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, BAHIA, BRASIL

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 Frias Suárez

Coorientador: PhD Vagner Fonseca

SALVADOR, BAHIA, BRASIL

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 Frias 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 Frias Suárez (Orientador)
Universidade do Estado da Bahia – UNEB

PhD Vagner Fonseca (Coorientador)
Universidade Coorientador - SIGLA

Membro da Banca 1
IES do Membro da Banca 1

Membro da Banca 2
IES do Membro da Banca 2

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. 17
Figura 2 – Tabela de Códons. 18
Figura 3 – Estrutura do coronavírus. 19
Figura 4 – Pipeline de Download das Sequências Genômicas. 24

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

A	adenina
BV-BRC	Bacterial and Viral Bioinformatics Resource Center
C	citosina
COVID-19	Coronavirus Disease 2019
DNA	Ácido Desoxirribonucleico *(Deoxyribonucleic Acid)*
DSR	Design Science Research
G	guanina
ML	Máxima Verossimilhança
mRNA	Ácido Ribonucleico Mensageiro *(Mensager Ribonucleic Acid)*
RNA	Ácido Ribonucleiro *(Ribonucleic Acid)*
T	timina
tRNA	Ácido Ribonucleiro Transportador *(Transporter Ribonucleic Acid)*

SUMÁRIO

1 INTRODUÇÃO 12
2 FUNDAMENTAÇÃO TEÓRICA 16
2.1 Biologia Molecular 16
2.2 Vírus 18
2.2.1 SARS-CoV-2 19
2.3 Filogenia 19
2.4 Trabalhos Correlatos 20
3 DESCRIÇÃO DO PROJETO 21
3.1 Metodologia 21
3.2 Materiais e Métodos 22
3.3 Plano de Implementação 23
3.3.1 Montagem e Preparação do Dataset 23
4 CONSIDERAÇÕES FINAIS 26
REFERÊNCIAS 27

1 INTRODUÇÃO

Os problemas impostos pela pandemia do 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. 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 (1).

Diante disso, 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. Segundo Hall e Barlow(2) 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."

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. Essas observações são realizadas com base em dados como sequências de Ácido Desoxirribonucleico *(Deoxyribonucleic Acid)* (DNA) e Ácido Ribonucleiro *(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 timina é substituída pela uracila (U) (3).

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(4) elas tem função importante porque apresentam de forma sucinta e particular a evolução dos descendentes partindo de ancestrais em comum.

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 (1).

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. Mesmo quando aplicados adequadamente, são mal explicados e, portanto, mal compreendidos. (2, p. 1) Além disso, por trabalhar com grandes quantidades de dados, os métodos utilizados devem ser avaliados também em relação ao seu custo computacional.

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.(1)
- Fall et al.(5)
- Shabbir et al.(6)
- Hudu et al.(7)
- Sallard et al.(8)
- Paez-Espino et al.(9)
- Tang et al.(10)
- Cho et al.(11)

Já em Yin et al.(12) e Bedoya-Pilozo et al.(13), 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.(14) 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(15) expos o

Frequency Chaos Game Representation e Kim et al.(16) a floresta aleatória. Por fim, Dimitrov et al.(17) comparou três modelos para reconstrução de árvores filogenéticas: junção de vizinhos; ML e inferência bayesiana.

As soluções até então desenvolvidas, são guiadas pela reconstrução das árvores filogenéticas construídas a partir das mutações de nucleotídeos. 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, que 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. Necessita-se então, de pesquisas e desenvolvimento de ferramentas que realizem uma classificação de sequências genéticas com base no uso/frequência 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 e validar 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.
  - (vi) Disponibilizar o modelo como uma ferramenta web de fácil acesso.

Atingir o objetivo de desenvolver um método de construção de árvores com base nas distâncias obtidas a partir da diferença do uso de códons, contribuiria com a tarefa de classificação de cepas para a vigilância sanitária, 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. Os métodos atuais ainda demandam de um alto custo computacional, sendo assim, existe a necessidade de desenvolver outros mais baratos e que

possam suportar o volume crescente de dados (sequências). Posto isso, 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. A hipótese referente à menor complexidade computacional do novo método deverá ser testada no trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 BIOLOGIA MOLECULAR

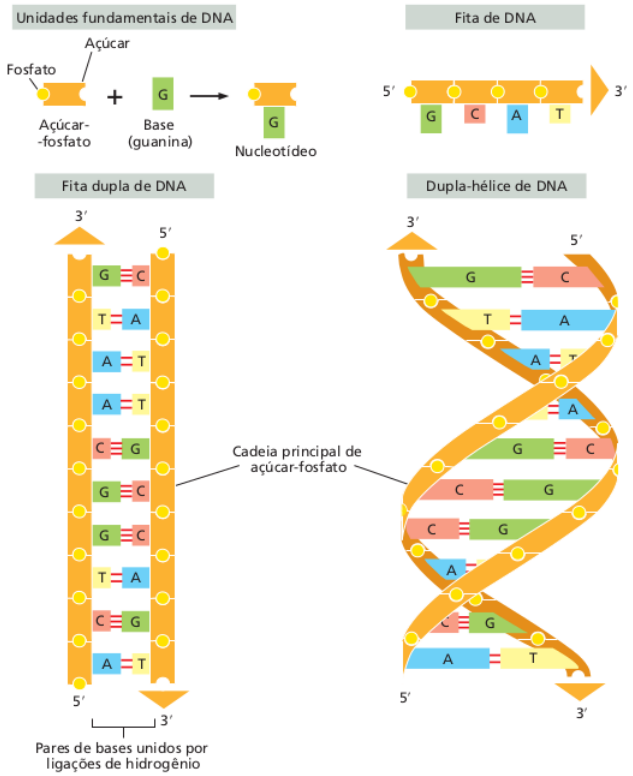
A Biologia Molecular é um ramo da biologia que lida e investigas os processos e mecanismos moleculares relacionados à estrutura, função e interações das biomoléculas presentes nos organismos vivos. Consiste principalmente em estudar as interações entre os vários sistemas da célula, partindo da relação 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 adenina (A), timina (T), citosina (C) ou guanina (G). A estrutura do DNA é mantida por pontes de hidrogênio entre as bases complementares, com a adenina pareando sempre com a timina e a citosina pareando sempre com a guanina.

A 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 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. A análise do genoma desempenha um papel fundamental na genética, na biologia molecular e na compreensão da hereditariedade e da evolução. (18)

As informações contidas no DNA é copiada 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 *(Mensager Ribonucleic Acid)* (mRNA). O mRNA

Figura 1 – Estrutura do DNA.



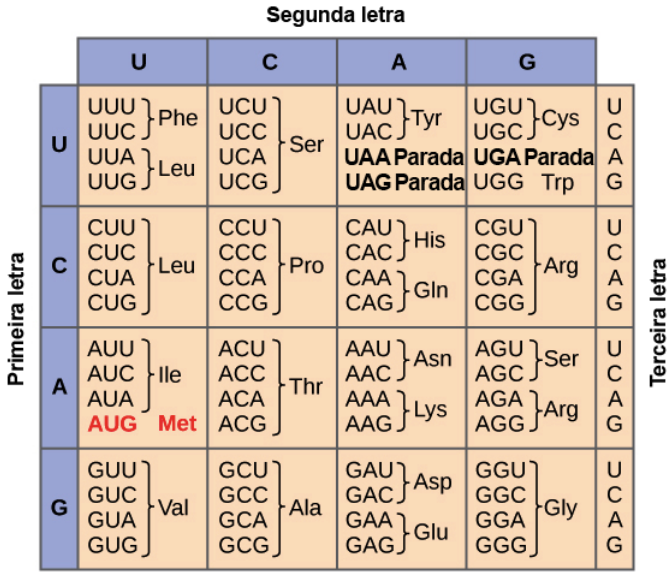
Fonte: Retirada de Alberts et al.(18)

é uma cópia do DNA que carrega a sequência de bases nitrogenadas correspondente a um gene específico. 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, chamados de códons. Os códons são sequências de três nucleotídeos consecutivos no RNA que correspondem a um aminoácido específico. Existem 64 códons possíveis, correspondentes a 20 aminoácidos diferentes como apresentado na Figura 2, além de sinais de início e parada da tradução. 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.

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.

Cada códon especifica um aminoácido distinto. Os aminoácidos são transportados para o ribossomo por moléculas de Ácido Ribonucleiro Transportador *(Transporter Ribonucleic*

Figura 2 – Tabela de Códons.



Fonte: Adaptade de OpenStax(19)

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 (18).

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 (20). 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 (3):

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 transcritas em moléculas de mRNA, que são

utilizadas para a síntese de proteínas virais.

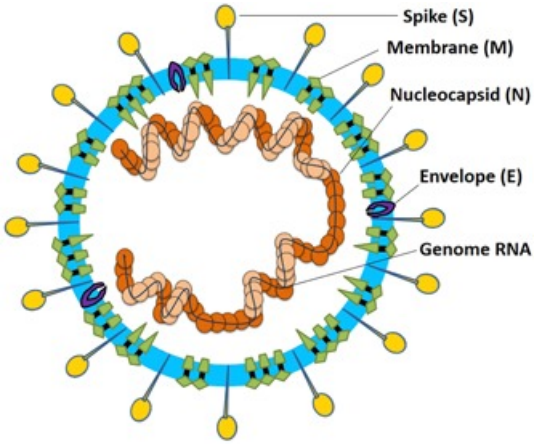
4. **Montagem:** as proteínas virais se unem para formar novas partículas virais.
5. **Liberação:** as novas partículas virais são liberadas da célula hospedeira, para a montagem de novos vírus e para a modificação do ambiente celular para garantir a sua replicação.

2.2.1 SARS-CoV-2

O SARS-CoV-2 é um vírus da família Coronaviridae, que causa a doença chamada Coronavirus Disease 2019 (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 (21, 22).

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), que são responsáveis pela ligação do vírus às células hospedeiras. 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.

Figura 3 – Estrutura do coronavírus.



Fonte: Retirada de Li et al.(23)

2.3 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 filogené-

tica molecular é uma abordagem utilizada para inferir a filogenia com base em informações moleculares, como sequências de DNA, RNA e proteínas(24).

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(25).

2.4 TRABALHOS CORRELATOS

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 Design Science Research (DSR). Essa metodologia, proporciona um framework teórico e prático para a criação de artefatos inovadores, como métodos, modelos ou frameworks, visando resolver problemas específicos (26). 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.
2. Desenvolvimento dos artefatos.
3. Avaliação do artefato.
4. Apresentar contribuições científicas.

Também será utilizada analises 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. (27)

3.2 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. (28)

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 (29).

O python possue uma gama de bibliotecas que facilitam a implementação de soluções complexas. A seguir serão apresentadas a 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. (30)
- **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. (31)

3.3 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 todos 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.

3.3.1 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.

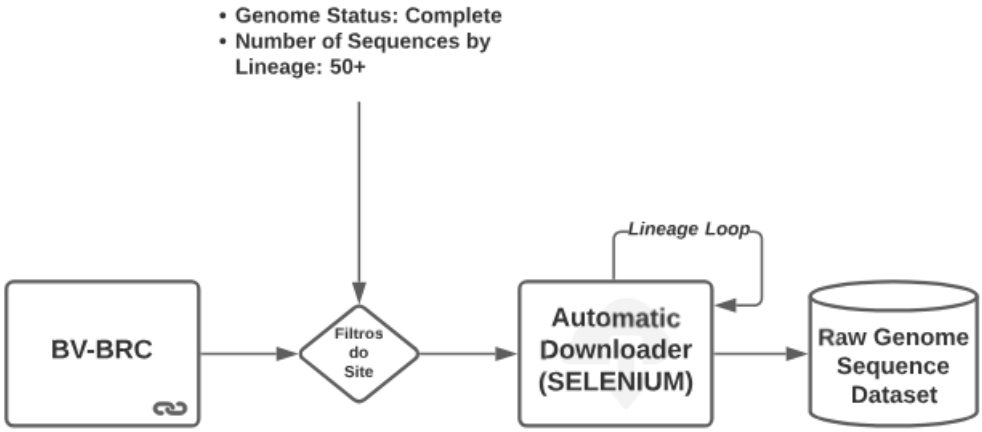
Inicialmente, foi realizada uma analise do BV-BRC, para entende a sua estrutura e verificar também se era possível realizar o download de todas as sequências queridas de forma manual. Foi verificado que o site possuía uma área de seleção de filtros, e foi definido que só seriam selecionadas sequências 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 4. Isto posto, foi realizado uma sequência de passos conhecidos como *Web Scrapping* utilizado o Python juntamente com o Selenium, apresentados em seguida:

1. teste 01
2. teste 02

Ao final do processo de montagem do dataset com sequências genômicas completas, o mesmo ficou com as seguintes informações apresentadas na tabela x.

Figura 4 – Pipeline de Download das Sequências Genômicas.



Fonte: O Autor

A montagem do dataset se deu com a construção de pipelines de processamento das sequências. Um pipeline inicial foi construído e será apresentado a seguir, o mesmo foi descontinuado após se verificar que o tempo de processamento ficaria inviável.

- **Montagem e Preparação do Dataset:** Abaixo, está apresentado o passo a passo que foi realizado nesta etapa:
  - Analise do site BV-BRC.
  - Desenvolvimento de um script Python juntamente com o Selenium, para o download automático das sequências genomicas no BV-BRC.
  - Filtragem das sequências genômicas duplicadas, ou seja, que contenham exatamente, a mesma sequência de nucleotídeos, mantendo apenas 1(uma) das repetidas e construindo um dataset de sequências genomicas únicas. (Filtro 01)
  - Alinhamento das sequências genomicas utilizando o Minimap2.
  - Implementado procedimento de extração do gene de interesse (Spike), com base em uma sequência Spike de referência, das sequências únicas.
  - Filtragem das sequências genicas duplicadas, ou seja, que contenham exatamente, a mesma sequência de nucleotídeos, mantendo apenas 1(uma) das repetidas e construindo um dataset de sequências genicas únicas. (Filtro 02)
  - Filtragem das sequências genicas de má qualidade. Foram removidas as sequências que possuiam mais de 30 N's consecultivos ou que ficaram com a quantidade de nucleotídeos discrepantes em relação a referência genica. (Filtro 03)
  - Criar arquivo de treinamento com linhagens diferentes e um arquivo de anotação.
- **Desenvolvimento do Modelo:** Para a realização do desenvolvimento do modelo serão

realizados os seguintes passos:

  - Levantamento dos requisitos.
  - Definir a arquitetura e a abordagem do modelo de classificação baseado em códons.
  - Implementar o modelo utilizando uma biblioteca ou framework adequado.
  - Desenvolver algoritmo para traduzir as sequências de DNA em sequências de códons.
  - Realizar treinamento do modelo utilizando os dados preparados.
  - Avaliar o desempenho do modelo utilizando métricas apropriadas.
  - Identificar possíveis problemas e realizar ajustes no modelo.

- **Análise comparativa entre o método proposto e outro método existente:** Será realizada uma analise com um conjunto de dados, onde será realizado analises estatísticas para verificação de melhorias, ou não, do novo método proposto analisando os seguintes aspectos:
  - Comparação dos métodos de agrupamento adotados, avaliando sua eficácia na formação de clusters e na identificação de padrões ou similaridades nas sequências.
  - Avaliação do custo computacional (tempo de execução e recursos requeridos) para a classificação das sequências em cada método.
  - Comparação da eficiência computacional entre os métodos, considerando a escalabilidade e o desempenho em grandes volumes de dados.

4 CONSIDERAÇÕES FINAIS

Nesta fase inicial do projeto, foi realizada com sucesso a montagem do dataset de genes virais a serem estudados. Através do uso de scripts e procedimentos adequados, foram baixadas sequências classificadas de genoma completo do vírus de uma base de dados pública, como o BV-BRC, e em seguida, filtradas para manter apenas as sequências únicas. Além disso, o gene de interesse foi extraído utilizando a técnica de blast.

A montagem do dataset é uma etapa crucial para o desenvolvimento da ferramenta de análise de genes virais baseada em códons. Ao obter uma base de dados representativa e de qualidade, garantimos que a análise subsequente seja feita em um conjunto abrangente de sequências relevantes.

No entanto, é importante ressaltar que essa é apenas uma etapa inicial do projeto e que há muito trabalho a ser feito nas próximas fases. A análise comparativa entre o método proposto e outro método existente, bem como a implementação do modelo de classificação e as etapas subsequentes do processo, ainda estão por vir.

Com base na conclusão desta etapa, temos uma base sólida de dados que nos permitirá avançar no desenvolvimento do projeto. O próximo passo será a implementação do modelo de análise de genes virais baseado em codons e a realização das etapas subsequentes, como o alinhamento, a tradução e a extração de códigos únicos para a classificação e agrupamento das sequências.

A montagem bem-sucedida do dataset é um marco importante para o progresso do projeto, fornecendo a base necessária para a etapa seguinte. Com um dataset representativo e de qualidade, podemos agora prosseguir para a implementação do modelo e a análise comparativa com outras abordagens existentes.

REFERÊNCIAS

1 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>.

2 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>.

3 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.

4 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>.

5 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>.

6 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>.

7 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>.

8 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>.

9 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>.

10 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>.

11 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>.

12 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>.

13 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>.

14 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>.

15 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>.

16 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>.

17 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>.

18 ALBERTS, B.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. **Biologia Molecular da Célula**. [S.l.]: Artmed, 2017.

19 OPENSTAX. **The Genetic Code**. último acesso em 05 de jul. de 2023. <https://openstax.org/books/biology/pages/15-1-the-genetic-code>.

20 KNIPE, P. M. H. D. M. **Fields Virology**. 6. ed. [S.l.]: LIPPINCOTT WILLIAMS & WILKINS, 2022. Vol I and II. ISBN 9781451105636,1451105630,2013003842.

21 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.

22 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.

23 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.

24 FELSENSTEIN, J. **Inferring Phylogenies**. 2. ed. [S.l.]: Sinauer Associates, 2004. ISBN 0878931775,9780878931774.

25 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.

26 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.

27 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>.

28 ROSSUM, G. van. **Python Programming Language**. Python Software Foundation, 1991. Disponível em: <https://www.python.org/>.

29 JUPYTER, P. **Jupyter Notebook: Interactive Computing**. [S.l.], 2001. Disponível em: <https://jupyter.org/>.

30 COCK, P. J. A.; ANTAO, 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.

31 MUTHUKADAN, B. **Selenium with Python**. [S.l.]. Disponível em: <https://selenium-python.readthedocs.io/>.