Fonte: O Autor

3.4.2 AGUA

O modelo proposto foi designado como *Ad hoc Genotyping with Unsupervised Algorithm* (AGUA). A subsequente seção discorrerá detalhadamente sobre os procedimentos metodológicos que foram seguidos para a formulação e concepção do referido modelo. Este processo envolveu uma série de etapas meticulosas e estrategicamente delineadas, as quais serão elucidadas a seguir, contribuindo para uma compreensão aprofundada de sua elaboração.

- Levantamento dos requisitos.
- Definição da arquitetura e abordagem do modelo de classificação baseado em códons.
- Implementação do modelo utilizando bibliotecas adequadas.
- Treinamento do modelo.
- Avaliação de desempenho utilizando métricas apropriadas.
- Identificação de problemas e realização de ajustes.

Prosseguindo com a sequência metodológica delineada anteriormente, será detalhada de maneira prática a implementação do AGUA. Este segmento apresentará o resultado tangível desse processo, traduzido em código. As implementações e estruturas técnicas discutidas

a seguir representam a concretização do modelo, oferecendo uma visão aprofundada de suas características fundamentais e contribuições efetivas no âmbito abordado. Para facilitar revisões e colaborações, o código-fonte encontra-se disponibilizado na plataforma de hospedagem GitHub, no repositório do G2BC ².

O AGUA foi particionado em 3 (três) fases distintas, a saber: Pré-processamento, análise quantitativa e análise qualitativa. Cada uma dessas etapas será detalhada nas seções subsequentes.

3.4.2.1 Pré-processamento

O processo inicia-se com a leitura das sequências do gene spike de um arquivo no formato FASTA. Após a leitura, as sequências são verificadas quanto ao seu tamanho, assegurando que sejam múltiplos de 3, um requisito fundamental para a correta representação de códons. Posteriormente, ocorre a conversão das sequências de nucleotídeos em códigos numéricos, representando códons. Esse processo é realizado através do algoritmo *Codon Based Unsupervised Classification* (CBUC), desenvolvido pelo Prof. Drº. Diego Gervasio Frías Suárez e o graduando Diego dos Santos Fonseca (71).

As sequências de códons resultantes passam por uma filtragem, removendo sequências que contenham códons de parada no final. A matriz de sequência de códons é então processada, convertida para um array *numpy* (72), e a contagem dos códons distintos por coluna é realizada. Calcula-se a variabilidade de códons em cada coluna, selecionando aquelas com variabilidade maior que 1, indicando a presença de mais de um tipo de códon.

Para cada sequência, uma lista de ID's é gerada contendo os códons nas colunas selecionadas, e após isso são convertidos em *strings*. A geração de ID's para as classes primárias das sequências segue a identificação e seleção de classes primárias únicas, associadas a números. Um *DataFrame* é criado usando a biblioteca Pandas para organizar as informações do *Gene Variant File* (GVF).

A aplicação do algoritmo de *clustering* clope otimiza a distribuição de transações em clusters para maximizar uma função objetivo(73). Apesar do modelo construído ser não supervisionado, o código otimiza a escolha do valor de repulsão para o algoritmo clope, buscando

² Url do repositório: <<https://github.com/G2BC/agua>>

a melhor resolução de agrupamento para um conjunto de dados de sequências genéticas. Este processo visa encontrar configurações que produzam agrupamentos com alta resolução, onde as sequências pertencentes ao mesmo cluster sejam homogêneas em termos de genótipo, em comparação com os rótulos reais fornecidos no ground truth de anotação.

O código, então, realiza a leitura de um arquivo de anotações e otimiza a escolha do valor de repulsão. A matriz de confusão entre as espécies (anotações) e os clusters é gerada, indicando como as espécies reais foram agrupadas nos clusters encontrados pelo algoritmo clope. A identificação de clusters contendo múltiplas espécies é realizada, seguida pelo cálculo de uma matriz de distâncias entre pares de classes primárias.

Por fim, o código utiliza a biblioteca SciPy para realizar a ligação hierárquica com diferentes métodos de ligação, gerando árvores filogenéticas que são plotadas a partir dos resultados da análise. Ao final do processo, a biblioteca Dill é utilizada para serializar o objeto e gravar o modelo (74).

3.4.3 Análise comparativa entre o método proposto (AGUA) e outro método existente

Nesta seção, apresentamos uma análise comparativa detalhada entre o método proposto e um método tradicional amplamente utilizado. O objetivo é avaliar o desempenho e a eficácia do nosso método em relação a uma abordagem estabelecida. Para esse fim, realizamos a análise em um conjunto de dados comum apresentado na figura 4. O método tradicional selecionado para comparação é o utilizando no *software* IQ-TREE, que é um estimador de máxima verossimilhança de alta performance, frequentemente empregado na filogenia molecular (75).

A escolha do IQ-TREE como método tradicional se baseia em sua prevalência na comunidade científica e sua eficácia comprovada em construir árvores filogenéticas a partir de dados de sequenciamento. A avaliação comparativa entre o método proposto e o IQ-TREE nos permitirá avaliar a capacidade do nosso modelo em produzir resultados relevantes e precisos em comparação com uma abordagem estabelecida.

Considerando o propósito de realizar uma comparação entre os métodos, procedeu-se à análise do tempo de execução de cada um deles, variando tanto o número de linhagens quanto a quantidade de sequências. No que diz respeito ao número de linhagens, foram consideradas 2 (Alpha e Beta) para o conjunto L1, 3 (Alpha, Beta e Gamma) para o conjunto L2, 4 (Alpha,

Beta, Gamma e Delta) para o conjunto L3 e 5 (Alpha, Beta, Gamma, Delta e Omicron) para o conjunto L4, em cada conjunto de dados. Em relação à quantidade de sequências, foram contemplados 100 (T100), 300 (T300), 500 (T500) e 700 (T700) sequências para cada categoria. A tabela 5 apresenta o total de sequências em cada um dos testes que foram montados. Dessa forma, estabeleceram-se cenários variados para avaliação dos métodos, considerando diferentes combinações de linhagens e quantidades de sequências.

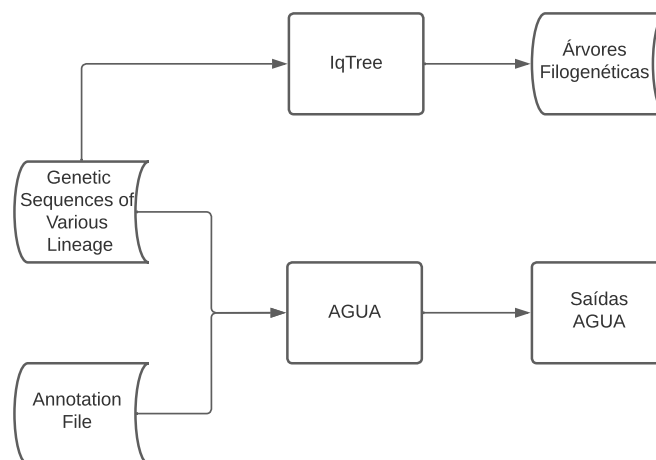
Tabela 5 – Quantidade de sequências por treinamento utilizados para testes de performance nos dois modelos.

Linhagens	T100	T300	T500	T700
L1	200	600	1000	1400
L2	300	900	1500	2100
L3	400	1200	2000	2800
L4	500	1500	2500	3500

Fonte: O Autor

A abordagem metodológica consiste em utilizar o mesmo conjunto de sequências do gene spike, que alimentou o AGUA para a construção de árvores filogenéticas usando o IQ-TREE como apresentado na figura 14. Comparamos, então, o tempo de processamento em ambos os casos. Essa análise comparativa nos permitirá determinar se o método proposto representa uma melhoria significativa em relação à abordagem tradicional, contribuindo assim para o avanço no campo da filogenia molecular.

Figura 14 – Arquivos de entrada do AGUA e do IQ-TREE.



Fonte: O Autor

Os testes foram executados conforme os parâmetros apresentados na tabela 5. Os tempos de processamento para o AGUA são apresentados na tabela 6 e os tempos do IQ-TREE na tabela 7.

Tabela 6 – Tempo de processamento, em segundos, das sequências gênicas com o AGUA.

Linhagens	T100	T300	T500	T700
L1	63.94	518.85	1309.27	2478.68
L2	160.76	1242.55	2992.54	5589.52
L3	298.78	2302.29	5804.33	9878.99
L4	438.31	3696.73	8910.84	15877.98

Fonte: O Autor

Tabela 7 – Tempo de processamento, em segundos, das sequências gênicas com o IQ-TREE.

Linhagens	T100	T300	T500	T700
L1	62.18	93.40	105.20	151.04
L2	76.38	95.33	121.06	177.27
L3	75.41	111.26	324.65	281.13
L4	76.69	131.25	210.89	393.44

Fonte: O Autor

Embora os resultados de tempo de processamento indiquem que o método alternativo apresentou um desempenho mais rápido em comparação com o AGUA em todos os cenários avaliados, é crucial ressaltar que a eficiência temporal não é o único critério para avaliar a robustez de um método. O AGUA, apesar de requerer mais tempo computacional, oferece uma abordagem inovadora para a classificação de linhagens, incorporando aprendizado automático que pode melhorar significativamente a precisão da classificação.

É fundamental destacar que a escolha entre métodos deve levar em consideração não apenas a rapidez da execução, mas também a capacidade de adaptação a diferentes contextos e a qualidade das análises produzidas. O método proposto, AGUA, prioriza a aprendizagem automática e, portanto, oferece uma solução mais robusta em termos de capacidade de classificação.

A rápida execução do método alternativo pode ser atribuída à sua natureza não supervisionada e pré-determinada, o que, embora eficiente em termos de tempo, limita sua adaptabilidade a diferentes conjuntos de dados. O AGUA, ao empregar aprendizado automático, apresenta maior flexibilidade para ajustar-se a variações nos dados, resultando em classificações

mais refinadas e adaptadas.

Em última análise, o AGUA busca proporcionar não apenas uma alternativa eficiente em termos de tempo, mas uma abordagem mais completa e contextualizada para a classificação de linhagens. Ao integrar aprendizado automático, o AGUA não só automatiza o processo, mas também aprimora a qualidade das análises, oferecendo uma contribuição valiosa para a compreensão da diversidade genética.

4 CONSIDERAÇÕES FINAIS

Em síntese, este estudo alcançou plenamente todos os objetivos propostos, abarcando tanto as metas específicas quanto a meta geral, resultando em avanços significativos no campo da análise da evolução molecular viral.

O objetivo geral de desenvolver um novo método para a análise da evolução molecular viral foi completamente atingido, culminando na concepção e implementação bem-sucedida do AGUA, uma ferramenta de análise de genes virais com base no uso de codons para classificação não supervisionada.

Quanto aos objetivos específicos, é possível notar destacar as seguintes conquistas:

- (i) Montagem do Dataset: Realizamos a compilação de um dataset abrangente, fundamental para as análises subsequentes.
- (ii) Definição de um Modelo para Validação: Estabelecemos um modelo robusto que serviu como base para a validação do método proposto, assegurando a solidez das avaliações.
- (iii) Desenvolvimento de uma Ferramenta para Caracterização/Validação: Criamos a ferramenta AGUA, uma implementação eficaz que permitiu não apenas caracterizar, mas também validar o método proposto de forma precisa e eficiente.
- (iv) Coleta de Dados Necessários para a Validação: Executamos a coleta e dos dados essenciais para validar o método, garantindo que as análises fossem realizadas com informações relevantes e representativas.
- (v) Comparação de Performance: Realizamos uma comparação abrangente da performance computacional do AGUA com um método do estado da arte (IQ-TREE).

Portanto, os resultados obtidos e os objetivos integralmente alcançados reforçam a contribuição significativa deste trabalho para o avanço da compreensão da evolução molecular viral, oferecendo uma abordagem inovadora e eficiente por meio do AGUA. Esses êxitos fortalecem a relevância do estudo no contexto científico e apontam para possíveis desenvolvimentos futuros na área.

5 TRABALHOS FUTUROS

Este estudo abre perspectivas para investigações futuras, destacando algumas possibilidades de continuidade:

1. Expansão para Outras Métricas e Espécies Virais:

- Realizar experimentos adicionais com o AGUA, explorando a sua aplicabilidade em métricas alternativas e em sequências de outras espécies virais, além do SARS-CoV-2.
- Avaliar o desempenho do AGUA em relação a sequências de nucleotídeos de diferentes comprimentos, analisando sua adaptabilidade a variabilidades no tamanho dos dados.

2. Análise Qualitativa da Eficiência do Modelo:

- Investigar a eficácia do AGUA na análise qualitativa, particularmente em sua capacidade de identificar novas espécies.
- Avaliar o desempenho do modelo em cenários de descoberta de espécies não previamente catalogadas, fornecendo insights sobre sua utilidade em contextos de pesquisa emergentes.

3. Desenvolvimento de uma Interface (Frontend) (Site) para o AGUA:

- Desenvolver um site que permita a fácil interação com o AGUA, tornando suas funcionalidades mais acessíveis a uma variedade de usuários, incluindo pesquisadores e profissionais da área.

REFERÊNCIAS

- 1 BINSFELD, P. C.; COLONELLO, N. A. Coronavirus - sars-cov-2: Risk group and biosafety consensus for laboratory with infectious samples. **SciELO Preprints**, May 2020. Disponível em: <<https://preprints.scielo.org/index.php/scielo/preprint/view/399>>.
- 2 BEHL, A.; NAIR, A.; MOHAGAONKAR, S.; YADAV, P.; GAMBHIR, K.; TYAGI, N.; SHARMA, R. K.; BUTOLA, B. S.; SHARMA, N. Threat, challenges, and preparedness for future pandemics: A descriptive review of phylogenetic analysis based predictions. **Infection, Genetics and Evolution**, v. 98, p. 105217, mar. 2022. ISSN 15671348. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S1567134822000144>>.
- 3 COLLABORATIVE, G. C. C. SARS-CoV-2 vaccination modelling for safe surgery to save lives: data from an international prospective cohort study. **British Journal of Surgery**, v. 108, n. 9, p. 1056–1063, 03 2021. ISSN 0007-1323. Disponível em: <<https://doi.org/10.1093/bjs/zgab101>>.
- 4 DEIGIN, Y.; SEGRETO, R. Sars-cov-2's claimed natural origin is undermined by issues with genome sequences of its relative strains. **BioEssays**, v. 43, 05 2021.
- 5 KUCHIPUDI, S.; NAIR, M. S.; RUDEN, R.; YON, M.; NISSLY, R.; VANDEGRIFT, K.; NELLI, R.; LI, L.; JAYARAO, B.; MARANAS, C.; LEVINE, N.; WILLGERT, K.; CONLAN, A.; OLSEN, R.; DAVIS, J.; MUSSER, J.; HUDSON, P.; KAPUR, V. Multiple spillovers from humans and onward transmission of sars-cov-2 in white-tailed deer. **Proceedings of the National Academy of Sciences**, v. 119, p. e2121644119, 02 2022.
- 6 MADHI, S.; BAILLIE, V.; CUTLAND, C.; VOYSEY, M.; KOEN, A.; FAIRLIE, L.; PADAYACHEE, S.; DHEDA, K.; BARNABAS, S.; BHORAT, Q.; BRINER, C.; KWATRA, G.; AHMED, K.; ALEY, P.; BHIKHA, S.; BHIMAN, J.; BHORAT, A.; PLESSIS, J.; ESMAIL, A.; IZU, A. Efficacy of the chadox1 ncov-19 covid-19 vaccine against the b.1.351 variant. **New England Journal of Medicine**, v. 384, 03 2021.
- 7 World Health Organization. **COVID-19 Dashboard**. último acesso em 20 de nov. de 2023. Disponível em: <<https://covid19.who.int/>>.
- 8 MADHI, S.; BAILLIE, V.; CUTLAND, C.; VOYSEY, M.; KOEN, A.; FAIRLIE, L.; PADAYACHEE, S.; DHEDA, K.; BARNABAS, S.; BHORAT, Q.; BRINER, C.; KWATRA, G.; AHMED, K.; ALEY, P.; BHIKHA, S.; BHIMAN, J.; BHORAT, A.; PLESSIS, J.; ESMAIL, A.; IZU, A. Efficacy of the chadox1 ncov-19 covid-19 vaccine against the b.1.351 variant. **New England Journal of Medicine**, v. 384, 03 2021.
- 9 WAGNER, E. K.; HEWLETT, M. J.; BLOOM, D. C.; CAMERINI, D. **Basic Virology, 3rd Edition**. 3. ed. [S.l.]: Wiley-Blackwell, 2007. ISBN 1405147156; 9781405147156.
- 10 FLORES, E.; WEIBLEN, R.; GIL, L.; TOBIAS, F.; LIMA, M.; GARCEZ, D.; BOTTON, S. Diversidade antigênica de amostras do vírus da diarreia viral bovina isoladas no Brasil: implicações para o diagnóstico e estratégias de imunização. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, SciELO Brasil, v. 52, p. 11–17, 2000.

- 11 SANTOS, C. M. D.; KLASSA, B. Sistemática filogenética hennigiana: revolução ou mudança no interior de um paradigma? **Scientiae Studia**, v. 10, n. 3, p. 593–612, 2012.
- 12 RIBAS, G. C. Considerações sobre a evolução filogenética do sistema nervoso, o comportamento e a emergência da consciência. **Brazilian Journal of Psychiatry**, SciELO Brasil, v. 28, p. 326–338, 2006.
- 13 SANTOS, C. M. D. d. Os dinossauros de hennig: sobre a importância do monofiletismo para a sistemática biológica. **Scientiae Studia**, SciELO Brasil, v. 6, p. 179–200, 2008.
- 14 ALBERTS, B.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. **Molecular Biology of the Cell**. 4th edition. ed. New York: Garland Science, 2002. ISBN 978-0815344643.
- 15 GLICK, C. L. P. B. R. **Molecular Biotechnology: Principles and Applications of Recombinant DNA, 6th Edition**. 6. ed. [S.l.]: Wiley, 2022. ISBN 9781683673644; 1683673646; 9781683673651; 1683673654; 9781683673668; 1683673662.
- 16 SNUSTAD, M. J. S. D. P. **Fundamentos de Genética**. 7. ed. [S.l.]: Guanabara Koogan, 2017.
- 17 HALL, B. G.; BARLOW, M. Phylogenetic analysis as a tool in molecular epidemiology of infectious diseases. **Annals of Epidemiology**, v. 16, n. 3, p. 157–169, 2006. ISSN 1047-2797. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1047279705001080>>.
- 18 MORRISON, D. A. Tree Thinking: An Introduction to Phylogenetic Biology. David A. Baum and Stacey D. Smith. **Systematic Biology**, v. 62, n. 4, p. 634–637, 05 2013. ISSN 1063-5157. Disponível em: <<https://doi.org/10.1093/sysbio/syt026>>.
- 19 FLINT, S. J.; RACANIELLO, V. R.; RALL, G. F.; SKALKA, A. M. **Principles of Virology**. 4th. ed. [S.l.]: ASM Press, 2015. ISBN 1555819516; 9781555819514.
- 20 JENSEN, F. **Introduction to Computational Chemistry**. 3rd. ed. [S.l.]: Wiley, 2017.
- 21 BROWN, T. A. **Gene Cloning and DNA Analysis: An Introduction**. 8. ed. [S.l.]: Wiley-Blackwell, 2020. ISBN 9781119640783; 1119640784.
- 22 FELSENSTEIN, J. **Inferring Phylogenies**. 2. ed. [S.l.]: Sinauer Associates, 2004. ISBN 0878931775, 9780878931774.
- 23 HUELSENBECK, J. P.; RONQUIST, F. Mrbayes: Bayesian inference of phylogenetic trees. **Bioinformatics**, Oxford University Press, v. 17, n. 8, p. 754–755, 2001.
- 24 LE, S. Q.; GASCUEL, O. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaeobacteria, eubacteria, and eukaryotes. **Critical Reviews in Biochemistry and Molecular Biology**, Taylor & Francis, v. 43, n. 2, p. 89–132, 2008.
- 25 SOKAL, R. R.; MICHENER, C. D. A statistical method for evaluating systematic relationships. **University of Kansas Science Bulletin**, v. 38, n. 22, p. 1409–1438, 1958.
- 26 ALBERTS, B.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. **Biologia Molecular da Célula**. 6. ed. [S.l.]: Artmed, 2017.

- 27 CORNISH-BOWDEN, A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. **Nucleic acids research**, Oxford University Press, v. 13, n. 9, p. 3021, 1985.
- 28 HOPPE-SEYLER, C. Nomenclature committee of iub (nc-iub) and iupac-iub joint commission on biochemical nomenclature (jcnb). **Biol. Chem. Hoppe-Seyler**, v. 367, p. 1–4, 1986.
- 29 PIERCE, B. A. **Genetics: A Conceptual Approach**. 6. ed. [S.l.]: W. H. Freeman, 2016. ISBN 1319050964; 9781319050962.
- 30 OPENSTAX. **The Genetic Code**. último acesso em 05 de jul. de 2023. <<https://openstax.org/books/biology/pages/15-1-the-genetic-code>>.
- 31 JAIN, M.; OLSEN, H. E.; PATEN, B.; AKESON, M. The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. **Genome Biology**, v. 17, n. 1, p. 239, 2016.
- 32 SANGER, F.; NICKLEN, S.; COULSON, A. R. Dna sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences**, v. 74, n. 12, p. 5463–5467, 1977.
- 33 GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics**, v. 17, n. 6, p. 333–351, 2016.
- 34 KNIPE, P. M. H. D. M. **Fields Virology**. 6. ed. [S.l.]: LIPPINCOTT WILLIAMS & WILKINS, 2022. Vol I and II. ISBN 9781451105636, 1451105630, 2013003842.
- 35 ZHU, N.; ZHANG, D.; WANG, W.; LI, X.; YANG, B.; SONG, J.; AL. et. A novel coronavirus from patients with pneumonia in china, 2019. **New England Journal of Medicine**, v. 382, n. 8, p. 727–733, 2020.
- 36 WU, F.; ZHAO, S.; YU, B.; CHEN, Y. M.; WANG, W.; SONG, Z. G.; AL. et. A new coronavirus associated with human respiratory disease in china. **Nature**, v. 579, n. 7798, p. 265–269, 2020.
- 37 LI, G.; FAN, Y.; LAI, Y.; HAN, T.; LI, Z.; ZHOU, P.; PAN, P.; WANG, W.; HU, D.; LIU, X.; ZHANG, Q.; WU, J. Coronavirus infections and immune responses. **J Med Virol**, v. 92, n. 4, p. 424–432, Apr 2020.
- 38 RAMBAUT, A.; HOLMES, E. C.; O'TOOLE, Á.; HILL, V.; MCCRONE, J. T.; RUIS, C.; PLESSIS, L. du; PYBUS, O. G. A dynamic nomenclature proposal for sars-cov-2 lineages to assist genomic epidemiology. **Nature Microbiology**, Nature Publishing Group, v. 5, n. 11, p. 1403–1407, 2020.
- 39 World Health Organization. **Tracking SARS-CoV-2 variants**. último acesso em 08 de nov. de 2023. Disponível em: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-%28covid-19%29-variants-of-sars-cov-2?gclid=Cj0KCQjwx5qoBhDyARIsAPbMagCLlJnvJwd7Z_Y3kIosIREkdjCv1BG3jU-ocG7HOPwNW0speOrwROIaAuY9EALw_wcB>.
- 40 SWOFFORD, D. L.; OLSEN, G. J.; WADDELL, P. J.; HILLIS, D. M. Phylogenetic inference. In: HILLIS, D. M.; MORITZ, C.; MABLE, B. K. (Ed.). **Molecular Systematics**. [S.l.]: Sinauer Associates, 1996. p. 407–514.

- 41 FITCH, W. M. Toward defining the course of evolution: Minimum change for a specific tree topology. **Systematic Biology**, v. 20, n. 4, p. 406–416, 1971.
- 42 FELSENSTEIN, J. Evolutionary trees from dna sequences: A maximum likelihood approach. **Journal of Molecular Evolution**, v. 17, n. 6, p. 368–376, 1981.
- 43 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.
- 44 HUELSENBECK, J. P.; RONQUIST, F.; NIELSEN, R.; BOLLBACK, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. **Science**, v. 294, n. 5550, p. 2310–2314, 2001.
- 45 KINGMAN, J. F. The coalescent. **Stochastic Processes and their Applications**, v. 13, n. 3, p. 235–248, 1982.
- 46 HUSON, D. H.; BRYANT, D. Application of phylogenetic networks in evolutionary studies. **Molecular Biology and Evolution**, v. 23, n. 2, p. 254–267, 2006.
- 47 EISEN, J. A. Horizontal gene transfer among bacteria and its role in biological evolution. **Life on Earth: An Encyclopedia of Biodiversity, Ecology, and Evolution**, v. 19, p. 237–245, 2000.
- 48 GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT press, 2016.
- 49 BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer, 2006.
- 50 MOUNT, D. W. **Bioinformatics: Sequence and Genome Analysis**. 2. ed. [S.l.]: Cold Spring Harbor Laboratory Press, 2004.
- 51 SCHÖLKOPF, B.; SMOLA, A. J. **Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond**. [S.l.]: MIT Press, 2002.
- 52 FALL, A.; ELAWAR, F.; HODCROFT, E. B.; JALLOW, M. M.; TOURE, C. T.; BARRY, M. A.; KIORI, D. E.; SY, S.; DIAW, Y.; GOUDIABY, D.; NIANG, M. N.; DIA, N. Genetic diversity and evolutionary dynamics of respiratory syncytial virus over eleven consecutive years of surveillance in Senegal. **Infection, Genetics and Evolution**, v. 91, p. 104864, jul. 2021. ISSN 15671348. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S1567134821001611>>.
- 53 SHABBIR, M. Z.; RAHMAN, A.-u.; MUNIR, M. A comprehensive global perspective on phylogenomics and evolutionary dynamics of Small ruminant morbillivirus. **Scientific Reports**, v. 10, n. 1, p. 17, dez. 2020. ISSN 2045-2322. Disponível em: <<http://www.nature.com/articles/s41598-019-54714-w>>.
- 54 HUDU, S. A.; NIAZLIN, M. T.; NORDIN, S. A.; HARMAL, N. S.; TAN, S. S.; OMAR, H.; SHAHAR, H.; MUTALIB, N. A.; SEKAWI, Z. Hepatitis E virus isolated from chronic hepatitis B patients in Malaysia: Sequences analysis and genetic diversity suggest zoonotic origin. **Alexandria Journal of Medicine**, v. 54, n. 4, p. 487–494, dez. 2018. ISSN 2090-5068, 2090-5076. Disponível em: <<https://www.tandfonline.com/doi/full/10.1016/j.ajme.2017.07.003>>.

- 55 SALLARD, E.; HALLOY, J.; CASANE, D.; DECROLY, E.; HELDEN, J. van. Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review. **Environmental Chemistry Letters**, v. 19, n. 2, p. 769–785, abr. 2021. ISSN 1610-3653, 1610-3661. Disponível em: <<https://link.springer.com/10.1007/s10311-020-01151-1>>.
- 56 PAEZ-ESPINO, D.; ZHOU, J.; ROUX, S.; NAYFACH, S.; PAVLOPOULOS, G. A.; SCHULZ, F.; MCMAHON, K. D.; WALSH, D.; WOYKE, T.; IVANOVA, N. N.; ELOE-FADROSH, E. A.; TRINGE, S. G.; KYRPIDES, N. C. Diversity, evolution, and classification of virophages uncovered through global metagenomics. **Microbiome**, v. 7, n. 1, p. 157, dez. 2019. ISSN 2049-2618. Disponível em: <<https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0768-5>>.
- 57 TANG, X.; YING, R.; YAO, X.; LI, G.; WU, C.; TANG, Y.; LI, Z.; KUANG, B.; WU, F.; CHI, C.; DU, X.; QIN, Y.; GAO, S.; HU, S.; MA, J.; LIU, T.; PANG, X.; WANG, J.; ZHAO, G.; TAN, W.; ZHANG, Y.; LU, X.; LU, J. Evolutionary analysis and lineage designation of SARS-CoV-2 genomes. **Science Bulletin**, v. 66, n. 22, p. 2297–2311, nov. 2021. ISSN 20959273. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S2095927321001250>>.
- 58 CHO, M.; MIN, X.; SON, H. S. Analysis of evolutionary and genetic patterns in structural genes of primate lentiviruses. **Genes & Genomics**, v. 44, n. 7, p. 773–791, jul. 2022. ISSN 1976-9571, 2092-9293. Disponível em: <<https://link.springer.com/10.1007/s13258-022-01257-6>>.
- 59 YIN, Y.; HE, K.; WU, B.; XU, M.; DU, L.; LIU, W.; LIAO, P.; LIU, Y.; HE, M. A systematic genotype and subgenotype re-ranking of hepatitis B virus under a novel classification standard. **Heliyon**, v. 5, n. 10, p. e02556, out. 2019. ISSN 24058440. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S2405844019362164>>.
- 60 BEDOYA-PILOZO, C. H.; MAGÜES, L. G. M.; ESPINOSA-GARCÍA, M.; SÁNCHEZ, M.; VALDIVIEZO, J. V. P.; MOLINA, D.; IBARRA, M. A.; QUIMIS-PONCE, M.; ESPAÑA, K.; MACIAS, K. E. P.; FLORES, N. V. C.; ORLANDO, S. A.; PENAHERRERA, J. A. R.; CHEDRAUI, P.; ESCOBAR, S.; CHANGO, R. D. L.; RAMIREZ-MORÁN, C.; ESPINOZA-CAICEDO, J.; SÁNCHEZ-GILER, S.; LIMIA, C. M.; ALEMÁN, Y.; SOTO, Y.; KOURI, V.; CULASSO, A. C.; BADANO, I. Molecular epidemiology and phylogenetic analysis of human papillomavirus infection in women with cervical lesions and cancer from the coastal region of Ecuador. **Revista Argentina de Microbiología**, v. 50, n. 2, p. 136–146, abr. 2018. ISSN 03257541. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S0325754117301372>>.
- 61 POTDAR, V.; VIPAT, V.; RAMDASI, A.; JADHAV, S.; PAWAR-PATIL, J.; WALIMBE, A.; PATIL, S.; CHOUDHURY, M.; SHASTRI, J.; AGRAWAL, S.; PAWAR, S.; LOLE, K.; ABRAHAM, P.; CHERIAN, S. Phylogenetic classification of the whole-genome sequences of SARS-CoV-2 from India & evolutionary trends. **Indian Journal of Medical Research**, v. 153, n. 1, p. 166, 2021. ISSN 0971-5916. Disponível em: <https://journals.lww.com/ijmr/Fulltext/2021/01000/Phylogenetic_classification_of_the_whole_genome.14.aspx>.
- 62 LICHTBLAU, D. Alignment-free genomic sequence comparison using FCGR and signal processing. **BMC Bioinformatics**, v. 20, n. 1, p. 742, dez. 2019. ISSN 1471-2105. Disponível em: <<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3330-3>>.
- 63 KIM, J.; CHEON, S.; AHN, I. NGS data vectorization, clustering, and finding key codons in SARS-CoV-2 variations. **BMC Bioinformatics**, v. 23, n. 1, p. 187, dez. 2022. ISSN 1471-2105. Disponível em: <<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04718-7>>.

- 64 DIMITROV, K. M.; ABOLNIK, C.; AFONSO, C. L.; ALBINA, E.; BAHL, J.; BERG, M.; BRIAND, F.-X.; BROWN, I. H.; CHOI, K.-S.; CHVALA, I.; DIEL, D. G.; DURR, P. A.; FERREIRA, H. L.; FUSARO, A.; GIL, P.; GOUJGOULOVA, G. V.; GRUND, C.; HICKS, J. T.; JOANNIS, T. M.; TORCHETTI, M. K.; KOLOSOV, S.; LAMBRECHT, B.; LEWIS, N. S.; LIU, H.; LIU, H.; MCCULLOUGH, S.; MILLER, P. J.; MONNE, I.; MULLER, C. P.; MUNIR, M.; REISCHAK, D.; SABRA, M.; SAMAL, S. K.; ALMEIDA, R. Servan de; SHITTU, I.; SNOECK, C. J.; SUAREZ, D. L.; BORM, S. V.; WANG, Z.; WONG, F. Y. Updated unified phylogenetic classification system and revised nomenclature for Newcastle disease virus. **Infection, Genetics and Evolution**, v. 74, p. 103917, out. 2019. ISSN 15671348. Disponível em: <<https://linkinghub.elsevier.com/retrieve/pii/S1567134819301388>>.
- 65 PEFFERS, K.; TUUNANEN, T.; ROTHENBERGER, M. A.; CHATTERJEE, S. A design science research methodology for information systems research. **Journal of management information systems**, Taylor & Francis, v. 24, n. 3, p. 45–77, 2007.
- 66 SILVA, D. D.; LOPES, E. L.; JUNIOR, S. S. B. Pesquisa Quantitativa: Elementos, Paradigmas e Definições. **Revista de Gestão e Secretariado**, v. 05, n. 01, p. 01–18, abr. 2014. ISSN 21789010, 21789010. Disponível em: <<http://www.revistagesec.org.br/ojs-2.4.5/index.php/secretariado/article/view/297>>.
- 67 ROSSUM, G. van. **Python Programming Language**. Python Software Foundation, 1991. Disponível em: <<https://www.python.org/>>.
- 68 JUPYTER, P. **Jupyter Notebook: Interactive Computing**. [S.l.], 2001. Disponível em: <<https://jupyter.org/>>.
- 69 COCK, P. J. A.; ANTÃO, T.; CHANG, J. T.; CHAPMAN, B. A.; COX, C. J.; DALKE, A.; FRIEDBERG, I.; HAMELRYCK, T.; KAUFF, F.; WILCZYNSKI, B.; HOON, M. J. L. de. Biopython: freely available python tools for computational molecular biology and bioinformatics. **Bioinformatics**, v. 25, n. 11, p. 1422–1423, 2009.
- 70 MUTHUKADAN, B. **Selenium with Python**. [S.l.]. Disponível em: <<https://selenium-python.readthedocs.io/>>.
- 71 FONSECA, D. d. S. Identificação das regiões gênicas com maior informação filogenética para simplificação da genotipagem in silico. 2021.
- 72 WALT, S. V. D.; COLBERT, S. C.; VAROQUAUX, G. The numpy array: a structure for efficient numerical computation. **Computing in science & engineering**, IEEE, v. 13, n. 2, p. 22–30, 2011.
- 73 YANG, Y.; GUAN, X.; YOU, J. Clope: a fast and effective clustering algorithm for transactional data. In: **Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.: s.n.], 2002. p. 682–687.
- 74 DILL: Serialize All of Python. último acesso em 30 de nov. de 2023. <<https://pypi.org/project/dill/>>.
- 75 MINH, B. Q.; SCHMIDT, H. A.; CHERNOMOR, O.; SCHREMPE, D.; WOODHAMS, M. D.; HAESELER, A. von; LANFEAR, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. **Molecular Biology and Evolution**, v. 37, n. 5, p. 1530–1534, 02 2020. ISSN 0737-4038. Disponível em: <<https://doi.org/10.1093/molbev/msaa015>>.