



UNIVERSIDADE DO ESTADO DA BAHIA (UNEB)
DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA, CAMPUS I
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

KYARA CARDOZO CAMPOS

DESENVOLVIMENTO DE UMA APLICAÇÃO WEB PARA A
CARACTERIZAÇÃO GENÔMICA DO VÍRUS OROPOUCHE.

SALVADOR
2024

KYARA CARDOZO CAMPOS

**DESENVOLVIMENTO DE UMA APLICAÇÃO WEB PARA A
CARACTERIZAÇÃO GENÔMICA DO VÍRUS OROPOUCHE.**

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação do Departamento de Ciências Exatas e da Terra (DCET) - Campus I, da Universidade do Estado da Bahia (UNEB), como requisito à obtenção do grau de bacharel em Sistemas de Informação.
Área de concentração: Sistemas de Informação.

Orientador: Prof. Dr. Vagner de Souza Fonseca

SALVADOR

2024

ATA DE DEFEFA: TRABALHO DE CONCLUSÃO DE CURSO: 2024.2

Aos vinte e sete dias do mês de dezembro do ano de dois mil e vinte e quatro, reuniu-se no Auditório José Rocha Lorangeira, Prof. Dr. Vagner de Souza Fonseca – Orientador, Profa Dra. Maria Inês Valderrama Restovic, Profa. Dra. Marta Giovanetti, Prof. Dr. Filipe Ferreira de Almeida Rego para avaliarem o trabalho da discente: **KYARA CARDOZO CAMPOS**, intitulado: **"DESENVOLVIMENTO DE UMA APLICAÇÃO WEB PARA A CARACTERIZAÇÃO GENÔMICA DO VÍRUS OROPOUCH"**, a banca efetuou os questionamentos pertinentes e, com base no desempenho do aluno, deliberou a seguinte nota: 9,5.

Salvador, vinte e sete dias do mês de dezembro de dois mil e vinte quatro

Assinaturas:



Prof. Dr. Vagner de Souza Fonseca
Orientador - Universidade do Estado da Bahia (UNEB)



Profa. Dra. Maria Inês Valderrama Restovic
Membro Interno - Universidade do Estado da Bahia (UNEB)



Profa. Dra. Marta Giovanetti
Membro Externo - University of Campus Bio-Medico di Roma (UCBM)



Prof. Dr. Filipe Ferreira de Almeida Rego
Membro Externo - Escola Bahiana de Medicina e Saúde Pública

AGRADECIMENTOS

Expresso minha mais sincera gratidão ao dedicar esta seção de agradecimentos do meu Trabalho de Conclusão de Curso. Esta trajetória acadêmica foi possível graças ao incomensurável apoio e incentivo de pessoas fundamentais, cuja presença e contribuição foram indispensáveis ao longo do percurso.

Agradeço primeiramente à minha família, a meu irmão Sérgio, a minha mãe Rosângela e a meu pai Dagoberto por serem a minha base, pelo amor incondicional, e por sempre acreditarem no meu potencial. Vocês foram a força que impulsionou cada passo desta caminhada.

Ao meu orientador, expresso minha profunda gratidão pela paciência, sabedoria e valiosa orientação ao longo de todo o processo de pesquisa. Seu suporte foi essencial para o desenvolvimento deste trabalho.

Aos professores e professoras que, com seus conhecimentos e experiências, contribuíram significativamente para a minha formação acadêmica, registro meus sinceros agradecimentos. Cada aula, conselho foi fundamental para o meu crescimento intelectual e acadêmico.

Aos meus colegas de curso, expresso minha mais profunda gratidão pelas ricas trocas de ideias, pelos momentos de estudo compartilhados e pelo apoio constante. As experiências que vivemos juntos – dos desafios às conquistas – tornaram essa jornada acadêmica não apenas mais significativa, mas também profundamente enriquecedora. Obrigado por fazerem parte dessa etapa tão especial.

Por fim, manifesto meus agradecimentos a todos aqueles que, de alguma forma, contribuíram para o êxito deste trabalho. Cada palavra de encorajamento e cada gesto de apoio foram essenciais para a concretização desta trajetória.

Este TCC não é apenas um marco acadêmico, mas uma realização coletiva. A todos vocês que estiveram comigo, o meu mais sincero agradecimento.

*"O importante não é ser o primeiro ou primeira, o importante é abrir
caminhos."
(Conceição Evaristo)*

RESUMO

O vírus Oropouche (OROV), um arbovírus pertencente à família Peribunyaviridae e ao gênero Orthobunyavirus, é um patógeno negligenciado de relevância significativa para a saúde pública no Brasil e em outras regiões da América Latina, frequentemente associado a surtos em áreas tropicais e subtropicais. As condições ambientais que favorecem sua disseminação, como altas temperaturas e elevada umidade, estão profundamente ligadas às mudanças climáticas, que intensificam o ambiente propício para a proliferação de seus vetores. O OROV tem se destacado por sua capacidade de provocar epidemias em áreas urbanas e rurais, apresentando sintomas semelhantes aos da dengue, como febre alta, cefaleia, dores musculares e articulares, além de erupções cutâneas. Em vista do crescente aumento da incidência de arboviroses e da necessidade de ferramentas eficientes para monitoramento e controle, este trabalho propõe o desenvolvimento de uma aplicação web de bioinformática para a caracterização genômica do OROV. A ferramenta foi projetada para facilitar o diagnóstico e o desenvolvimento de estratégias de prevenção e tratamento de forma eficiente, possibilitando a genotipagem do OROV com base em modelos evolutivos apropriados, de forma automática, precisa e rápida. Para isso, sequências do vírus foram selecionadas em bancos de dados públicos por meio de um sistema automatizado de mineração de dados, resultando em um conjunto eficiente de sequências de referência usadas no desenvolvimento da aplicação, que caracteriza os três segmentos do OROV (L, M e S). A aplicação web demonstrou alta eficácia na caracterização de linhagens genéticas, oferecendo resultados que sustentam o entendimento da variabilidade genética do vírus. A ferramenta possibilitou a criação de árvores filogenéticas, permitindo uma análise detalhada da evolução viral e a identificação de potenciais mutações de interesse epidemiológico. Os resultados indicam que a ferramenta desenvolvida é uma solução promissora para o estudo de arbovírus negligenciados como o OROV, oferecendo suporte ao monitoramento epidemiológico e à implementação de estratégias de resposta rápida em casos de surtos. A pesquisa ressalta a importância de manter a atualização contínua da aplicação para incorporar novos dados e melhorias, adaptando-se às evoluções tecnológicas e às mudanças nas dinâmicas virais. Essa ferramenta contribui para o fortalecimento da infraestrutura de vigilância em saúde pública e evidencia a bioinformática como um aliado estratégico na luta contra doenças virais impulsionadas por mudanças climáticas e outros fatores ambientais.

Palavras-chave: Vírus Oropouche; Ferramenta Web; Filogenia; Genotipagem; Caracterização Genômica.

ABSTRACT

The Oropouche virus (OROV), an arbovirus belonging to the Peribunyaviridae family and the Orthobunyavirus genus, is a neglected pathogen of significant relevance to public health in Brazil and other regions of Latin America, frequently associated with outbreaks in tropical and subtropical areas. The environmental conditions that favor its spread, such as high temperatures and humidity, are deeply linked to climate change, which intensifies the environment conducive to the proliferation of its vectors. OROV has gained attention for its ability to trigger epidemics in both urban and rural areas, presenting symptoms similar to dengue, such as high fever, headache, muscle and joint pain, and skin rashes. Given the increasing incidence of arboviruses and the need for efficient tools for monitoring and control, this study proposes the development of a bioinformatics web application for the genomic characterization of OROV. The tool was designed to facilitate diagnosis and the efficient development of prevention and treatment strategies, enabling the genotyping of OROV based on appropriate evolutionary models automatically, accurately, and rapidly. To achieve this, virus sequences were selected from public databases through an automated data mining system, resulting in an efficient set of reference sequences used for the development of the application, which characterizes the three segments of OROV (L, M, and S). The web application demonstrated high efficacy in the characterization of genetic lineages, providing results that support the understanding of the virus's genetic variability. The tool enabled the creation of phylogenetic trees, allowing detailed analysis of viral evolution and the identification of potential mutations of epidemiological interest. The results indicate that the developed tool is a promising solution for the study of neglected arboviruses such as OROV, supporting epidemiological monitoring and the implementation of rapid response strategies in outbreak scenarios. The research emphasizes the importance of continuous updates to the application to incorporate new data and improvements, adapting to technological advancements and changes in viral dynamics. This tool contributes to strengthening the public health surveillance infrastructure and highlights bioinformatics as a strategic ally in the fight against viral diseases driven by climate change and other environmental factors.

Key-words: Oropouche virus; Typingtool; Phylogenetics; Genotyping; Genomic Characterization.

LISTA DE ILUSTRAÇÕES

Figura 1 – Número de casos confirmados para o vírus Oropouche, Brasil, 2023 e 2024.	15
Figura 2 – Cidades na Bahia com casos confirmados da doença.	16
Figura 3 – Estrutura do vírus nus e envelopados.	22
Figura 4 – Ciclo de replicação do vírus.	23
Figura 5 – Glóbulos brancos liberando substâncias que desencadeiam o processo inflamatório no local infectado.	24
Figura 6 – Ligação do linfócito ao antígeno do patógeno.	25
Figura 7 – Estrutura do vírus Oropouche.	27
Figura 8 – Representação esquemática da partícula de <i>Orthobunyavirus</i> evidenciando os três segmentos de RNA (PRNA, MRNA e GRNA) associados à proteína L. Circundando cada molécula de RNA, observa-se a proteína N, e no detalhe, as duas glicoproteínas Gn e Gc emergindo do envelope (E).	28
Figura 9 – Ciclo de vida do vírus Oropouche na interface selvagem e doméstica.	29
Figura 10 – Aplicações de ferramentas de bioinformática.	34
Figura 11 – Etapas de pré-processamento, redução e transformação de dados.	35
Figura 12 – Ferramentas de alinhamentos mais comuns em análises de sequências genéticas.	35
Figura 13 – Número aproximado de artigos de aprendizagem profunda publicados por ano. O número de artigos é baseado nos resultados da pesquisa em http://www.scopus.com com as duas consultas: ‘ <i>Deep Learning</i> ’, ‘ <i>Deep Learning</i> ’ AND ‘ <i>bio*</i> ’.	39
Figura 14 – Ilustração da SNPs.	43
Figura 15 – Nomenclatura padronizada sugerida.	48
Figura 16 – Resultado da nova nomenclatura padronizada sugerida.	49
Figura 17 – Processo de mineração de dados do Genbank.	52
Figura 18 – Processo de mineração de sequências e construção do conjunto de dados utilizado na ferramenta de genotipagem do vírus do Oropouche.	56
Figura 19 – Escolha da sequência do vírus do Oropouche na base de dados GenBank.	58
Figura 20 – Processo de verificação após a seleção da sequência.	59
Figura 21 – Script para inserção, deleção e substituição.	60
Figura 22 – Adição da sequência alterada na plataforma.	61
Figura 23 – A interface web da ferramenta de genotipagem do vírus do Oropouche.	67

Figura 24 – Processo inserção de dados.	68
Figura 25 – Formatos de exportações disponíveis.	68
Figura 26 – Exemplo de exportação em formato Excel na interface web da ferramenta de genotipagem do vírus do Oropouche.	69
Figura 27 – Exemplo de exportação em formato XML na interface web da ferramenta de genotipagem do vírus do Oropouche.	69
Figura 28 – Exemplo da classificação da sequências e visualização dos resultados na ferramenta de genotipagem do vírus do Oropouche.	70
Figura 29 – Análise filogenética de todas as sequências de entrada.	71
Figura 30 – Atribuição de sequência.	72
Figura 31 – Processo de recuperar Id e execução de trabalhos anteriores.	72
Figura 32 – Interface web e suas disposições disponíveis.	73

LISTA DE TABELAS

Tabela 1	– Segmento L considerando sequências com baixa qualidade	75
Tabela 2	– Segmento L removendo sequências com baixa qualidade	75
Tabela 3	– Segmento M considerando sequências com baixa qualidade	75
Tabela 4	– Segmento M removendo sequências com baixa qualidade	76
Tabela 5	– Segmento S considerando sequências com baixa qualidade	76
Tabela 6	– Segmento S removendo sequências com baixa qualidade	77

LISTA DE ABREVIATURAS E SIGLAS

DTN	Doenças Tropicais Negligenciadas
OMS	Organização Mundial da Saúde
OROV	Vírus Oropouche
CHIKV	Vírus Chikungunya
ZIKV	Vírus da Zika
DENV	Vírus da Dengue
MAYV	Vírus Mayaro
VLAC	Vírus La Crosse
ARN	Ácido Ribonucleico
YFV	Vírus da febre amarela
ROCV	Vírus Rocio
IH	Inibição de Hemaglutinação
FC	Fixação do Complemento
SN	Soroneutralização em Camundongos
ELISA	Enzyme-linked immunosorbent assay
RT-PCR	transcription-polymerase chain reaction
PCA	Análise de Componentes Principais
BLAST	Basic Local Alignment Search Tool
ClustalW	Multiple Sequence Alignment
PAUP*	Phylogenetic Analysis Using
MAFFT	Multiple Alignment using Fast Fourier Transform
BWA	Burrows-Wheeler Aligner
NGS	Sequenciamento de Nova Geração
SNPs	Polimorfismos de Nucleotídeo Único

INDELs	Inserções e Deleções
GATK	Genome Analysis Toolkit
DNA	Ácido Desoxirribonucleico
RNA	Ácido Ribonucleico
NJ	Neighbor Joining
ML	Máxima verossimilhança
GN	Gene de Glicoproteína
CRFs	Formas Recombinantes Circulantes
RVFV	Rift Valley Fever Virus
ML	Machine Learning
IA	Inteligência Artificial
EHRs	Registros Eletrônicos de Saúde
NGS	Sequenciamento de Nova Geração
BWA	Burrows-Wheeler Aligner
RNA-seq	Sequenciamento de RNA
DSL	Linguagem Específica de Domínio
JWT	Java Web Toolkit
EST	Expressed Sequence Tags
PCR	Reação em Cadeia da Polimerase
NGS	Sequenciamento de Nova Geração
MNPs	Polimorfismos de Nucleotídeos Múltiplos
IGSP	Projeto de Sequenciamento do Genoma da Gripe
TRF	Transformada Rápida de Fourier
BME	Evolução Mínima Equilibrada
LGPD	Lei Geral de Proteção de Dados
LACEN	Laboratórios Centrais de Saúde Pública

SUMÁRIO

1	INTRODUÇÃO	13
1.1	Justificativa	16
1.2	Objetivos	19
1.2.1	<i>Objetivo Geral</i>	19
1.2.2	<i>Objetivos Específicos</i>	19
2	MAPEAMENTO DO VÍRUS OROPOUCHE: EXPLORANDO UM PATÓGENO NEGLIGENCIADO	21
2.1	Estrutura do vírus	21
2.2	Métodos de infecção e replicação	27
2.3	Gênero da família do Oropouche	30
2.4	Atuação da Bioinformática	33
2.5	Métodos de Machine Learning e Deep Learning em Bioinfor- mática: Inovações em Análises Filogenéticas e Predição de Genótipos	38
2.6	Unindo Ferramentas: Pipelines Bioinformáticos Flexíveis . . .	40
2.7	Genes e Justiça: Ética, Privacidade e Segurança em Dados Genômicos Humanos	44
2.8	Genomas em Comparação: Explorando a Evolução e Adaptação Viral	46
2.9	Trabalhos Correlatos	49
3	METODOLOGIA	52
3.1	Validação e Experimentos: Onde a Teoria Encontra os Dados	57
3.2	Da Ideia ao Código: Os Estágios Essenciais do Desenvolvimento da Oropouche Virus Typing Tool	63
4	RESULTADOS E DISCUSSÕES	64
4.1	Interface e Usuário	66
4.2	<i>Benchmarking</i> da ferramenta	74
4.3	Limitações Técnicas e Impacto nos Resultados	77
4.4	Direções Futuras e Oportunidades de Melhoria	78
5	CONCLUSÃO	80
	REFERÊNCIAS	81

1 INTRODUÇÃO

Negligenciados é o plural de negligenciado, o mesmo que: abandonados, deixados, descuidados, desprezados, esquecidos, ignorados, largados (Dicio, 2024).

Doenças negligenciadas são enfermidades que não só prevalecem em condições menos favorecidas, como contribuem para a manutenção da desigualdade. Representam um forte entrave ao desenvolvimento dos países segundo Vasconcelos, Kovaleski e Junior (2016), incapacitam ou matam milhões de pessoas e retratam uma necessidade médica importante que permanece não resolvida (Jayashankar *et al.*, 2024).

As doenças negligenciadas fazem parte do grupo das doenças tropicais e subtropicais endêmicas, especialmente nas populações desassistidas da África, Ásia e América Latina. Originalmente proposto na década de 1970 por um programa da Fundação Rockefeller denominado “*The Great Neglected Diseases*”¹, coordenado por Kenneth Warren, este conceito foi retomado em 2001 pela Organização Não Governamental Médicos Sem Fronteiras (MSF), em seu documento “*Fatal Imbalance*”², a MSF, em 2001, propôs dividir as doenças em Globais, Negligenciadas e Mais Negligenciadas (Souza, 2010).

A maioria dessas doenças fazem parte de um conjunto de patógenos virais transmitidos por artrópodes, também conhecidos como arbovirose. O termo arbovirose é derivado de arbovírus e seus vetores incluem insetos e aracnídeos, como mosquitos, carrapatos e aranhas afirma Figueiredo *et al.* (2017), esses vírus são assim designados porque parte de seu ciclo de replicação ocorre nos insetos, podendo ser transmitidos aos seres humanos e a outros animais pela sem picada de artrópodes hematófagos (Lopes; Nozawa; Linhares, 2014).

Arbovírus urbanos como vírus da Dengue (DENV), vírus da Chikungunya (CHIKV) e o vírus da Zika (ZIKV) são preocupações constantes de saúde pública no Brasil, expandindo-se em condições favorecidas por uma confluência de fatores ecológicos e humanos, tais fatores incluem temperaturas elevadas, alta umidade, vegetação abundante e chuvas significativas, todos agravados pela distribuição desigual dos serviços de tratamento de água (Lee *et al.*, 2021; Santos, 2016).

De acordo com Lee *et al.* (2021), tais condições criam um ambiente ideal para a dispersão generalizada de mosquitos, que são os principais vetores de sua transmissão. As ocorrências iniciais de DENV, CHIKV e ZIKV no Brasil foram documentadas em 1986 pelos pesquisadores Nogueira, Araújo e Schatzmayr (2007), e em 2015 por Simião *et al.* (2019), Lowe *et al.* (2018). Os avanços no conhecimento sobre os arbovírus emergentes e

¹ As grandes doenças negligenciadas.

² Desequilíbrio fatal.

re-emergentes no Brasil destacam questões fundamentais sobre seus vetores, reservatórios, patogênese e diversidade genética, no entanto, permanecem muitas perguntas sem respostas, especialmente sobre os potenciais efeitos sinérgicos da infecção ou co-infecção entre esses vírus ou com outros vírus circulantes (Tajudeen *et al.*, 2022; Marchi *et al.*, 2018). Isso ressalta a necessidade de pesquisas em rede para otimizar a vigilância, o acompanhamento dos doentes e as intervenções de saúde pública diante da atual epidemia causada por esses arbovírus.

Além destes três arbovírus, no Brasil ainda são encontrados os vírus causadores das febres Amarela (YFV), Mayaro (MAYV), Vírus Oropouche (OROV) e o vírus do Oeste do Nilo (WNV) (Li *et al.*, 2021; Silva *et al.*, 2024). A maior preocupação em relação à infecção causada pelo WNV, em relação a sua patogenicidade (Lim *et al.*, 2011). Pereira *et al.* (2020) ressalta que a doença no homem pode ocorrer de forma assintomática ou apresentar sintomas que podem variar desde uma febre passageira a uma encefalite grave, a notificação de casos de infecção por estes vírus é esporádica e concentra-se na região amazônica.

Contudo, acredita-se que uma grande parcela das infecções por estes vírus é subnotificada, devido a sobreposição de sintomas com arboviroses mais comuns no Brasil, por exemplo a DENV, ou de grande foco como o ZIKV e CHIKV. Os vírus de MAYV e OROV mantêm ciclos silváticos onde as transmissões ocorrem através de vetores diferentes do gênero *Aedes*, como o gênero *Haemagogus* no caso do MAYV (Mota *et al.*, 2016). Entretanto, a adaptação destes vírus a vetores de hábitos urbanos, como os do gênero *Aedes* e *Culicoides*, pode ocorrer, e estudos em vigilância genômica e/ou epidemiológica devem manter-se vigilantes a surtos por estas arboviroses, dessa forma, é crucial que esses programas de vigilância estejam atentos não apenas aos surtos das arboviroses mais conhecidas, como DENV e CHIKV, mas também às emergentes, que circulam de forma silenciosa como MAYV e OROV.

No ano de 2024, o Brasil evidenciou casos autóctones da febre do Oropouche, um vírus transmitido pela picada do mosquito do gênero *Culicoides*, popularmente conhecido como "maruim", "muruim" ou "mosquito pólvora", trata-se de um arbovírus do gênero *Orthobunyavirus*, da família *Peribunyaviridae*.

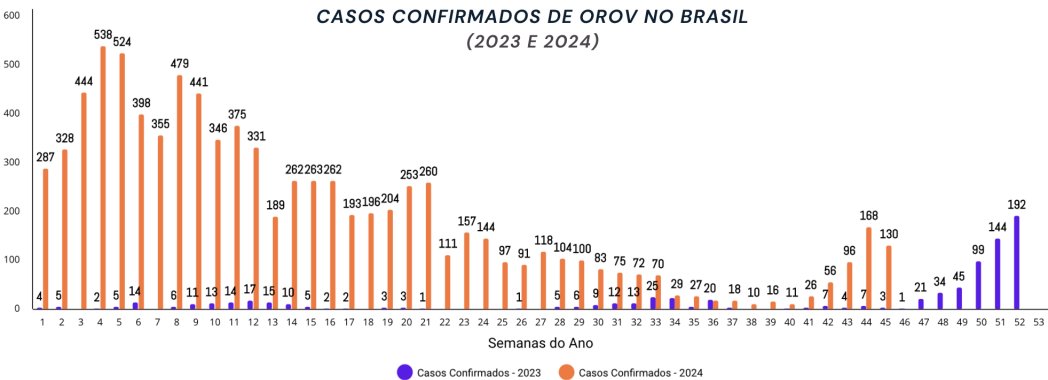
Essa doença febril aguda, conhecida como Oropouche, apresenta sintomas clínicos semelhantes aos da Dengue, incluindo febre, dor de cabeça, dores musculares e articulares, e erupção cutânea (Sakkas *et al.*, 2018). Nos últimos 60 anos, conforme Vasconcelos *et al.* (2009), Rosa *et al.* (2017), mais de 30 epidemias foram registradas, com casos notificados no Brasil, Peru, Panamá, Trinidad e Tobago, o vírus Oropouche tem uma ampla distribuição geográfica, com casos relatados em várias regiões da América do Sul e Central (Pereira *et al.*, 2022).

Embora a doença apresente maior prevalência na região Norte do Brasil, casos

foram identificados em mais de 24 unidades da federação. Até o dia 19 de novembro de 2024, o país registrava um total de 9.126 casos da enfermidade, o que representa um aumento de aproximadamente 998,19% vezes em relação ao número de casos contabilizados no ano anterior, em 2023, que foi de 831 (Ministério da Saúde, 2024a).

A análise comparativa dos dados dos anos de 2023 e 2024, ilustrada na Figura 1, além de revelar um aumento significativo na incidência da doença, evidencia possíveis padrões de disseminação e impacto em diferentes regiões do país.

Figura 1 – Número de casos confirmados para o vírus Oropouche, Brasil, 2023 e 2024.



Fonte: Elaborada pela autora, 2024, com base dos casos confirmados em: Ministério da Saúde (2024a)

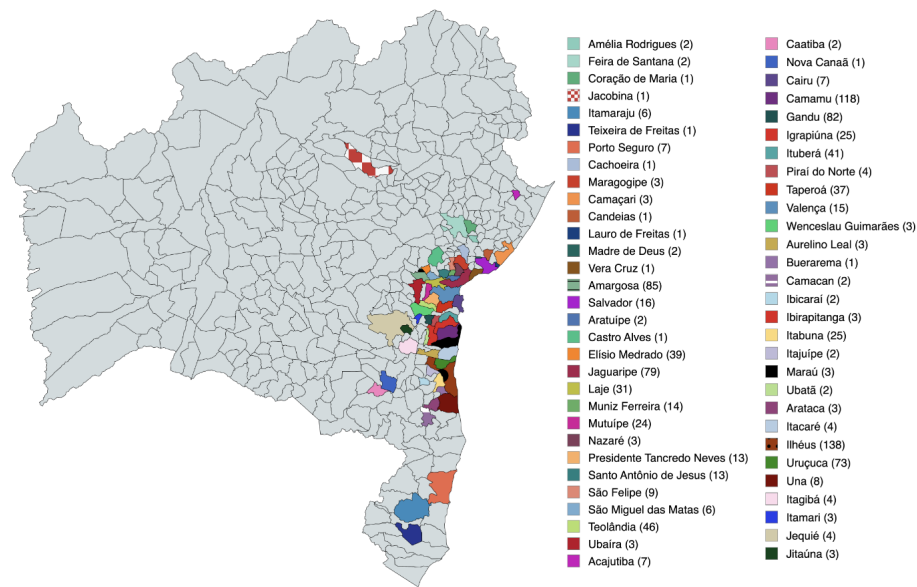
Além disso, foram registrados casos de infecção pelo vírus Oropouche ao longo dos anos (Figueiredo; Figueiredo, 2014), com a presença do vírus sendo identificada em mosquitos coletados em áreas rurais e periurbanas no estado da Bahia, a rápida propagação recente do OROV em diferentes regiões do Brasil levanta preocupações sobre o potencial de complicações clínicas e epidemiológicas que esse patógeno pode causar, com base na afirmação de Naveca *et al.* (2017), a falta de diagnóstico eficaz, vacina e terapia para sua infecção aumenta ainda mais a incerteza sobre o impacto que ele poderá provocar no serviço de saúde pública brasileiro.

Casos esporádicos de OROV foram relatados em áreas não endêmicas do Brasil, curiosamente, o vírus também foi detectado na região metropolitana de Salvador, Bahia, localizada na região Nordeste do Brasil, especificamente em amostras de soro, saliva e urina de pacientes que apresentam sintomas de infecção por arbovírus (Zhang *et al.*, 2024).

No início de abril de 2024, o Brasil registrou mais de 3 mil casos confirmados de Febre Oropouche. A disseminação da doença avançou para outras regiões do país, ultrapassando os limites da Região Norte (Digital, 2024). Além disso, foram confirmados na Bahia dois óbitos relacionados à doença, constituindo as primeiras mortes registradas mundialmente, conforme destacado por um estudo realizado pelos pesquisadores Bandeira *et al.* (2024).

O primeiro caso confirmado na Bahia ocorreu em 10 de abril de 2024. A Figura 2 apresenta uma visão das cidades baianas com registros da enfermidade, conforme reportagem da Metro1 (2024). Até setembro do mesmo ano, a Bahia contabilizou notificações em mais de 27 municípios, totalizando 1.041 casos (G1 Bahia, 2024).

Figura 2 – Cidades na Bahia com casos confirmados da doença.



Fonte: Elaborada pela autora, 2024, com base no número de casos confirmados disponível em: G1 Bahia (2024)

1.1 Justificativa

Com o crescente volume de dados gerados por infecções causadas por esses vírus, e as pesquisas realizadas sobre o genoma humano, de outras espécies e dos patógenos, acentua-se, ainda mais, a necessidade do uso de tecnologias da informação para auxiliar no processamento deste volume de dados, nesse sentido pesquisas bioinformáticas processam, em geral, um volume considerável de informações que devem ser analisadas em conjunto para proporcionar inferências precisas sobre possíveis hipóteses e possibilitar, assim, a construção de novas teses sobre o papel ou funcionamento biológico de determinada proteína ou organismo (Baxevanis; Bader; Wishart, 2020).

A bioinformática é uma ciência que utiliza soluções tecnológicas a fim de lidar com informações biológicas, desenvolvendo abordagens que incluem coleta de dados biológicos, busca e construção de bancos de dados, análise e construção de software, entre outras, ela desempenha um papel fundamental na era da genômica e da biologia computacional. Segundo Lengauer (2007), a bioinformática pode ser definida como "a aplicação de métodos computacionais para entender e organizar os dados biológicos", além

disso, a sua importância se destaca na análise de sequências genômicas, na modelagem de estruturas de proteínas e na compreensão de sistemas biológicos complexos (Bult, 1998).

Degrave e Machado (2007) afirmam que os conjuntos de dados sempre crescentes, novas fronteiras na ciência, o desenvolvimento de novas abordagens para análise e apresentação de dados, e a investigação de novas e complexas perguntas são as forças motrizes da bioinformática, as pesquisas nos campos genômicas, proteômicas, transcriptômicas, metabolômicas e fenômicas tiveram um aumento explosivo no volume de dados, tornando informações essenciais sobre a biologia molecular em conjuntos de dados em grande escala.

Essa revolução no domínio da biologia molecular resulta na necessidade de ferramentas e métodos robustos para gerenciar, visualizar, integrar, analisar e modelar esses dados (Schadt *et al.*, 2010), com o aprimoramento da tecnologia de sequenciamento e a introdução de técnicas de alta resolução, os pesquisadores têm agora acesso a conjuntos de dados complexos e multidimensionais, exigindo abordagens avançadas para a extração de informações significativas (Hasin; Seldin; Lusi, 2017).

O desenvolvimento de abordagens que incluem coleta de dados biológicos, busca e construção de bancos de dados, análise e construção de software, ela desempenha um papel fundamental na era da genômica e da biologia computacional, Bult (1998) destaca a importância da bioinformática na análise de sequências genômicas, na modelagem de estruturas de proteínas e na compreensão de sistemas biológicos complexos.

Na vasta gama de possibilidades que a bioinformática oferece, destaca-se uma esfera específica dedicada ao estudo dos vírus negligenciados. Doenças negligenciadas, como malária, doença de Chagas, Dengue e Oropouche, continuam sendo uma das principais razões por trás da morbidade e mortalidade em escala global. Ao aplicar técnicas bioinformáticas, como o sequenciamento de genomas virais e modelagem computacional de interações entre hospedeiro e patógeno, os pesquisadores podem identificar alvos terapêuticos promissores e projetar tratamentos mais eficazes contra essas doenças (Hotez, 2021; Molyneux; Savioli; Engels, 2017).

Podemos entender que esta área é uma ciência que visa sempre o desenvolvimento de ferramentas mais úteis e amigáveis, um exemplo a ser citado é a IBM, que desenvolveu o *Watson for Oncology*³, uma plataforma que analisa as informações médicas dos pacientes e auxilia os médicos com opções de tratamento (International Business Machines Corporation, 2024). A integração interdisciplinar proporcionada pela bioinformática é crucial para enfrentar os desafios impostos pelos arbovírus emergentes e negligenciados, tanto no Brasil quanto em escala global, ao fortalecer a capacidade de resposta às epidemias virais e aprimorar a saúde pública, as investigações nesse campo não apenas preenchem lacunas críticas de conhecimento, mas também colaboram para o desenvolvimento de políticas de

³ Watson para oncologia.

saúde mais eficientes e embasadas em evidências.

Portanto, torna-se imprescindível continuar investindo em pesquisas no campo da bioinformática, visando enfrentar os desafios complexos impostos pelos arbovírus no Brasil e promover uma abordagem proativa e baseada em evidências para o fortalecimento da saúde pública em escala global.

Diante do elevado número de casos de infecções causadas por arbovírus, algumas questões precisam ser avaliadas. Existem condições ambientais favoráveis para a propagação do Oropouche, como temperaturas elevadas e alta umidade, que poderiam torná-lo um problema mundial? A distribuição desigual dos serviços de tratamento de água pode contribuir para a disseminação do Oropouche em populações menos favorecidas? A bioinformática pode desempenhar um papel crucial na identificação e análise de dados genômicos, como a caracterização genômica, auxiliando, assim, no desenvolvimento de estratégias de prevenção e controle do Oropouche?

Diante dos problemas elencados nesta pesquisa, os dois primeiros citados envolvem questões amplas, principalmente relacionadas a políticas públicas para reduzir as emissões de gases que contribuem para o aquecimento global e ao saneamento básico em populações menos favorecidas. No entanto, o último problema se encaixa nos propósitos desta pesquisa, cujo objetivo central é apoiar as investigações relacionadas a esse arbovírus. Este trabalho propõe o desenvolvimento de uma ferramenta de bioinformática de livre acesso para a caracterização genômica do vírus Oropouche, utilizando sequências públicas disponíveis em bancos de dados primários.

Além desse recurso tecnológico, outras estratégias inovadoras na bioinformática têm sido empregadas para enfrentar desafios virais. Por exemplo, a utilização de algoritmos de aprendizado de máquina e inteligência artificial (IA) tem permitido a identificação de padrões complexos em grandes conjuntos de dados genômicos, facilitando a detecção precoce de mutações virais relevantes para a resistência a medicamentos e o desenvolvimento de vacinas mais eficazes (Montesinos-López *et al.*, 2021; Alharbi; Rashid, 2022), da mesma forma, técnicas avançadas de reconstrução filogenética têm sido empregadas para entender novos meios de transmissão e dispersão viral (Giovanetti *et al.*, 2023).

Nesse sentido, as ferramentas filogenéticas são recursos de extrema importância na bioinformática para estudar a evolução viral, traçar a origem de epidemias, estabelecer o modo de transmissão, pesquisar a ocorrência de resistência a medicamentos ou determinar a origem do vírus nos diferentes compartimentos corporais conforme os autores, Fonseca *et al.* (2019a), Juma *et al.* (2022a). O uso de técnicas de bioinformática é fundamental para acompanhar a evolução da diversidade viral, dando suporte aos estudos de análise de sequências genômicas e sendo crucial para a vigilância o polimorfismo viral, o desenvolvimento de novas estratégias terapêuticas, o desenvolvimento de produtos vacinais e a escolha adequada desses produtos.

Desta forma, mais uma vez, torna-se evidente a importância do avanço contínuo na pesquisa de bioinformática para enfrentar os desafios representados pelos arbovírus emergentes e negligenciados. A interdisciplinaridade, já citada, que a bioinformática proporciona possibilita não apenas a compreensão mais profunda da epidemiologia e patogênese desses vírus, mas também o desenvolvimento de ferramentas e estratégias eficazes de vigilância, diagnóstico, prevenção e controle, com isso, fortalece as capacidades de resposta a epidemias virais, contribuindo para a saúde pública global (Ferrucci *et al.*, 2010).

1.2 Objetivos

1.2.1 *Objetivo Geral*

Desenvolver uma ferramenta de ambiente web, de fácil acesso e amigável, que permita a classificação genômica do vírus Oropouche.

1.2.2 *Objetivos Específicos*

- a) Avaliar e compreender técnicas de identificação de marcadores genéticos em sequências de cepas virais do vírus do Oropouche;
- b) Desenvolver um *pipeline* automatizado para a classificação de linhagens em sequências genéticas do Oropouche, permitindo uma análise precisa e rápida das diferentes cepas;
- c) Criar uma interface intuitiva para a classificação genética, baseada em análises dos grupos monofiléticos do Oropouche, facilitando a visualização e interpretação dos resultados por pesquisadores e profissionais da área.

Esta monografia aborda o desenvolvimento de uma ferramenta de bioinformática voltada para a caracterização genômica do vírus Oropouche, um arbovírus negligenciado de crescente relevância para a saúde pública. Inicialmente, são apresentados os fundamentos teóricos que sustentam a pesquisa, com uma análise detalhada da estrutura do vírus, seus métodos de transmissão e replicação, e as condições ambientais que favorecem sua propagação, destaca-se o papel crucial da bioinformática no estudo de doenças virais, evidenciando como essa ciência contribui para avanços na vigilância epidemiológica.

Em seguida, é detalhada a metodologia utilizada para conceber a ferramenta, desde a coleta e organização dos dados genômicos em banco de dados públicos até o desenvolvimento de um pipeline automatizado para genotipagem e a criação de uma interface web amigável. Os resultados obtidos demonstram a eficiência da aplicação, que possibilita análises precisas e rápidas das linhagens genéticas do vírus, oferecendo suporte tanto para pesquisadores quanto para profissionais de saúde.

Por fim, o trabalho discute com ênfase as contribuições da pesquisa para a saúde pública, suas limitações e as possibilidades de aprimoramento futuro da ferramenta, reforçando a importância da interdisciplinaridade para enfrentar os desafios impostos por arbovírus negligenciados.

2 MAPEAMENTO DO VÍRUS OROPOUCHE: EXPLORANDO UM PATÓGENO NEGLIGENCIADO

2.1 Estrutura do vírus

Os vírus são pequenos agentes infecciosos (Meneguetti, 2024). Não há consenso na comunidade científica sobre serem considerados seres vivos, pois não possuem estrutura celular afirma Microbiologando (2024), dependem de uma célula viva para se replicarem, sendo assim conhecidos como parasitas intracelulares obrigatórios (Educação, 2024).

A estrutura essencial de um vírus é composta por dois componentes primordiais: um ácido nucléico específico e um envoltório proteico conhecido como capsídeo. A união desses ácidos nucléicos ao capsídeo é denominada nucleocapsídeo, o envelope viral é derivado da membrana celular do hospedeiro durante a liberação do vírus ao término do ciclo de replicação, resultando em vírus desnudos, sem envelope, e envelopados (Meneguetti, 2024), conforme ilustrado na Figura 3.

Existe também uma distinção quando nos referimos aos vírus no ambiente externo (fora do hospedeiro), momento em que são chamados de vírions (Meneguetti, 2024).

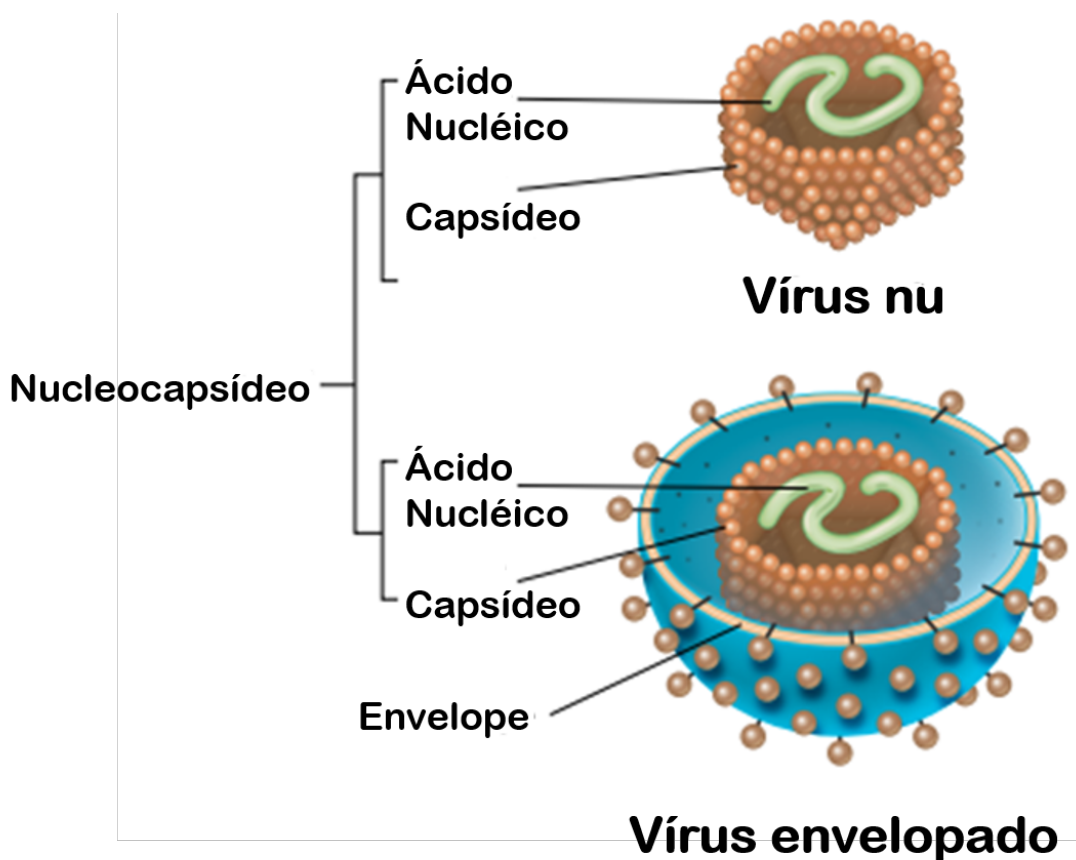
Os vírus são classificados com base no tipo de ácido nucleico que possuem, na estrutura do capsídeo e nos organismos que têm capacidade de infectar (Toda Matéria, 2024):

- a) Adenovírus: formado por ácido desoxirribonucleico (DNA);
- b) Retrovírus: formados por ácido ribonucleico (RNA);
- c) Arbovírus: transmitidos por insetos;
- d) Bacteriófagos: vírus que infectam bactérias;
- e) Micófitos: vírus que infectam fungos.

Sensi Saúde (2024), relata que a transmissão de vírus pode ocorrer de diversas maneiras, alguns vírus são transmitidos pelo ar, através de gotículas respiratórias expelidas por indivíduos infectados ao tossir ou espirrar, outros podem ser transmitidos pelo contato direto com fluidos corporais, como sangue, saliva, urina e secreções genitais. Certos vírus podem ser disseminados por meio de alimentos contaminados, picadas de insetos ou contato com superfícies contaminadas.

Aproximadamente 50% de todos os resfriados são causados por um dos 100 sorotipos do rinovírus (Toda Matéria, 2024). Em 2018, de acordo com o Tribunal de Justiça do Distrito Federal e Territórios (2024), até 30 de maio, foram registrados 4.704 casos de influenza de todos os tipos no Brasil, deste total, 3.978 foram influenza A (H1N1),

Figura 3 – Estrutura do vírus nus e envelopados.



Fonte: Madigan *et al.* (2019, p. 3).

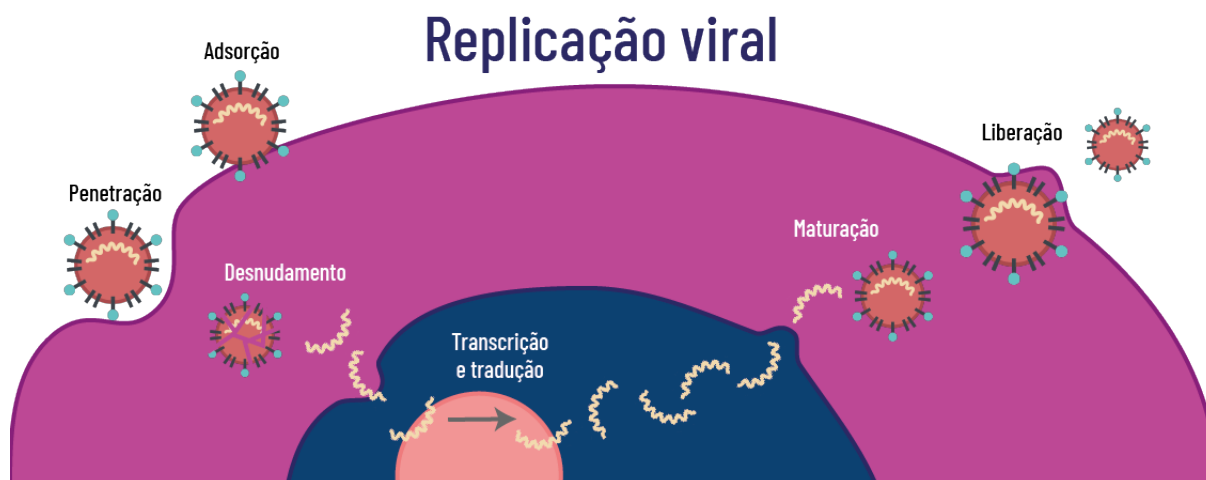
resultando em 764 óbitos. As partículas de influenza são envelopadas de RNA de fita simples segmentada e subdividem-se nos tipos A, B e C (Forleo-Neto *et al.*, 2003).

Por não possuírem células próprias, os vírus precisam invadir uma célula hospedeira viva para se reproduzir, em um processo chamado infecção viral (National Geographic Brasil, 2023). De acordo com Manuals (2023), após a infecção, o vírus adere a uma célula, chamada célula hospedeira, e libera seu material genético.

Os vírus são capazes de infectar todos os tipos de organismos vivos, a estratégia fundamental de um vírus pode ser compreendida em seis estágios principais, ilustrados na Figura 4. Lopes (2014), fala que, embora cada vírus siga uma abordagem distinta em cada um desses estágios, o princípio infeccioso geral permanece consistente: utilizar a célula hospedeira para produzir mais partículas virais.

Na primeira etapa do ciclo de replicação, as partículas virais se ligam à célula hospedeira por meio de proteínas localizadas em seus capsídeos, Santos (2024) alega que essas proteínas interagem com proteínas da membrana plasmática da célula-alvo, facilitando a adesão.

Figura 4 – Ciclo de replicação do vírus.



Fonte: Abdalla (2020).

Durante a penetração, a partícula viral insere seu material genético no interior da célula. Diferentes vírus utilizam diferentes métodos para alcançar esse objetivo. Nos vírus envelopados, toda a partícula viral pode ser englobada pela célula por endocitose, ou o envelope viral pode se fundir com a membrana plasmática da célula hospedeira (Lopes, 2014).

A informação trazida por Santos (2024), releva que a remoção do capsídeo é o processo onde o ácido nucleico do vírus é liberado, geralmente devido à fragmentação do capsídeo por enzimas da célula hospedeira, nem todos os vírus passam por essa etapa, pois nem todos penetram na célula com seus capsídeos, Mariz (2020) complementa dizendo que, durante a replicação genética, ocorre a formação de proteínas virais.

Na montagem de partículas virais, novos vírus são formados pela união do ácido nucleico com os polipeptídeos que compõem o capsídeo. Na etapa final da multiplicação, as partículas virais são liberadas da célula hospedeira, e essa liberação pode ocorrer pela lise da célula infectada ou por brotamento, sendo este último típico de vírus envelopados. Tanto a lise quanto o brotamento podem resultar na morte da célula hospedeira (Santos, 2024).

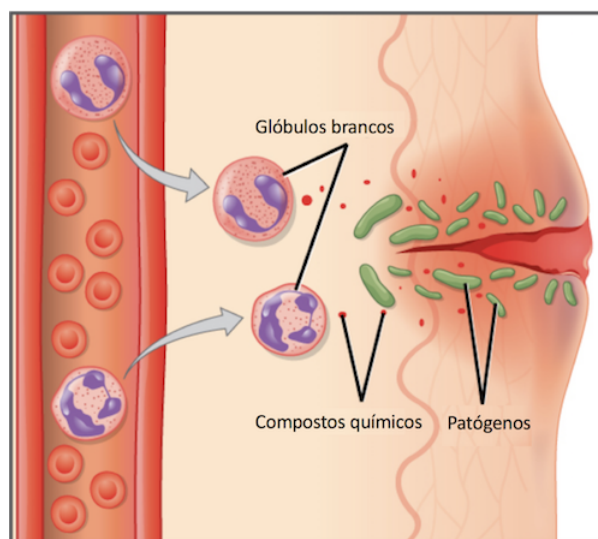
O sistema imunológico é constituído por uma intrincada rede de órgãos, células e moléculas, e tem por finalidade manter a homeostase do organismo, combatendo agressões em geral (Cruvinel *et al.*, 2010). Em seu estudo, Oceandrop (2024) fala que resposta imunológica visa manter a saúde do organismo, garantindo a sobrevivência e combatendo qualquer ameaça. Assim como os sistemas renal, cardíaco e hepático, existe o sistema imune.

A imunidade inata atua em conjunto com a imunidade adaptativa e caracteriza-se

pela rápida resposta à agressão, independentemente de estímulo prévio, sendo a primeira linha de defesa do organismo conforme Cruvinel *et al.* (2010), Benegrip (2024) complementa que a imunidade inata é aquela que nasce conosco, oferecendo uma defesa imediata para frear o avanço de invasores no corpo, mesmo que de maneira genérica.

O sistema imunológico inato é composto por barreiras físicas, químicas e biológicas, células especializadas e moléculas solúveis, presentes em todos os indivíduos, independentemente de contato prévio com agentes agressores. As barreiras físicas, como a pele, a saliva, o muco, as lágrimas e os pelos, formam a primeira linha de defesa do organismo, na Figura 5, é possível ver o esquema mostrando os glóbulos brancos liberando substâncias que desencadeiam o processo inflamatório no local infectado, esses fluidos contêm enzimas que destroem as membranas dos patógenos. Se o patógeno ultrapassar essas barreiras, a segunda linha de defesa, a biológica, é ativada. A presença do patógeno desencadeia uma resposta inflamatória, aumentando o fluxo sanguíneo e direcionando glóbulos brancos para a área infectada. Esses glóbulos brancos realizam fagocitose e liberam substâncias inflamatórias, causando vermelhidão, suor e dor. A destruição dos patógenos pode levar à produção de compostos químicos que elevam a temperatura corporal, resultando em febre, o que ajuda a inibir a proliferação dos patógenos (Khan Academy, 2018).

Figura 5 – Glóbulos brancos liberando substâncias que desencadeiam o processo inflamatório no local infectado.



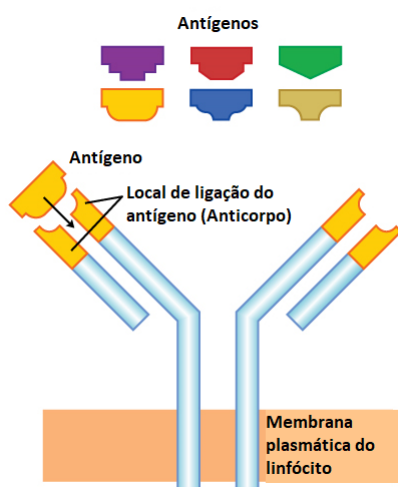
Fonte: OpenStax (2021).

A imunidade adaptativa precisa ser desenvolvida com o tempo, mas promove uma resposta mais específica e eficaz, DASA (2024) confirma que esta imunidade também pode se formar após a imunização, tornando as vacinas extremamente importantes para o fortalecimento do sistema imunológico.

Quando os patógenos conseguem superar as defesas do sistema imunológico inato, o sistema imunológico adaptativo é acionado. As principais características da resposta imune adquirida incluem especificidade e diversidade de reconhecimento, memória imunológica, especialização de resposta, autolimitação e tolerância aos componentes do próprio organismo, diferentemente da resposta inata, a resposta imune adaptativa depende da ativação de leucócitos especializados e das substâncias por eles produzidas, nesse caso a resposta imune específica é desencadeada por antígenos, que estão presentes na superfície dos patógenos e são exclusivos de cada um deles (Cruvinel *et al.*, 2010).

Assim, a resposta imune específica é ativada pelos antígenos, normalmente encontrados na superfície dos patógenos e exclusivos de cada um. Conforme ilustrado na Figura 6, o sistema imunológico responde à presença dos antígenos produzindo células que atacam diretamente o patógeno ou produzindo proteínas específicas chamadas anticorpos. Esses anticorpos se ligam aos antígenos e levam o patógeno aos locais onde serão destruídos (Khan Academy, 2017).

Figura 6 – Ligação do linfócito ao antígeno do patógeno.



Fonte: Khan Academy (2017).

O sistema imunológico possui a capacidade de memorizar cada antígeno com o qual entra em contato. Após a exposição inicial, alguns linfócitos se diferenciam em células de memória, que podem perdurar por anos ou até mesmo décadas. Quando essas células de memória encontram o mesmo antígeno novamente, o reconhecimento ocorre de forma imediata, resultando em uma resposta rápida, vigorosa e específica contra aquele antígeno. Esta resposta imunológica específica é a razão pela qual os indivíduos não contraem varicela ou sarampo mais de uma vez e também explica a eficácia das vacinas na prevenção de determinadas doenças. Esses linfócitos podem ser células T ou células B (MSD Manual Consumer Version, 2017).

As células T, que mediam a defesa contra estes organismos, incluem diversos tipos de linfócitos T auxiliares CD4+. Um papel importante dessas células é ajudar os linfócitos B a produzir anticorpos, como parte da resposta imunológica humoral, alguns subconjuntos de células T auxiliares promovem respostas inflamatórias ricas em leucócitos ativados, sendo particularmente eficientes na eliminação de microrganismos extracelulares, confirma SanarMed (2023).

O vírus Oropouche é um vírus envelopado com capsídeo helicoidal, pertencente ao domínio *Riboviria*, reino *Orthornavirae*, filo *Negarnaviricota*, subfilo *Polyploviricotina*, classe *Ellioviricetes*, ordem *Bunyavirales*, família *Peribunyaviridae*, gênero *Orthobunyavirus* e subgrupo sorológico *Simbu*. Este vírus apresenta um genoma composto por três segmentos de RNA de cadeia simples com polaridade negativa: L, M e S (Murphy *et al.*, 2012), conforme ilustrado na Figura 7.

Como todos os *bunyavírus*, o genoma do OROV é composto por três segmentos de RNA de fita simples e polaridade negativa, cada um nomeado de acordo com seu tamanho: o segmento L (*large*)¹ possui 6852 nucleotídeos, o segmento M (*medium*)² 4385 nucleotídeos e o segmento S (*small*)³ 958 nucleotídeos (Acrani *et al.*, 2015).

Simião *et al.* (2019) ressalta que, embora não existam estudos específicos de microscopia eletrônica em tecidos humanos para o OROV, acredita-se que suas partículas virais tenham morfologia similar à observada para outros *ortobunyavírus*, como demonstrado em estudos de microscopia eletrônica para o VLAC.

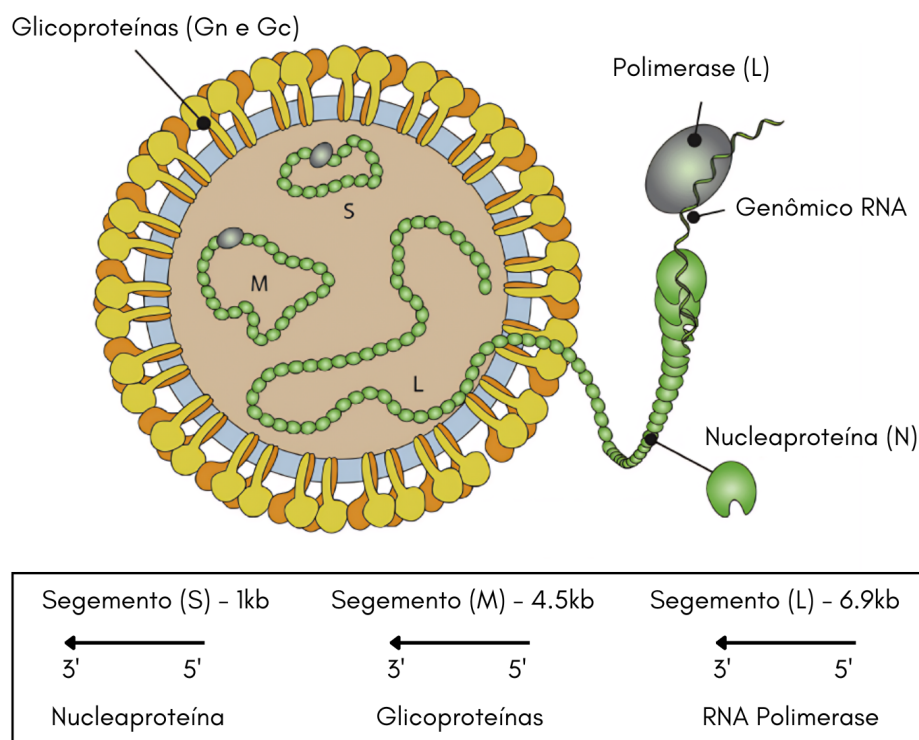
As partículas desse vírus são esféricas, medindo entre 80 a 110 nm de diâmetro, e são envolvidas por um envelope lipídico, conforme a Figura 8. O envelope viral deriva das membranas do complexo de Golgi e, ocasionalmente, da membrana plasmática celular. Internamente, a partícula viral contém três segmentos de RNA de diferentes tamanhos, cada um ligado a uma proteína L (possivelmente RNA polimerase viral dependente de RNA) e circundado por uma proteína denominada N (nucleocapsídio), formando três nucleoproteínas (Simião *et al.*, 2019).

¹ Grande.

² Médio.

³ Pequeno.

Figura 7 – Estrutura do vírus Oropouche.



Fonte: Romero-Alvarez e Escobar (2018, p. 3).

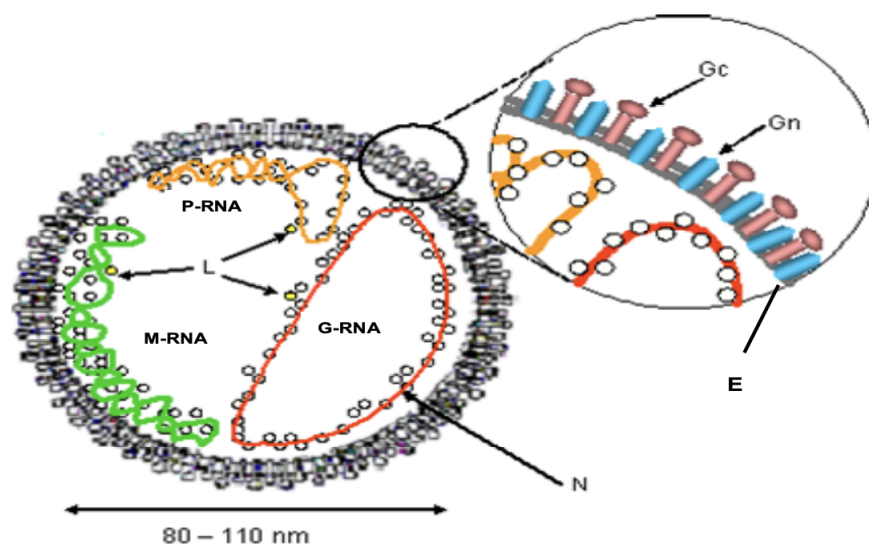
2.2 Métodos de infecção e replicação

Os estudos de Romero-Alvarez e Escobar (2018) esclarecem que, ciclo silvestre do OROV ocorre entre hospedeiros vertebrados como bichos-preguiça, primatas não-humanos e possivelmente aves silvestres, sendo transmitido por diversos vetores. A infecção humana ocorre em regiões periflorestais e, ao retornar ao meio urbano com viremia, o indivíduo pode servir como fonte de vírus para a infecção de artrópodes, principalmente o vetor urbano *Culicoides paraensis*.

Alguns arbovírus adotam um mecanismo alternativo de multiplicação, mantendo-se nos artrópodes através da transmissão viral pelos ovos (transmissão transovariana) e durante os estágios imaturos. Posteriormente, os artrópodes podem transmitir esses vírus para humanos ou outros vertebrados durante seu primeiro repasto sanguíneo, permanecendo infectados ao longo de suas vidas (Shope, 1994).

No ciclo urbano ou epidêmico, demonstrados na Figura 9, o vírus é transmitido de pessoa a pessoa pela picada do inseto antropofílico *Culicoides paraensis*, conhecido popularmente na região Amazônica brasileira como "maruim". Estudos experimentais demonstraram a capacidade do *Culicoides paraensis* de transmitir o OROV para hamsters após se alimentarem do sangue de pacientes virêmicos, com um intervalo de exposição de

Figura 8 – Representação esquemática da partícula de *Orthobunyavirus* evidenciando os três segmentos de RNA (P-RNA, MRNA e GRNA) associados à proteína L. Circundando cada molécula de RNA, observa-se a proteína N, e no detalhe, as duas glicoproteínas Gn e Gc emergindo do envelope (E).



Fonte: Murphy *et al.* (2012).

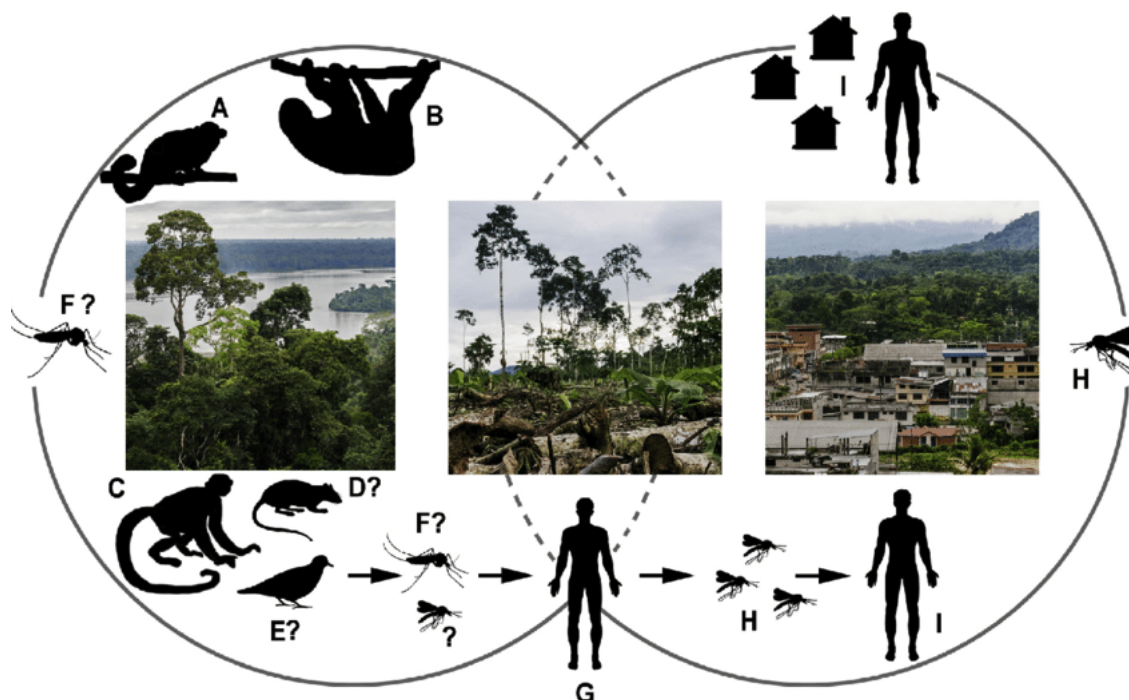
cinco dias ou mais (Haaster *et al.*, 1982).

Exceto pelos DENV, ZIKV, CHIKV e OROV, que possuem ciclos de transmissão urbana, todos os outros arbovírus associados a doenças humanas na Amazônia brasileira são mantidos na natureza por meio de ciclos silvestres. Nestes ciclos, várias espécies de insetos hematófagos atuam como vetores e vertebrados silvestres como hospedeiros (Rosa; Vasconcelos; Rosa, 1998).

Os maruins possuem hábitos diurnos, são mais ativos no final da tarde e têm uma forte preferência por sangue humano. São encontrados em grande densidade durante períodos epidêmicos e se reproduzem principalmente em matéria orgânica em decomposição, como troncos de bananeiras, cascas de cacau e cupuaçu, e detritos acumulados em depressões de árvores, sendo comuns em áreas tropicais e subtropicais das Américas (Hoch; Roberts; Pinheiro, 1986; Linley; Hoch; Pinheiro, 1983).

Romero-Alvarez e Escobar (2018) descrevem o ciclo de vida do OROV em sete fases, conforme ilustrado na Figura 9, que abrange as fases de A a D na interface silvestre e de E a G na fase urbana. As características de cada fase são detalhadas a seguir:

- a) o ciclo silvestre do vírus Oropouche envolve mamíferos selvagens como *Callithrix penicillata* (Sagui de tufos pretos), *Bradypus tridactylus* (Preguiça-de-bentinho), *Sapajus alloata* (Macaco-prego), *Allouatta caraya* (Bugio-preto) e o gênero *Proechimys*, um roedor da família *Echimyidae*;



Pinheiro, Rosa e Rosa (1981), Nunes *et al.* (2005) retratam que estudos com animais domésticos durante surtos excluíram seu papel como amplificadores do vírus, sugerindo que o homem seja o único vertebrado envolvido no ciclo urbano da virose. No ciclo silvestre, há evidências de que primatas não-humanos e possivelmente algumas espécies de aves

silvestres atuam como hospedeiros vertebrados (Nunes *et al.*, 2005; LeDuc *et al.*, 1981), embora o vetor ainda não seja completamente identificado, sendo necessário investigar a possível participação de *Culicoides paraensis* no ciclo silvestre do vírus (Cherry *et al.*, 2013).

Rosa, Vasconcelos e Rosa (1998) evidencia que o vírus Oropouche, juntamente com os DENV, está associado a doenças epidêmicas em áreas urbanas, enquanto ROCV, MAYV e YFV predominam em áreas rurais. O vínculo entre os dois ciclos é provavelmente o homem. Ao adquirir a infecção em áreas silvestres enzoóticas e retornar ao meio urbano durante a fase virêmica, ele serve como fonte de infecção para outros maruins, estabelecendo uma cadeia de transmissão que pode resultar em epidemias.

No Brasil, especialmente na Amazônia, a alta densidade de diversas espécies de dípteros hematófagos (mosquitos, flebotomíneos, carrapatos, maruins) e vertebrados silvestres cria condições ambientais altamente favoráveis para a manutenção de vírus, especialmente os arbovírus, na natureza retrata Leão (1997).

A doença febril aguda conhecida como Oropouche apresenta sintomas clínicos semelhantes aos da Dengue (Sakkas *et al.*, 2018), clinicamente, manifesta-se como um quadro febril agudo, frequentemente acompanhado por dores de cabeça, mialgias, artralgias, anorexia, tontura, congestão conjuntival, calafrios e fotofobia. Alguns pacientes podem apresentar exantema morbiliforme semelhante ao da rubéola, além de náuseas, vômitos, diarreia, dor epigástrica e dor retroocular. Manifestações sistêmicas adicionais podem ocorrer (Pinheiro; Rosa; Rosa, 1981).

Após o término do episódio febril inicial, é comum a recorrência dos sintomas, geralmente com menor intensidade, alguns pacientes podem desenvolver meningite asséptica (Pinheiro *et al.*, 1982). A recuperação dos pacientes é geralmente completa, sem sequelas aparentes, mesmo nos casos mais graves, e não há relatos de mortalidade diretamente atribuída à febre do Oropouche (Pinheiro *et al.*, 1982).

2.3 Gênero da família do Oropouche

Os arbovírus, abreviação de "*arthropod-borne viruses*"⁴, compreendem uma diversidade de vírus que são transmitidos pela picada de artrópodes hematófagos, como mosquitos culicídeos, carrapatos e flebotomíneos, entre hospedeiros vertebrados infectados virêmicos e outros vertebrados susceptíveis, incluindo o ser humano (Santos; Romanos; Wigg, 2015).

O termo "*arthropod-borne virus*" foi introduzido em 1942 para descrever um grupo de vírus animais que se propagavam em artrópodes e eram transmitidos biologicamente a hospedeiros vertebrados. Duas décadas depois, o Subcomitê Internacional para Nomenclatura Viral recomendou oficialmente a adoção do termo "arbovírus" para designar os vírus

⁴ Vírus transmitidos por artrópodes.

mantidos na natureza em ciclos envolvendo vetores artrópodes hematófagos e hospedeiros vertebrados (Karabatsos, 1985).

Entre as arboviroses que afetam os seres humanos, a maioria pertence às famílias *Flaviviridae*, *Togaviridae*, *Peribunyaviridae*, *Reoviridae* e *Rhabdoviridae*. Destas, destacam-se as famílias *Flaviviridae*, *Togaviridae* e *Peribunyaviridae* por sua importância em saúde pública no Brasil (Fundação Oswaldo Cruz, 2024).

A família *Reoviridae* inclui 11 gêneros, entre eles o gênero *Orbivirus*, que abrange arbovírus principalmente transmitidos por carrapatos, culicídeos e mosquitos. Os três orbivírus economicamente mais importantes são o vírus da língua azul, o vírus da doença equina africana e o vírus da doença hemorrágica epizootica, todos transmitidos por espécies de Culicídeos, causando doenças em equinos e ruminantes domésticos e selvagens (Fundação Oswaldo Cruz, 2024).

A maioria das doenças causadas por arbovírus, com exceção da Dengue, teve origem zoonótica, ou seja, os vírus circulavam naturalmente entre animais vertebrados silvestres e, devido a mudanças ambientais e pressões seletivas que levaram a mudanças genéticas no patógeno, começaram a se adaptar e infectar populações humanas (Lopes; Nozawa; Linhares, 2014). Os arbovírus possuem critérios específicos para serem classificados como tais:

- a) infectar vertebrados e invertebrados;
- b) iniciar uma viremia, ou seja, a presença do vírus no sangue circulante de um ser vivo;
- c) manter títulos virais suficientemente altos em um hospedeiro vertebrado por tempo suficiente para permitir a infecção do vetor invertebrado;
- d) iniciar uma infecção produtiva e persistente na glândula salivar do invertebrado, possibilitando a transmissão do vírus a outros hospedeiros vertebrados.

Murray, Rosenthal e Pfaller (2006) afirma que os arbovírus têm uma ampla gama de hospedeiros, incluindo vertebrados (mamíferos, aves, anfíbios, répteis) e invertebrados (mosquitos, carrapatos). Sua distribuição geográfica é ampla, abrangendo todos os continentes, com predominância em regiões tropicais devido às condições ecológicas favoráveis, nos trópicos, os vetores e hospedeiros vertebrados coexistem durante todo o ano, enquanto em climas temperados o ciclo de transmissão é interrompido durante o inverno, reiniciando-se na primavera ou verão (Leão, 1997).

O desequilíbrio atual dos ecossistemas pode levar ao surgimento de um maior número de doenças relacionadas ao manejo inadequado dos ecossistemas naturais, contribuindo para o aparecimento de diversos arbovírus. Alguns desses vírus são responsáveis por importantes problemas de saúde pública em níveis regional e nacional (Vasconcelos *et al.*,

2001).

Os arbovírus têm sido extensivamente estudados quanto às suas propriedades físico-químicas, possuem genoma constituído por RNA e podem ter genomas segmentados ou não, com uma ou duas fitas nucleotídicas (Beaty; Trent; Roehrig, 2020), os arbovírus, na sua maioria, são vírus com genoma RNA e grande plasticidade genética e alta frequência de mutações, o que permite adaptações a hospedeiros vertebrados e invertebrados (Ministério da Saúde, 2024b). A maioria dos arbovírus conhecidos está classificada em seis famílias distintas: *Rhabdoviridae*, *Bunyaviridae*, *Togaviridae*, *Flaviviridae*, *Reoviridae* (Pinheiro *et al.*, 1996) e *Asfaviridae* é o único vírus de DNA conhecido transmitido por artrópodes.

No Brasil, pelo menos 210 tipos de arbovírus foram isolados, com 196 deles inicialmente identificados na Amazônia brasileira. Muitos desses vírus nunca foram encontrados fora dessa região (Pinheiro *et al.*, 1996; Vasconcelos *et al.*, 1998), Leão (1997), Vasconcelos *et al.* (2001) complementa que na Amazônia, pelo menos 110 espécies são consideradas novas para a ciência, e 34 delas estão associadas a infecções humanas.

A Pan-Amazônia é reconhecida como o maior reservatório de arbovírus do mundo, com a região Amazônica brasileira destacando-se pela maior variedade de arbovírus isolados globalmente, epidemias em áreas urbanas e rurais, especialmente causadas pelos vírus DENV, YFV, OROV, ROCV e MAYV, representam um significativo risco à saúde de uma parcela expressiva da população (Leão, 1997).

Os métodos sorológicos específicos para diagnóstico se baseiam principalmente em técnicas como inibição de hemaglutinação, fixação do complemento e soroneutralização em camundongos (Schmidt, 1995; Paula; Fonseca, 2004). O teste de inibição de hemaglutinação em microplacas é especialmente recomendado para sorologia de rotina, sendo sensível, fácil de executar e requerendo equipamentos simples, utilizando a capacidade dos anticorpos contra arbovírus específicos para inibir a hemaglutinação de hemácias de ganso (Clarke; Casals, 1958).

Outros métodos incluem o ensaio imunoenzimático (ELISA) para detecção de antígenos virais (Silva-Nunes *et al.*, 2006) e métodos de biologia molecular, como a reação em cadeia da polimerase com transcriptase reversa (RT-PCR), seguida de sequenciamento nucleotídico, além da microscopia eletrônica (Samuel; Tyagi, 2006).

Os arbovírus são agentes significativos de enfermidades tanto em animais quanto em seres humanos, especialmente na Amazônia brasileira, onde algumas doenças ainda não foram determinadas em animais domésticos, compreender a relação dessas viroses é crucial, especialmente em casos de doenças encefálicas e abortos em herbívoros domésticos, cuja etiologia pode estar relacionada a arbovírus.

2.4 Atuação da Bioinformática

No século atual, vivenciamos a emergência da era genômica e pós-genômica, onde a Bioinformática se destaca como uma ferramenta essencial. Através das plataformas online disponíveis, é possível adquirir, armazenar, classificar e interpretar os volumosos dados gerados pelo sequenciamento de DNA e proteínas, caracterizando as ômicas, como a genômica e a proteômica (Junior; Oliveira; Ceccatto, 2012).

Uma das áreas primordiais da biologia computacional envolve o desenvolvimento de ferramentas que suportam todo o ciclo da pesquisa biomédica, desde a aquisição e curadoria dos dados até a análise, visualização e integração das informações obtidas (Ambrósio, 2021). Novas técnicas de processamento de dados biológicos, modelagem e simulação computacional de sistemas biológicos têm possibilitado o surgimento de inovações tecnológicas na biologia, permitindo prever com alta confiabilidade o comportamento de sistemas biológicos diante de doenças ou intervenções (Chicco, 2017).

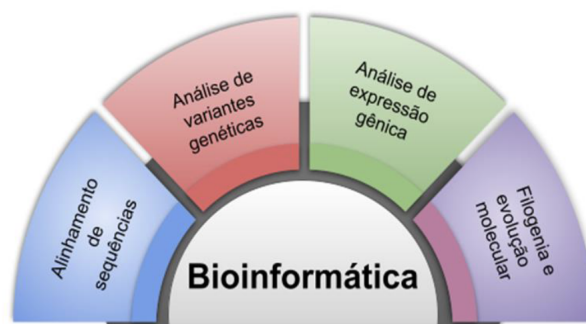
A precisão dos modelos e simulações depende diretamente da qualidade dos dados biológicos disponíveis, assim como da qualificação dos profissionais, uso de softwares especializados e equipamentos adequados, dessa forma a biologia computacional engloba estudos em áreas como bioinformática, modelagem e simulação de sistemas biológicos, interpretação e visualização de dados biológicos, armazenamento digital de informações biológicas, mineração de bases de dados biológicos, e reconhecimento e classificação de padrões em dados biológicos, vistos na Figura 10, Ridder, Ridder e Reinders (2013) complementa que seu objetivo é auxiliar no estudo de genômica, biologia sistêmica e biologia estrutural.

A integração da genética com a bioinformática tem possibilitado diagnósticos específicos para diversas doenças, permitindo previsões mais precisas sobre seu desenvolvimento. Esses avanços são resultado da capacidade de examinar detalhadamente o conjunto completo de genes de um indivíduo, abrangendo desde alterações cromossômicas até modificações em níveis individuais de DNA (Pereira; Oliveira; Sousa, 2020).

No contexto médico, ferramentas de bioinformática, conhecidas como *pipelines*, são utilizadas para analisar dados clínicos de pacientes, incluindo informações genéticas. Isso permite personalizar tratamentos e compreender melhor as bases genéticas de doenças hereditárias (Aronson; Rehm, 2015). O campo é diversificado e utiliza abordagens computacionais para interpretar e extrair informações dos dados biológicos, beneficiando pesquisas não apenas em saúde, mas também em diversas áreas do conhecimento.

Ao trabalhar com análise de dados, o pré-processamento é uma etapa fundamental. Ele auxilia na preparação dos dados brutos para análise, tornando-os mais limpos, organizados e adequados para a aplicação de técnicas analíticas. Em geral, essa fase envolve várias etapas, como detecção de erros, tratamento de valores ausentes, identificação de

Figura 10 – Aplicações de ferramentas de bioinformática.



Fonte: Silva e Alves (2024), Disponível em: <https://sea.ufr.edu.br/index.php/SEA/article/view/1872/1897>.

outliers e outros problemas de qualidade dos dados. Os dados são corrigidos ou removidos conforme necessário (Pandey; Perumal, 2023).

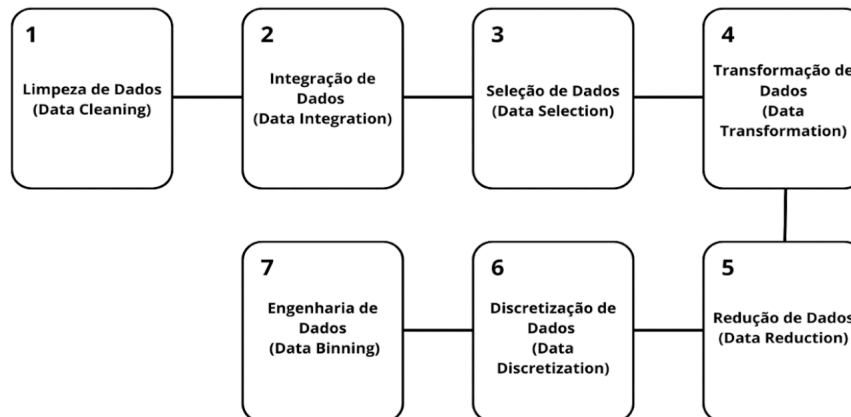
Além disso, o pré-processamento pode incluir a redução e transformação dos dados. Isso envolve técnicas como normalização, discretização e seleção de características., a normalização ajusta a escala dos dados para uma faixa específica, enquanto a discretização agrupa valores contínuos em intervalos discretos. A seleção de características é importante para escolher apenas aquelas mais relevantes e informativas para análises e modelagens subsequentes (Deelen *et al.*, 2015).

A agregação de dados também é uma etapa significativa, onde informações de diversas fontes são consolidadas em um conjunto único e coeso. Isso pode incluir a combinação de tabelas, resolução de discrepâncias de esquema e unificação dos dados para garantir consistência e integridade nas análises (Duggal *et al.*, 2019), essas etapas de pré-processamento, explicadas na Figura 11, garantem que os dados estejam prontos para serem utilizados de forma eficaz em análises posteriores, melhorando a qualidade e a precisão dos resultados obtidos.

Na etapa de limpeza de dados, ocorre a transformação dos dados para prepará-los para análise. Isso pode incluir normalização para padronizar as escalas das características, discretização de variáveis contínuas para simplificar a análise, e aplicação de técnicas matemáticas para melhorar a qualidade dos dados (Dutheil; Figuet, 2015). Em conjuntos de dados extensos, reduzir a quantidade de dados pode ser necessário para diminuir complexidade e tempo de processamento. Técnicas como a Análise de Componentes Principais (PCA) são úteis para reduzir a dimensionalidade dos dados, mantendo as informações mais relevantes (Frith; Wan; Horton, 2010).

Além disso, o tratamento de dados pode envolver a transformação de variáveis contínuas em variáveis discretas, simplificando assim a análise e a modelagem, a engenharia

Figura 11 – Etapas de pré-processamento, redução e transformação de dados.

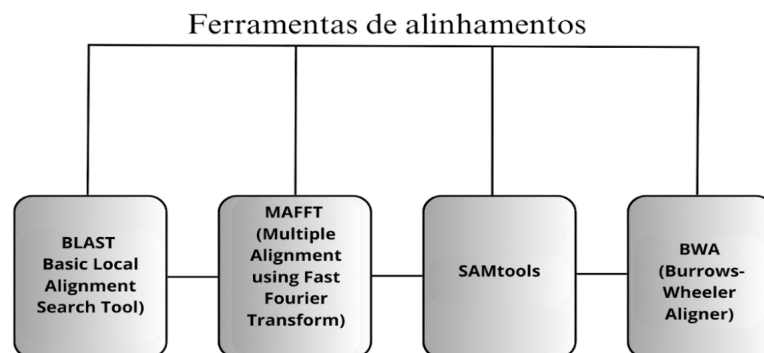


Fonte: Silva e Alves (2024), Disponível em: <https://sea.ufr.edu.br/index.php/SEA/article/view/1872/1897>.

de dados é outra etapa importante, onde novos atributos são criados a partir dos existentes para capturar informações importantes ou criar características mais significativas para tarefas de aprendizado de máquina (Lan *et al.*, 2018).

No contexto da bioinformática, o alinhamento de sequências genéticas é essencial para comparar e identificar semelhanças e diferenças entre sequências de DNA ou proteínas. Existem várias ferramentas de bioinformática que são amplamente usadas, conforme ilustrado na Figura 12, amplamente utilizadas para alinhamento, tanto local quanto múltiplo, que são fundamentais para pesquisas e estudos genômicos (Altschul *et al.*, 1990; Li; Godzik, 2006).

Figura 12 – Ferramentas de alinhamentos mais comuns em análises de sequências genéticas.



Fonte: Silva e Alves (2024), Disponível em: <https://sea.ufr.edu.br/index.php/SEA/article/view/1872/1897>.

O algoritmo BLAST (*Basic Local Alignment Search Tool*) foi desenvolvido como uma maneira de realizar buscas de similaridade de sequência de DNA e proteína por um algoritmo (Mount, 2007), é um método de alinhamento local utilizado para identificar

regiões homólogas em sequências biológicas. Este método é fundamental para a anotação de genes, busca por similaridades funcionais e inferência de evolução molecular, podendo ser aplicado em diversas áreas, desde pesquisa em genômica até a identificação de patógenos em amostras clínicas (Pearson, 2014).

O MAFFT (*Multiple Alignment using Fast Fourier Transform*) é uma ferramenta amplamente reconhecida para o alinhamento de múltiplas sequências biológicas devido à sua eficiência computacional e precisão (Katoh; Asimenos; Toh, 2009), a dificuldade em alinhar as sequências distantemente relacionadas corretamente sem inspeções manuais por conhecimento especializado. Muitos esforços foram feitos nos problemas relativos à otimização do alinhamento de sequências (Katoh *et al.*, 2002).

A ferramenta SAMtools, consiste em um suíte de programas amplamente reconhecida por sua capacidade de processar, manipular e analisar dados provenientes do Sequenciamento de Nova Geração (NGS) (Spencer *et al.*, 2014), e o BWA (Burrows-Wheeler Aligner) altamente eficiente para o alinhamento de sequências genômicas, amplamente reconhecida por sua capacidade de mapear rapidamente sequências de DNA em um genoma de referência (Jo; Koh, 2015).

A análise de variantes pode ser integrada para identificar tanto Polimorfismos de Nucleotídeo Único (SNPs) quanto Inserções e Deleções (Indels) (Lam *et al.*, 2012). As ferramentas de detecção identificam as posições onde as sequências diferem no conjunto de dados comparados; algoritmos como GATK (*Genome Analysis Toolkit*), SAMtools e FreeBayes são frequentemente utilizados para essa finalidade (Yao *et al.*, 2020).

O emprego de ferramentas de bioinformática na análise de dados genéticos é essencial para avanços expressivos na pesquisa genômica. No entanto, alguns pontos precisam ser avaliados, entre os principais desafios está a complexidade dos dados genéticos, que podem ser extensos e heterogêneos. Isso exige o desenvolvimento de algoritmos e métodos computacionais sofisticados para o processamento e interpretação dessas informações, além disso, a qualidade dos dados é crucial, e erros na coleta ou sequenciamento podem comprometer a precisão das análises. Dessa maneira, as ferramentas de bioinformática estão constantemente se aprimorando, o que demanda atualização contínua de conhecimento e habilidades por parte dos pesquisadores.

Steiger, Ritchie e Laurance (2016), Foley *et al.* (2007) afirmam que o surgimento e o ressurgimento de arbovírus geralmente não estão relacionados somente as mudanças ambientais, como o desmatamento, mas também ao uso da terra e à localização geográfica das famílias dentro de uma área, a necessidade de vetores para a transmissão de várias das viroses emergentes e reemergentes introduz fatores ecológicos de importância na discussão que se efetiva nos países de clima tropical (Schatzmayer, 2001).

O surgimento de novos agentes de doenças é resultado das mudanças sociais e

ambientais ao longo da história humana, fazendo com que os patógenos sejam capazes de adquirir acesso a novas populações hospedeiras ou de se tornarem mais virulentos em indivíduos comprometidos imunologicamente (Pignatti, 2004), as doenças infecciosas por tanto são marcadores de processos ecológicos que participam ao menos duas populações, a do hospedeiro e a do parasita e, frequentemente, várias outras, entre vetores e reservatórios complementa Souza, Minayo e Minayo (1995).

Na região amazônica, as espécies de mosquitos são diversas e a relação entre a diversidade de mosquitos e a transmissão de doenças é influenciada por mudanças na paisagem relacionadas ao desmatamento, construção de estradas, estabelecimento de assentamentos, diversificação de atividades produtivas relacionadas à agricultura familiar e moradias não planejadas (Foley *et al.*, 2007). O ambiente pode influenciar de forma positiva no desenvolvimento de novos surtos na América Latina da febre do Oropouche, pois a falta de vigilância em locais remotos acaba aumentando a fronteira de expansão do Oropouche (Casseb *et al.*, 2015).

Algumas pesquisas retratando os fatores ambientais e ecológicos que condiciona na sua disseminação, aspectos ecológicos comuns associados tanto aos locais envolvidos na detecção de OROV, as características ecológicas tiveram o cultivo de safras como o uso antropogênico mais comum da terra, rios como a fonte de água mais comum e floresta primária como o tipo mais proeminente de cobertura de terra (Walsh; Robert; Christoferson, 2021). Esses fatores ambientais podem ser considerados, incluindo precipitação anual, temperatura média anual, elevação, sazonalidade térmica e pluviométrica, amplitude térmica e variação da temperatura diurna.

Estudos sobre a associação de fatores ambientais com a probabilidade de ocorrência de surtos de febre Oropouche, Mayaro, Saint Louis e Rocio, descobriu que a temperatura é uma variável central que determina a distribuição de áreas de alto risco. Esse fato é preocupante, devido a temperatura média de algumas áreas terem aumentado significativamente ao longo do tempo, como resultados, os dados sugerem de forma robusta que essas quatro doenças têm o potencial de se tornarem importantes problemas de saúde pública ou se tornarem cada vez mais relevantes no Brasil e em outras áreas tropicais nos próximos anos e devem ser monitoradas como parte de programas de controle eficazes (Lorenz *et al.*, 2017).

A definição mais popular de *Big Data* são os 5 Vs, que são Volume, Velocidade, Variedade, Verificação/Veracidade e Valor afirma Huang *et al.* (2015), *Big Data* é um termo usado para descrever conjuntos de dados com volume ou complexidade tão grandes que os métodos convencionais não são o suficiente para lidar com eles, o conceito de *Big Data* foi descrito de forma diferente por pessoas diferentes (Baro *et al.*, 2015).

A análise de *Big Data* abrange a coleta, manipulação e análises de conjuntos de dados massivos e diversos que contêm uma variedade de tipos de dados, incluindo

dados genômicos e Registros Eletrônicos de Saúde (EHRs) para revelar padrões ocultos e correlações enigmáticas (Raghupathi; Raghupathi, 2014), por sua eficácia, a análise de *Big Data* é amplamente utilizada em diferentes campos de pesquisa (Chute *et al.*, 2013).

Embora o desenvolvimento das tecnologias em Sequenciamento de Nova Geração (NGS) tenha tornado cada vez mais fácil sequenciar um genoma inteiro ou exoma, ainda existem desafios consideráveis em termos de manuseio, análise e interpretação das informações genômicas geradas pelo NGS (He; Ge; He, 2017).

As infraestruturas de *Big Data* apresentam um potencial significativo para otimizar a análise de grandes volumes de dados genômicos. Um exemplo relevante é o uso do *Burrows-Wheeler Aligner* (BWA), um pacote de software para mapear sequências pouco divergentes contra um grande genoma de referência, como o genoma humano (Li, 2024).

O BWA foi adaptado para plataformas de *Big Data*, o qual demonstra uma capacidade de aumentar a velocidade de alinhamento genético em até 36 vezes quando comparado à versão original do BWA. Essa melhoria substancial de desempenho evidencia o impacto das tecnologias de *Big Data* na aceleração de processos computacionais complexos, possibilitando análises mais rápidas e eficientes em pesquisas genômicas (Abuín *et al.*, 2015).

2.5 Métodos de Machine Learning e Deep Learning em Bioinformática: Inovações em Análises Filogenéticas e Predição de Genótipos

O crescimento exponencial da quantidade de dados biológicos disponíveis levanta dois problemas: primeiro, o armazenamento e gerenciamento eficientes de informações e, por outro, a extração de informações úteis desses dados. O segundo problema é um dos principais desafios da biologia computacional, que requer o desenvolvimento de ferramentas e métodos capazes de transformar todos esses dados heterogêneos em conhecimento biológico sobre o mecanismo subjacente (Larranaga *et al.*, 2006).

Ferramentas e métodos podem transcender a simples descrição dos dados, proporcionando a obtenção de conhecimento por meio de modelos passíveis de verificação empírica por meio dessa abstração simplificadora, representada pelos modelos, será possível gerar previsões sobre o comportamento do sistema em estudo.

Para Mitchell (2006), Friedman (2009), aprendizado de máquina, ou *Machine Learning* (ML), é um ramo da Inteligência Artificial que explora o estudo e a construção de algoritmos computacionais a partir do aprendizado através dos dados, tais aprendizados de máquina é um subcampo da ciência da computação que busca uma interseção de técnicas matemáticas e estatísticas com algoritmos computacionais (Deo, 2015; Obermeyer;

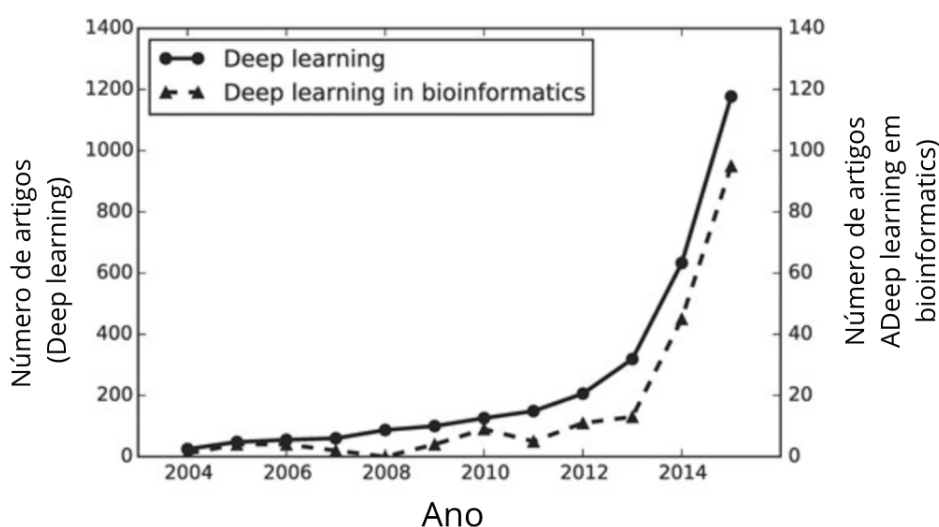
Emanuel, 2016).

Os algoritmos de ML já estão entrelaçado em diversas áreas, como sistemas bancários para detecção de fraudes, mecanismos de busca na internet, sistemas de vigilância em vídeo, segurança de dados, logística de empresas, robótica e, na medicina, para diagnóstico e prognóstico (Zhou *et al.*, 2017). Conforme mencionado pelos autores, Darcy, Louie e Roberts (2016), Waljee e Higgins (2010), esses algoritmos com o conceito de IA pode ser aplicada em determinadas situações em que se busca padrões em um conjunto de variáveis com o intuito de prever um resultado específico de interesse.

A *Deep Learning* por Oracle (2024), é um subconjunto da ML, os modelos computacionais e suas múltiplas camadas de processamento aprendam representações de dados com múltiplos níveis de abstração, ate o ponto de melhorar drasticamente o estado da arte em reconhecimento de fala, reconhecimento de objetos visuais, detecção de objetos e muitos outros domínios, como descoberta de medicamentos e genômica (LeCun; Bengio; Hinton, 2015).

O desempenho adequado dos algoritmos convencionais da ML depende muito de representações de dados para Goodfellow, Bengio e Courville (2016), a *Deep Learning* emergiu no campo da IA, impulsionada pelo crescimento das bases de dados em larga escala, *Big Data* pelo autor Sagiroglu e Sinanc (2013), desde o início dos anos 2000, o interesse acadêmico por *Deep Learning* tem aumentado de forma exponencial, refletindo seu potencial transformador em diversas áreas de pesquisa e aplicação visto na Figura 13.

Figura 13 – Número aproximado de artigos de aprendizagem profunda publicados por ano. O número de artigos é baseado nos resultados da pesquisa em <http://www.scopus.com> com as duas consultas: ‘Deep Learning’, ‘Deep Learning’ AND ‘bio*’..



Fonte: Min, Lee e Yoon (2017).

A bioinformática tem passado por uma revolução contínua em suas técnicas, com o objetivo de alcançar resultados mais precisos e robustos. Paralelamente, o aprendizado de máquina vem sendo amplamente aplicado em diversos contextos, devido à sua capacidade de lidar com grandes volumes de dados e gerar percepções significativas. Nesse sentido, a aplicação de algoritmos de ML e *Deep Learning* em análises filogenéticas e na predição de genótipos se destaca como uma abordagem promissora, capaz de aprimorar a precisão e a eficiência dessas análises, proporcionando avanços significativos no campo.

As principais abordagens para reconstrução da filogenia dependem de modelos evolutivos probabilísticos que descrevem os processos estocásticos de substituições de nucleotídeos, aminoácidos e códonos (Thorne, 2000).

Como afirma Chor e Tuller (2005), a busca computacional pela topologia de árvore de ML mostrou ser difícil, uma abordagem geral para uma busca heurística de máxima verossimilhança é começar com uma árvore inicial aleatória ou com uma árvore inicial obtida por métodos rápidos e geralmente menos precisos, como NJ (Trees, 1987; Ogden; Rosenberg, 2006).

A criação do ModelTeller, uma metodologia computacional para seleção de modelos filogenéticos, desenvolvida dentro da estrutura de aprendizado de máquina, otimizada para prever o modelo de substituição de nucleotídeos mais preciso para estimativa de comprimento de ramo, o mesmo depende de um modelo de aprendizado de máquina prontamente implementado e, portanto, a previsão de acordo com recursos extraídos dos dados de sequência resulta em uma diminuição substancial no tempo de execução em comparação com as estratégias existentes afirma Abadi *et al.* (2020).

Foi desenvolvido um algoritmo de aprendizado de máquina com o objetivo de otimizar a busca em árvores filogenéticas no contexto de máxima verossimilhança ML. Especificamente, utilizou-se um modelo de *Deep Learning* combinado com regressão baseada em floresta aleatória para prever a classificação de todas as modificações possíveis durante o processo de poda e reencaixe de subárvores. Nas palavras de Azouri *et al.* (2021), o modelo treinado foi capaz de fornecer uma estimativa precisa das classificações para cada movimento possível, com base nos respectivos valores de mérito, facilitando assim a seleção eficiente das melhores estratégias de reconstrução filogenética.

2.6 Unindo Ferramentas: Pipelines Bioinformáticos Flexíveis

Em diversas áreas, inclusive na ciência, os avanços tecnológicos e conceituais têm levado a uma geração acelerada e contínua com grandes volumes de dados. Embora a pesquisa científica sempre tenha envolvido a coleta e organização de informações, o cenário atual, caracterizado pela próspera quantidade, diversidade e velocidade de produção de dados científicos, dá início a novas oportunidades e desafios, tais desafios são

particularmente evidentes em relação à escalabilidade e à complexidade dos dados, exigindo abordagens inovadoras para o gerenciamento e a análise eficientes dessas informações.

A interoperabilidade e integração de ferramentas bioinformáticas, como *Galaxy*, *Nextflow* e *Snakemake*, são fundamentais para enfrentar os desafios da análise genômica moderna. Pequenas variações entre plataformas computacionais também contribuem para irreprodutibilidade computacional ao produzir instabilidade numérica, que é especialmente relevante para ambientes computacionais de alto desempenho (luigi, 2024).

Nos últimos anos, diversos *pipelines* para análise de sequenciamento de RNA (RNA-Seq) foram desenvolvidos e publicados, certos aspectos de implementação e a escolha das ferramentas utilizadas podem se sobrepor, a instalação e execução reprodutíveis, assim como a facilidade de uso, apresentam diferenças, tais destaques ressaltam a importância de abordagens específicas para garantir a reprodutibilidade e a acessibilidade nos workflows de análise de dados transcriptômicos.

O *pipeline* de RNA-Seq fornecido pela comunidade nf-core, se resume no esforço para coletar um conjunto selecionado de *pipelines* de análise criados usando o Nextflow (Nextflow, 2024), quando a análise é feita em um conjunto de dados muito grandes, a principal fonte de irreprodutibilidade computacional surge de uma falta de boas práticas relativas ao uso de software e banco de dados conforme os autores Albrecht *et al.* (2012), McKenna *et al.* (2010), uma solução para essa instabilidade é o Nextflow, um sistema de gerenciamento de fluxo de trabalho que utiliza a tecnologia Docker para o manuseio multi-escala de computação em contêineres, complementa Tommaso *et al.* (2017).

Conforme explicado por Mölder *et al.* (2021), *Nextflow* e *Snakemake*, são fluxos de trabalho especificados usando uma linguagem específica de domínio (DSL), que fornece instruções e declarações que modelam especificamente os componentes centrais do gerenciamento de fluxo de trabalho, evitando operadores supérfluos ou código boilerplate. Uma vez que a DSL é implementada como uma extensão para uma linguagem de programação genérica (Groovy e Python), o acesso ao poder total da linguagem de programação subjacente é mantido.

O Galaxy é uma plataforma web que tem como principal objetivo tornar simples o manuseio de diversas ferramentas para a utilização, ferramentas que são complexas para o manuseio, configuração e instalação por um usuário comum, afirma Jalili *et al.* (2020), com o objetivo de tornar a análise de grandes volumes de dados mais acessível, transparente e reprodutível, por meio de um ambiente web, em que os usuários podem executar análises computacionais complexas e ter todos os detalhes de cada etapa da execução registrados para posterior inspeção, publicação ou reutilização (NETO *et al.*, 2016).

Essas plataformas permitem a construção de *pipelines* modulares e escaláveis, facilitando a personalização e automação de workflows complexos. Ao combinar diferentes

ferramentas em um ambiente unificado, é possível melhorar a eficiência, promover a reutilização de código e garantir a adaptabilidade das análises a diferentes conjuntos de dados e requisitos de pesquisa.

Nas reflexões dos autores Digiampietri, Setubal e Medeiros (2008), nos últimos anos, várias ferramentas para o processamento de atividades em bioinformática foram disponibilizadas como serviços web. No entanto, a ausência de padronização das interfaces desses serviços e da descrição de suas funcionalidades representa um obstáculo significativo para o seu uso eficaz. Para promover a interoperabilidade entre os serviços e facilitar sua integração em workflows bioinformáticos, é essencial estabelecer padrões que definam claramente as interfaces e as funcionalidades oferecidas, isso possibilitará uma melhor comunicação entre diferentes ferramentas e a construção de pipelines mais eficientes e reproduzíveis.

Um problema suplementar na interpretação dos dados genômicos surge da falta de integração entre as diferentes análises e ferramentas, sendo que muitas delas sequer são projetadas para interagirem e complementarem outras, afirma Bromberg e Capriotti (2013).

A formalização de análises por meio de fluxogramas e a integração de ferramentas em pipelines contribuem significativamente para a redução do tempo de execução das análises e para a minimização da manipulação manual dos dados, resultando em maior eficiência e reprodutibilidade dos estudos científicos. Laboratórios de médio e pequeno porte, que muitas vezes carecem de suporte especializado em bioinformática, dependem frequentemente de ferramentas desenvolvidas por terceiros, no entanto, mesmo com a disponibilidade de plataformas já citadas nesse texto, o uso dessas soluções pode representar um desafio para pesquisadores com formação limitada em computação, evidenciando a necessidade de interfaces mais intuitivas e de suporte adequado para facilitar a adoção dessas tecnologias (Marx, 2013; D'Antonio *et al.*, 2013).

A integração do conhecimento biológico é realizada por meio da mineração de dados em larga escala, caracterizando-se pelo uso de informações provenientes de múltiplos conjuntos de dados com o objetivo de responder a questões biológicas específicas. No contexto da associação entre variantes genéticas e fenótipos, a integração requer metodologias que sejam não apenas escaláveis, mas também eficientes na consulta e recuperação de informações, além disso, é essencial a capacidade de combinar dados oriundos de diferentes experimentos e retornar informações relevantes dos repositórios apropriados, garantindo uma análise abrangente e rigorosa, afirmam os autores Goldstein *et al.* (2013), Thorisson, Muilu e Brookes (2009).

Conforme explicado por Souza *et al.* (2014), a integração de bases de dados e análises genéticas pode, não apenas responder e validar questões complexas no contexto biológico, como também inspirar novos paradigmas computacionais. Além disso, aplicações

com grande volume de dados relacionados e grande potencial de distribuição e colaboração apresentam um desafio para o desenvolvimento de ferramentas computacionais de integração eficientes.

Existem duas metodologias importantes que são responsáveis pela variação genética: 1ª mutação e 2ª recombinação. A mutação é especialmente relevante pois é a primeira fonte da mudança evolutiva, os novos alelos produzidos por mutação resultando em molde para um segundo nível de organização, efetuado por recombinação (Griffiths *et al.*, 2006).

Polimorfismos de Nucleotídeo Único (SNPs), discutidos em seção 2.4 são posições de pares de bases simples no DNA genômico onde diferentes alternativas de sequência (alelos) existem em indivíduos normais em algumas populações, em que o alelo menos frequente tem uma abundância de 1% ou mais (Brookes, 1999). Alguns autores como Kahl *et al.* (2005) consideram Indels (adição de nucleotídeos extras ou a exclusão de um nucleotídeo) como SNPs, embora eles certamente ocorram por um mecanismo diferente .

As alternativas são a forma mais frequente de diversidade genética entre indivíduos de uma espécie, um SNP pode substituir o nucleotídeo citosina (C) pela timina (T) numa determinada sequência do DNA (Health *et al.*, 2019) ilustrado na Figura 14, ou seja, um SNP é um lugar de uma letra onde seu genoma varia de outra sequência de genoma (National Human Genome Research Institute, 2024).

Figura 14 – Ilustração da SNPs.



Fonte: Biomedicina Padrão (2020).

Conforme mencionado pelos autores Picoult-Newberg *et al.* (1999), Oeveren e Janssen (2009) para, o processo de identificação dos SNPs uma das metodologias tais como o sequenciamento de produtos de Reação em Cadeia da Polimerase (PCR), a identificação eletrônica de SNPs utilizando como base, por exemplo, as bibliotecas de EST (*Expressed*

Sequence Tags) ou bibliotecas genômicas.

Novas tecnologias como as de Sequenciamento de Nova Geração (NGS), juntamente com ferramentas computacionais existentes, são demasiadamente eficazes e robustas nas descobertas de SNPs sem um genoma de referência (Zolet *et al.*, 2017).

Além disso, as novas tecnologias e os dados de acesso à informação genômica das espécies constituem recursos essenciais no processo de busca e identificação de marcadores SNPs, seja em genes candidatos ou dispersos ao longo do genoma. Ferramentas como SAMtools Li *et al.* (2009) e GATK (McKenna *et al.*, 2010), que empregam diferentes algoritmos de detecção de variantes, foram desenvolvidas para análise tanto de amostras únicas quanto múltiplas de forma paralela (Nielsen *et al.*, 2011; Kumar; Banks; Cloutier, 2016).

Afim de validar as variantes genéticas que possuem variações individuais como distintas do genoma de referência, uma série de ferramentas foram desenvolvidas, incluindo GATK, uma estrutura de programação estruturada projetada para facilitar o desenvolvimento de ferramentas de análise eficientes e robustas para sequenciadores de DNA (Yang *et al.*, 2007).

Como detector de variantes genéticas bayesianas existe o Freebayes, projetado para encontrar pequenos polimorfismos, indels, MNPs (polimorfismos de nucleotídeos múltiplos), marcadores potencialmente mais eficientes, capazes de determinar com precisão a identidade entre diferentes variedades vegetais, bem como no nível de espécie afirma Fang *et al.* (2021), e eventos complexos, onde ele chama variantes com base nas sequências literais de leituras alinhadas a um alvo específico, não em seu alinhamento preciso, sendo um modelo de generalização direta de modelos anteriores, por exemplo GATK, (Garrison; Marth, 2012).

No entendimento dos autores Keel e Snelling (2018), Fonseca *et al.* (2012), algumas ferramentas possuem uma alta cobertura de leituras de sequências curtas que foram geradas nas mesmas regiões de um genoma, a primeira tarefa é alinhar essas leituras de sequência às regiões correspondentes de um genoma de referência, na atualidade já existem mais de 60 algoritmos diferentes para mapear leituras de sequência para um genoma de referência.

2.7 Genes e Justiça: Ética, Privacidade e Segurança em Dados Genômicos Humanos

Pelas palavras de Capurro (2005), atualmente a transparência de dados e a garantia da proteção sobre os mesmos estão inseridos no mundo atual, a privacidade fundamentalmente sob a denominação “autonomia informacional”, que consiste no poder de escolha do indivíduo acerca do uso da informação em um ambiente eletrônico, o uso de dados genômicos levanta uma série de implicações éticas e legais que são fundamentais para a

pesquisa, desenvolvimento de tratamentos e políticas de saúde pública.

O uso de dados genômicos possui um potencial significativo para avanços em pesquisas, porém requer uma abordagem cuidadosa a fim de garantir que as implicações éticas e legais sejam devidamente consideradas e geridas. Pesquisadores, profissionais de saúde trabalham juntos para criar um ambiente que respeite os direitos dos indivíduos, promovendo ao mesmo tempo a inovação e o progresso na área da genômica.

A pesquisa genética no Brasil ganhou espaço na Constituição Federal de 1988, todos têm direito ao meio ambiente ecologicamente equilibrado, bem de uso comum do povo e essencial à sadia qualidade de vida, impondo-se ao poder público e à coletividade o dever de defendê-lo e preservá-lo para as presentes e futuras gerações (Brasil, 1988), deixando exposto no 1º parágrafo:

- II - preservar a diversidade e a integridade do patrimônio genético do País e fiscalizar as entidades dedicadas à pesquisa e manipulação de material genético (Brasil, 1988);

Conforme o autor Clotet (1997), todo ser humano tem uma identidade genética própria, o genoma humano, ou a série completa de genes da espécie humana, é propriedade inalienável da pessoa e, por sua vez, um componente fundamental do patrimônio comum da humanidade, a discussão sobre banco de dados não é recente, em 1998 já se articula sobre as vantagens e os desvantagens de se ter um banco nacional com os dados de DNA (“*fingerprint*” ou impressões genéticas) da população, seria um benefício ou uma ameaça?, complementa o autor Dawkins (1998).

As agências internacionais responsáveis pelo monitoramento das legislações nesta área catalogam sete características onde as pessoas apresentam vulnerabilidade das informações genéticas pessoais e devem ser normatizadas pelo legislador a fim de evitar abusos, são elas: bancos de dados de DNA, bancos de dados forenses, relação de emprego, proteção de dados e privacidade, pesquisas em geral, adoção e seguradoras (Crosbie, 2000)

Os dados genéticos humanos são informações obtidas após a realização de testes genéticos, no entanto, nunca podem ser interpretados de forma isolada, pois eles não são absolutos e devem ser analisados com muita cautela por profissional habilitado na área, os dados genéticos podem apresentar informações de riscos que determinada pessoa pode ter de desenvolver uma doença, antes mesmo da sua manifestação (Dantas; Andrade *et al.*, 2022).

2.8 Genomas em Comparação: Explorando a Evolução e Adaptação Viral

A genômica comparativa desempenha um papel central na compreensão da evolução viral, da adaptação ao hospedeiro e da emergência de novas variantes. Ao comparar genomas de diferentes linhagens virais do Oropouche, bem como de espécies relacionadas dentro do gênero, é possível identificar padrões evolutivos, traçar relações filogenéticas e revelar mutações críticas que influenciam o comportamento do vírus em termos de transmissão, patogenicidade e resposta a intervenções.

O conceito de genômica comparativa, que se baseia principalmente no estudo da homologia e da dinâmica evolutiva dos organismos, seus respectivos genes e proteínas, o que apresenta grande utilidade no entendimento da evolução das espécies pela comparação de seus genomas completos ou de genes específicos de cada espécie (Hardison, 2003).

Para Touchman (2010), a genômica comparativa é uma área das ciências biológicas dedicada à análise comparativa das sequências genômicas de diferentes espécies, os genomas são conhecidos por passarem por diversos eventos de rearranjo e recombinação ao longo do processo evolutivo, os quais podem resultar na duplicação de genes, formando famílias multigênicas, ou na perda de genes ortólogos.

A análise comparativa dos genomas podem fornecer *insights* valiosos sobre os processos evolutivos que moldam a diversidade genômica do vírus, a partir da comparação de sequências genômicas de linhagens coletadas em diferentes regiões geográficas e períodos de tempo, é possível.

Esses eventos permitem que comparações genômicas ofereçam uma compreensão aprofundada sobre a dinâmica da organização estrutural dos genomas, podem revelar mutações específicas que conferem vantagens evolutivas, como maior eficiência na transmissão entre vetores e hospedeiros humanos ou maior resistência a respostas imunes.

Como destacado por Darling *et al.* (2004), fornecendo evidências essenciais para estudos evolutivos ao identificar genes conservados e exclusivos entre diferentes organismos e a disponibilidade de sequências de genomas nos bancos de dados públicos pressionam novos desafios para a análise comparativa dessas sequências, fazendo surgir a necessidade de ferramentas de visualização que facilitem as análises.

Possibilitando a aplicação de descobertas realizadas em espécies já sequenciadas e analisadas a organismos em estudo a genômica comparativa, permitido avanços significativos, sobretudo na identificação de genes e RNAs não codificadores, representando algumas das aplicações mais comuns da genômica comparativa (Coimbra, 2013).

Em diversos estudos, conforme a fala dos autores Junior (2002), Delcher *et al.* (1999), encontram-se distintas metodologias para comparação genômica, muitas das quais

utilizam exclusivamente informações de sequências de DNA, enquanto outras se baseiam em sequências de genes codificantes, alguns estudos utilizam o algoritmo BLAST como ferramenta comparativa (Viana, 2006).

Atualmente, existe um programa que possibilita a comparação de genomas entre diferentes espécies, sem se restringir apenas a fungos. Essa ferramenta permite a caracterização de genes de qualquer organismo, gerando informações em poucos segundos e possibilitando a análise simultânea de múltiplos genomas, sendo viável identificar, por exemplo, quais espécies de fungos apresentam maior virulência ou menor capacidade infecciosa (Coimbra, 2013).

É de total conhecimento a existência do vírus da gripe, me arrisco a dizer que é um dos mais conhecidos, e é um exemplo de vírus cujas sequências genômicas tiveram impactos significantes. Um evento crucial na revolução genômica do vírus da influenza A, onde a implementação do Projeto de Sequenciamento do Genoma da Gripe (IGSP) possibilitou o sequenciamento sistemático e a análise detalhada de seu genoma (Ghedin *et al.*, 2005).

Como mencionado pelo autor Nelson *et al.* (2006), Nelson *et al.* (2007), o processo de análise da sequência completa do genoma revela como, ao colocar segmentos de genes em novas configurações genômicas, a recombinação pode vir a gerar isolados com propriedades antigênicas alteradas, o que, por sua vez, pode levar à falha da vacina.

A definição de genótipos do vírus Oropouche tem sido historicamente limitada devido à ausência de consenso e à dependência de estudos baseados em sequenciamento genômico parcial ou conjuntos de dados disponíveis publicamente. Conforme a evolução do vírus, se tornam linhagens mais complexas, torna-se evidente a necessidade de uma estrutura mais robusta e prospectiva para a classificação genotípica.

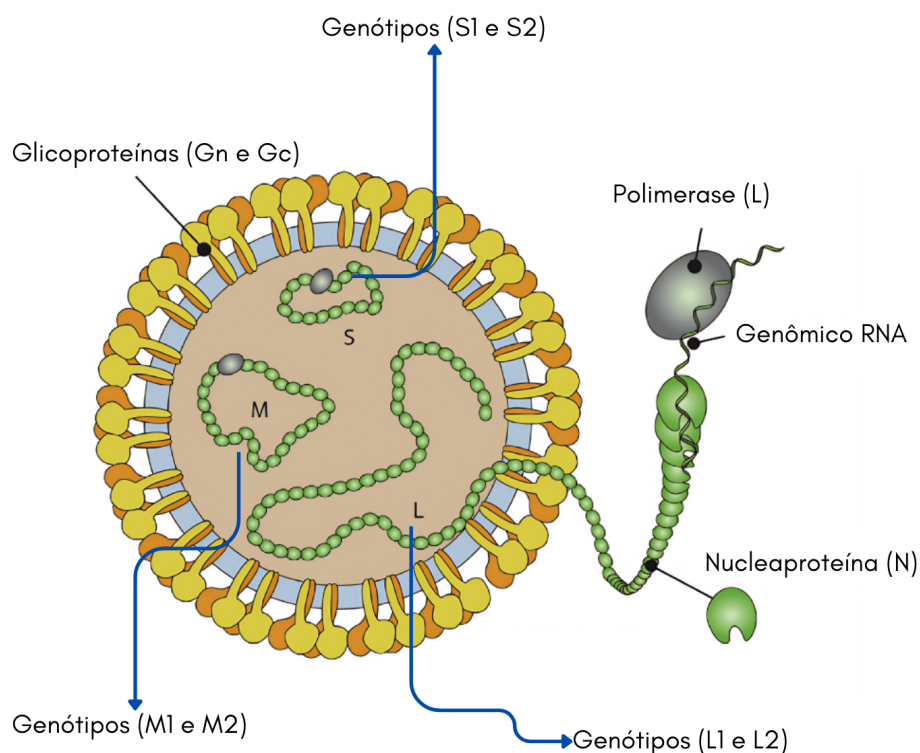
A ferramenta desenvolvida incorpora aspectos de genômica comparativa ao utilizar bancos de dados genômicos para classificação de linhagens e análise evolutiva, dessa forma permitindo, por exemplo, na comparação de sequências genômicas, traçar relações evolutivas entre linhagens e identificar agrupamentos de variantes emergentes.

Dessa forma, uma abordagem inovadora para definição de genótipos, baseada em múltiplos critérios, tais como monofilia, distâncias genéticas intra e intergrupos e diversidade genética líquida dentro de grupos. Essa estrutura visa aprimorar a resolução taxonômica e estabelecer diretrizes mais claras e replicáveis para a classificação genotípica.

Dentre os desafios na classificação do vírus Oropouche, a ausência de uma convenção de nomenclatura amplamente aceita, as nomenclaturas historicamente empregadas frequentemente associam genótipos a localizações geográficas, o que se mostra inadequado diante da disseminação global do vírus. Nesse contexto, propomos uma convenção de nomenclatura padronizada, ilustrada na Figura 15, que utiliza identificadores baseados

em segmentos genômicos combinados com numerais, alinhando-se à estrutura genética do vírus e permitindo a incorporação de novas linhagens emergentes.

Figura 15 – Nomenclatura padronizada sugerida.



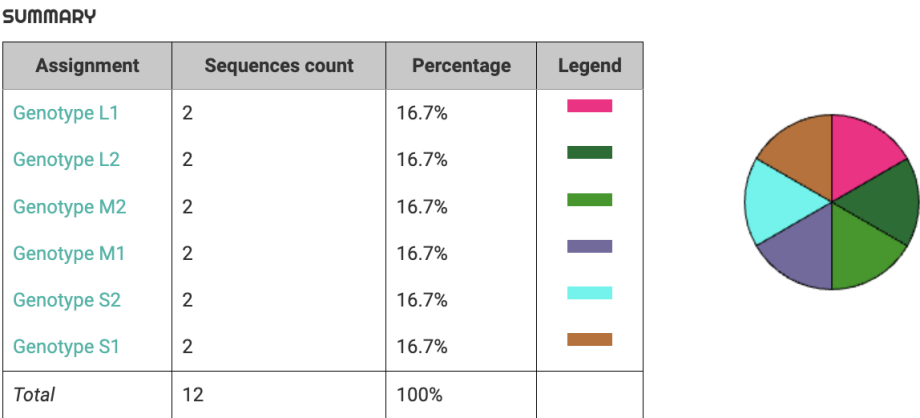
Fonte: Romero-Alvarez e Escobar (2018, p. 3), adaptado pelo autora.

Para o vírus Oropouche, cujo genoma é segmentado em três partes (Acrani *et al.*, 2015) sugerimos a identificação de dois genótipos para cada segmento:

- a) S1 e S2:
 - segmento Pequeno(S);
- b) M1 e M2:
 - segmento Médio(M);
- c) L1 e L2:
 - segmento Grande(L).

Dispondo como resultado a Figura 16, com os detalhes dos segmentos com a nova nomenclatura proposta neste trabalho.

Figura 16 – Resultado da nova nomenclatura padronizada sugerida.



Fonte: Autora, 2024.

A proposta dessa nomenclatura foi promover a clareza e consistência, além de ser escalável para futuras descobertas genômicas. Como parte do desenvolvimento dessa estrutura, a seleção das sequências de referência que encapsulam a diversidade genética de cada genótipo foi proposta, a realização de testes de validação para assegurar a robustez da abordagem e a precisão na classificação de novas sequências virais.

Ao redefinir os critérios e a nomenclatura dos genótipos do vírus Oropouche, teve como objetivo estabelecer uma estrutura que acompanhe a evolução do vírus, promovendo maior compreensão e colaboração científica. Esse esforço visa não apenas aprimorar a classificação genotípica, mas também oferecer uma base sólida para investigações futuras sobre a diversidade e a evolução viral.

O pipeline permite identificar mutações recorrentes em diferentes linhagens, destacando possíveis regiões de interesse para estudos adicionais, como a investigação de mecanismos de virulência ou transmissibilidade.

De forma objetiva, a genômica comparativa oferece um caminho essencial para compreender a biologia do OROV e antecipar suas mudanças evolutivas, permitindo que a ferramenta desenvolvida forneça informações críticas para a vigilância epidemiológica e o controle do vírus.

2.9 Trabalhos Correlatos

Trabalhos correlatos, pesquisas e publicações, utilizando abordagens metodológicas similares e compartilhando objetivos comuns com este estudo. Essa base de conhecimento pré-existente proporcionou um contexto rico para a contextualização desta pesquisa.

Fonseca *et al.* (2019b) realizou o desenvolvimento de um método com o objetivo de

classificar sequências de vírus com relação às suas espécies e subespécies, a ferramenta foi validada em um grande conjunto de dados avaliando o desempenho da classificação com relação às sequências do genoma completo e sequências do genoma parcial. O *pipeline* de classificação descrito na pesquisa compreende dois componentes distintos. O primeiro componente permite a atribuição de espécies e subespécies, utilizando uma análise de classificação que submete uma sequência de consulta a uma busca BLAST contra um conjunto de sequências de referência. O segundo componente do *pipeline* envolve a construção de uma árvore filogenética pelo método Neighbor Joining (NJ), possibilitando a atribuição de genótipos e/ou subtipos.

A elaboração do método realizado por Juma *et al.* (2022b) que permite a classificação e atribuição de linhagens de sequências parciais de consenso gene de glicoproteína (GN) e sequências de genoma completo. Para isso, foi construído um banco de dados local utilizando o DIAMOND, que associa nomes taxonômicos, nós e acessos de proteínas aos identificadores taxonômicos obtidos do NCBI. A ferramenta foi avaliada utilizando sequências completas do genoma do Rift Valley Fever Virus (RVFV), gerado a partir de amostras clínicas de gado de um surto recente no Quênia, tais dados de sequência foram gerados usando tecnologias Illumina. Essas tecnologias têm sido usadas para conduzir a epidemiologia genômica de patógenos em diversas escalas de surtos. A análise filogenética é conduzida utilizando um conjunto de dados de referência filogenética padrão gerado por métodos de Neighbor-Joining (NJ), Máxima Verossimilhança (ML) e árvore Bayesiana. Os usuários podem especificar o conjunto de dados de referência filogenética desejado, onde as sequências de consulta são alinhadas ao alinhamento de sequências múltiplas do conjunto de dados utilizando MAFFT. Em seguida, uma árvore filogenética ML é construída e as linhagens são atribuídas. Um arquivo de saída no formato CSV é gerado, contendo as atribuições de linhagem, valores de bootstrap e resultados do teste de verossimilhança.

O Genome Detective, desenvolvido por Cleemput *et al.* (2020), é um aplicativo de software amigável e baseado na web para montar de forma rápida e precisa todos os genomas de vírus conhecidos a partir de conjuntos de dados de sequenciamento de última geração permitindo a identificação de clusters filogenéticos e genótipos a partir de genomas montados no formato FASTA. Essa ferramenta de tipagem de coronavírus do Genome Detective que pode identificar com precisão as novas sequências de coronavírus relacionadas à síndrome respiratória aguda grave, testada e validada com centenas de genomas completos de 10 espécies de coronavírus e classificou corretamente todos os coronavírus relacionados à SARS (SARSr-CoV) e todos os dados públicos disponíveis para SARS-CoV-2, a ferramenta também permite o rastreamento de novas mutações virais à medida que o surto se expande globalmente, o que pode ajudar a acelerar o desenvolvimento de novos diagnósticos, medicamentos e vacinas para deter a doença COVID-19.

Pineda-Peña *et al.* (2013) investigou diferenças na patogênese, diagnóstico e vias

de resistência entre subtipos de HIV-1, onde é necessário uma ferramenta de subtipagem precisa para grandes conjuntos de dados, o principal objetivo foi avaliar o desempenho de ferramentas de subtipagem automatizadas para classificar os diferentes subtipos e formas recombinantes circulantes usando pol, a região mais sequenciada na prática clínica. Também apresentamos a versão 3 atualizada da ferramenta de subtipagem Rega HIV (REGAv3), a mesma teve um desempenho melhorado para o subtipo B e CRF02_AG em comparação com REGAv2 e sendo capaz de identificar todos as formas recombinantes circulantes (CRFs) epidemiologicamente relevantes.

3 METODOLOGIA

A metodologia predominante para esta pesquisa é baseada na estratégia denominada como abordagem quantitativa e computacional, onde a busca por resultados precisos é conduzida por meio de variáveis previamente definidas, nas quais se verifica e se explica a influência exercida sobre tais variáveis, através da análise da frequência de ocorrências e de correlações estatísticas (Doxsey; Riz, 2002), assim a análise foi feita de forma exata e objetiva, tornando a possibilidade de identificações de padrões e as correlações possíveis.

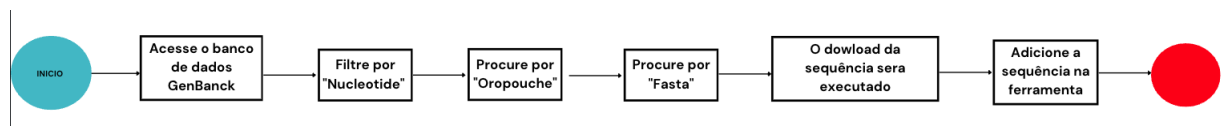
Por essa razão a análise de grandes volumes de dados com rigor estatístico e reprodutibilidade são realizáveis. Os dados utilizados foram capturados de banco de dados públicos, e processados para remover valores atípicos e padronizados para garantir a consistência nas análises onde alguns métodos foram utilizados.

O projeto inicia-se com o estudo de tópicos especiais nas seguintes temáticas entender melhor o vírus Oropouche e suas características genéticas:

- a) estrutura e mecanismos de acesso automatizado a bancos públicos para download:
 - sequências genéticas dos 3 segmentos do vírus do Oropouche;
 - metadados como data de coleta, país de origem e strain da amostra.

O processo para a construção do *pipeline* se inicia com a grande coleta de sequências genômicas, conforme o fluxograma Figura 17, que foram coletadas em bancos de dados públicos, como por exemplo o GenBank ou RefSeq, tais sequências algumas ações de pré-processamento garantem que apenas sequências relevantes e de alta qualidade sigam.

Figura 17 – Processo de mineração de dados do Genbank.



Fonte: Elaborada pela autora, 2024.

A escolha do BLAST e do ClustalW para essa etapa se deriva pela sua busca de similaridade e anotação das sequências de forma rápida e eficiente na identificação de regiões conservadas e homólogas em grandes bancos de dados de sequência.

Foi realizado uma análise básica de pesquisa de alinhamento local, popularmente conhecida no ramo da bioinformática como BLAST para encontra e identificar regiões de similaridade local entre sequências e assim identificar membros de famílias de genes e, assim, separa-las corretamente por cada segmento (Altschul *et al.*, 1990).

Nesse processo as informações e o potencial de relações evolutivas entre as sequências é possível, esses resultados fazem parte do processo de seleção das sequências para o alinhamento, assim ajudando a definição de um conjunto mais refinado para análises mais profundas.

As sequências que foram selecionadas passaram por um alinhamento múltiplo utilizando MAFFT, a mesma baseia-se na Transformada Rápida de Fourier (TRF) para efetuar o alinhamento múltiplo de DNA e proteínas, a ferramenta possui duas características que se destacam das demais:

- a) utilização de TRF para detecção de regiões homólogas;
- b) abordagem eficiente de pontuação das sequências para o alinhamento múltiplo (Katoh *et al.*, 2002).

Esta ferramenta possui o tempo de processamento que supera consideravelmente outras ferramentas que temos disponível, e ainda produz um alinhamento resultante de alta qualidade do ponto de vista biológico, contra ponto, quando as sequências selecionadas são pouco relacionadas ou nelas existem grandes trechos com inserção, o resultado é degradado (Katoh; Standley, 2013).

ClustalW é um algoritmo básico, que consiste em três estágios principais, afirma Peres (2006):

- a) todos os pares de sequências são alinhados separadamente para calcular a matriz de distância fornecendo assim a divergência entre cada par de sequências;
- b) uma árvore guia é calculada a partir desta matriz de distância;
- c) as sequências são progressivamente alinhadas de acordo com a topologia na árvore guia.

Dessa forma as árvores guias usadas para direcionar o processo final de alinhamento múltiplo são calculadas a partir da matriz de distância utilizando-se o método NJ, as árvores serão produzidas e em seguida enraizadas com comprimentos de ramos proporcionais a divergência entre as sequências, em seguida a árvore é enraizada no ponto médio, onde a distância média dos comprimentos de ramos de cada lado da árvore é equivalente (Thompson; Higgins; Gibson, 1994).

O alinhamento múltiplo permite identificar conservações e variações em regiões específicas das sequências, sendo um passo essencial para a construção da árvore filogenética. Durante essa etapa, parâmetros como a matriz de substituição e o método de alinhamento são ajustados para otimizar a precisão do alinhamento.

Todas as sequências serão separadas por segmentos e posteriormente alinhadas utilizando estratégias de alinhamento de sequências múltiplas, a estratégia a ser utilizada de alinhamento múltiplo são classificadas em três tipos:

- a) o método progressivo;
- b) o método de refinamento iterativo com a pontuação ponderadas da soma dos pares;
- c) o método de refinamento iterativo usando as pontuações ponderadas da soma dos pares e de consistência(Long; Li; Fu, 2016);

Com base nas características descritas, a combinação entre elas e suas vantagens, torna-se possível afirmar que as ferramentas atuaram de forma robusta para uma análise bioinformática. O BLAST foi responsável por tornar a busca e anotação biológica rápida e objetiva, já o MAFFT, por permitir que o alinhamento múltiplo acontecesse de modo eficiente, e o ClustalW, por ter proporcionado instrumentalização ordenada do alinhamento, inserção dinâmicas evolutivas ideais para consulta da pesquisa.

Para isso o cálculo do sinal filogenético que pode ser definido como a covariação entre diferenças na medida de um atributo entre espécies e a distância filogenética que as separa (Blomberg; Jr; Ives, 2003).

Conforme Pagel (1999), um sinal filogenético elevado (por exemplo, próximo a 1 ou superior a 1 no caso da estatística K) indica que o atributo de interesse é amplamente compartilhado entre espécies evolutivamente próximas, refletindo uma alta correlação entre similaridade fenotípica e proximidade filogenética. Em contraste, um sinal filogenético baixo (próximo de 0) sugere que espécies filogeneticamente distantes apresentam menor semelhança nos atributos estudados, indicando uma fraca correspondência entre o traço e a história evolutiva das espécies.

Essa informação é relevante para a tomada de decisão sobre quais atributos escolher para calcular diversidade funcional, além do mais é útil em análises de diversidade filogenética baseadas em atributos conservados evolutivamente, afirma Webb (2000).

Desta forma o cálculo do sinal filogenético, a construção de árvores intermediárias e sua combinação em uma árvore final é um processo rigoroso, que possui o objetivo maximizar a precisão e robustez das relações evolutivas inferidas. A escolha do método de mapeamento de verossimilhança e das técnicas de consenso para a combinação das árvores foi fundamental para garantir que a árvore final represente de forma confiável o histórico evolutivo das sequências estudadas, minimizando o impacto de variações e ruídos nos dados de entrada.

Conforme explicado por Felsenstein (2006), PHYML Online é uma interface web para o algoritmo PHYML, por determinação, os dados de entrada consistem em somente em um arquivo de texto contendo um ou mais alinhamentos de sequências de DNA ou proteína no formato PHYLIP.

Esse software desempenha o processo de reconstrução, utilizando método de máxima verossimilhança, que é um método que calcula a probabilidade de os dados de uma árvore

específica representarem a realidade de acordo com um modelo evolutivo estabelecido. Este programa tem sido amplamente utilizado devido sua simplicidade, além de apresentar eficiência e exatidão (Delgado, 2020).

Tal algoritmo possui a capacidade de avaliar a confiabilidade internas utilizando *bootstrap*, mesmo com um grande conjuntos de dados, sendo possível graças a sua velocidade de otimização (Felsenstein, 1985). O *pipeline* tornou-se eficaz e consolidado aqui já explicitadas de maneira que favoreceu a averiguação da similaridade, variabilidade e identificação de relações filogenéticas entre as sequências, de forma otimizada às necessidades que regem o desafio diante do fato desta conclusão.

Utilizando essas estratégias será possível determinar o início e término do processo de tradução e posteriormente será possível checar a presença de polimorfismos genéticos associados aos genótipos e regiões geográficas, e assim estimar por contagem direta dos nucleotídeos visualizando o alinhamento e as sequências serão avaliadas pelo cálculo de qui-quadrado ou pelo teste exato de Fisher (Kearse *et al.*, 2012).

A escolha dos métodos filogenéticos Neighbor-Joining e Máxima Verossimilhança, tem como base as utilidades específicas de cada abordagem para a construção e interpretação de árvores evolutivas, alinhando-se aos objetivos do estudo em questão. O algoritmo de Neighbor-Joining apresentou melhores resultados em simulações e é computacionalmente o mais eficiente algoritmo de distâncias (Saitou; Nei, 1987), pode ser visto como um algoritmo ganancioso por querer otimizar uma árvore seguindo o critério da Evolução Mínima Equilibrada (BME) (Carriço, 2015).

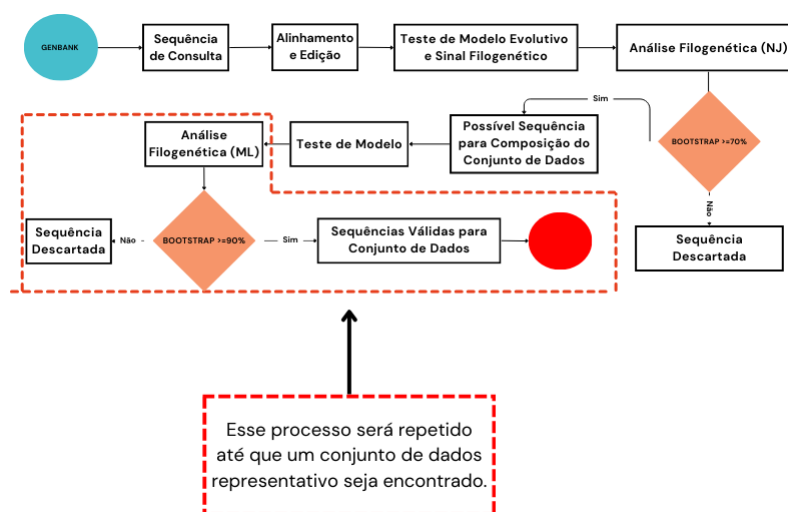
Segundo o critério, a árvore buscada é aquela que melhor explica a diferença observada nas folhas. Esse é o caso do algoritmo da verossimilhança máxima, sendo sempre possível obter a verossimilhança de cada proposta de topologia que se apresente, desta forma a verossimilhança é uma medida proporcional à probabilidade dos dados observados uma vez definidos (Silva, 2007).

O conjunto de cada segmento das sequências coletadas de OROV foram submetidas checadadas o sinal filogenético das mesmas. Para calcular o sinal filogenético, foi usado um método de mapeamento de verossimilhança (Strimmer; Haeseler, 1997). Esse cálculo foi realizado por um procedimento de três etapas, primeiro reconstruindo todas as árvores de máxima verossimilhança possíveis do quarteto, depois combinando repetidamente as árvores do quarteto em uma árvore geral e, finalmente, calculando o consenso da regra da maioria de todos os intermediários da árvores dando ao quarteto uma árvore intrigante (Schmidt *et al.*, 2002).

A determinação do modelo evolutivo foi realizada testando os alinhamentos através de teste da razão de verossimilhança hierárquico, utilizando o modelo NJ com o modelo de distâncias corrigidas LogDet. As sequências serão assim submetidas ao método de NJ e

com suporte estatístico para ramos internos nas árvores pelo método de 1.000 replicações, sequências com similaridade abaixo de 70% serão removidas do conjunto de sequências. Após essa acurácia as sequências restante serão submetidas a metodologia de ML serão inferidas com o modelo selecionado e parâmetros de substituição estimados e realizadas com método de 1.000 replicações. Serão selecionadas apenas sequências com 90% de agrupamentos de clados monofiléticos, sendo possível visualizar a estrutura do processo na Figura 18.

Figura 18 – Processo de mineração de sequências e construção do conjunto de dados utilizado na ferramenta de genotipagem do vírus do Oropouche.



Fonte: Elaborada pela autora, 2024.

Para o processo de identificação de SNPs foi realizada a partir dos sequenciamento genômico, a partir de algumas etapas de bioinformática, após a definição e sequenciamento foi feito a análise estatística, considerado uma técnica ampla, favorável e eficaz quando o objetivo é uma análise de variáveis qualitativas (Callegari-Jacques, 2009).

Na serventia do teste Qui-Quadrado foi preciso estabelecer duas hipóteses contraditórias acerca da pergunta a ser respondida ou testada, conforme Callegari-Jacques (2009) e Soares e Mendonça (2003). Essas suposições se definem na hipótese de nulidade (H_0), afirma a inexistência de diferenças entre os grupos testados, enquanto a hipótese alternativa (H_1), a qual rebate a hipótese nula H_0 e, normalmente, representa o que o pesquisador busca demonstrar efetivamente (Vieira, 1997).

A próxima etapa consiste na coleta e análise meticulosa de requisitos, definindo as funcionalidades essenciais da aplicação e seus objetivos específicos. Incorporando um método para a evolução consciente e sustentável de um sistema, afirma Anderson (2010), mantendo um máximo de tarefas em andamento simultaneamente, evitando a sobrecarga e melhorando o fluxo de desenvolvimento, limitando o trabalho em progresso e facilitando a identificar gargalos no processo (Kniberg; Skarin, 2010).

Levando em consideração momentos de adaptação visando que algum aspecto do processo deixa de ser útil ou diminui a transparência sobre o trabalho realizado, ele deve ser modificado (Sutherland; Schwaber, 2013).

Com o objetivo de identificar os genótipos virais, foi realizado um alinhamento múltiplo de sequências para cada espécie viral, utilizando um conjunto abrangente de dados. Em seguida, foi conduzida uma análise filogenética baseada no método de máxima verossimilhança, com 1000 réplicas de *Bootstrap*. O método *Bootstrap*, amplamente utilizado em diversas aplicações estatísticas, consiste em uma técnica de reamostragem que gera novos conjuntos de dados a partir do conjunto original (Tibshirani; Efron, 1993).

A programação será desenvolvida utilizando a tecnologia Java, amplamente adotada pela comunidade global de desenvolvedores de software e provedores de conteúdo na Internet. Essa tecnologia oferece aos usuários da Internet e da World Wide Web acesso a aplicativos seguros e independentes de plataforma, acessíveis a partir de qualquer local na rede (Arnold, 2005).

A aplicação terá como resultado em uma interface web que permitirá aos usuários realizar a classificação de suas sequências e visualizar os resultados, incluindo a atribuição de genótipos às sequências analisadas. O processo de análise será conduzido de forma individual para cada sequência, iniciando com a identificação utilizando o programa BLAST, seguida pelo alinhamento com o programa Multiple Sequence Alignment (ClustalW) e pela reconstrução filogenética por meio do programa Phylogenetic Analysis Using (PAUP*).

3.1 Validação e Experimentos: Onde a Teoria Encontra os Dados

Para a criação do ambiente de experimentação que foi baseado em uma infraestrutura flexível e vigorosa, estando capacitado para suportar os métodos analíticos e manipulação de grandes volumes de dados genômicos. Esse ambiente inclui ferramentas de software, linguagens de programação, sistemas operacionais, com o objetivo de assegurar a eficiência quanto a reprodutibilidade das análises.

A infraestrutura foi configurada de forma que a aplicação seja web, dessa forma será possível o acesso e gerenciamento dos recursos analíticos e computacionais. Essa aplicação web foi desenvolvida com o *framework Java Web Toolkit (JWT)* de forma que tenha uma escalabilidade, flexibilidade, e uma estrutura desenvolvida e conveniente que possibilitou os múltiplos usuários executarem e monitorem as análises simultaneamente.

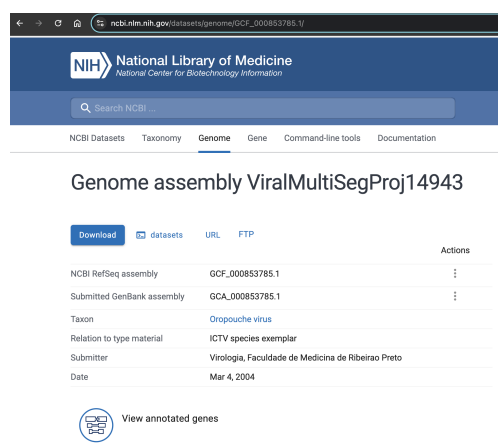
Essa infraestrutura baseada na web também garantirá a segurança e controle de acesso, além disso, a plataforma permitirá a integração e opções de armazenamento escalável para lidar com grandes volumes de dados bioinformáticos. Como já citado nessa pesquisa, o ambiente incluirá ferramentas especializadas para cada etapa do *pipeline*, garantindo precisão e eficiência em análises de dados complexos, seja para alinhamento de

sequências, chamada de variantes, montagem de genomas e a análise filogenética.

Dessa forma o ambiente de experimentação foi capaz de garantir que os dados fosse analisados com precisão e eficiência, proporcionando a aplicação com diferentes métodos em bioinformática com essa infraestrutura, o projeto pode ser reproduzível e acessível para futuros estudos. Os resultados obtidos por esse *pipeline* foram validados através de uma perspectiva que envolve a comparação entre bases de dados públicos e os experimentos, sendo possível através de alguns processos. Afim de garantir que a ferramenta web tenha atendido aos objetivos propostos e funcionou de maneira concreta, foi importante implementar uma ação afim de garantir que a validação tenha sido feita de forma abrangente.

Para garantia da precisão e eficácia, foi fundamental testar o desempenho com a utilização de dados já conhecidos, no caso desta pesquisa foi envolvido a seleção de sequências genômicas com suas classificações de espécie e linhagem bem estabelecidas. Para o feito da validação foi feita a escolha de sequências de vírus Oropouche cujas características taxonômicas, como espécie e linhagem, estão bem documentas na base de dados públicos do GenBank conforme a Figura 19, essa sequência serviu como um padrão de referência para avaliar a precisão da ferramenta desenvolvida.

Figura 19 – Escolha da sequência do vírus do Oropouche na base de dados GenBank.



Fonte: Elaborada pela autora, 2024.

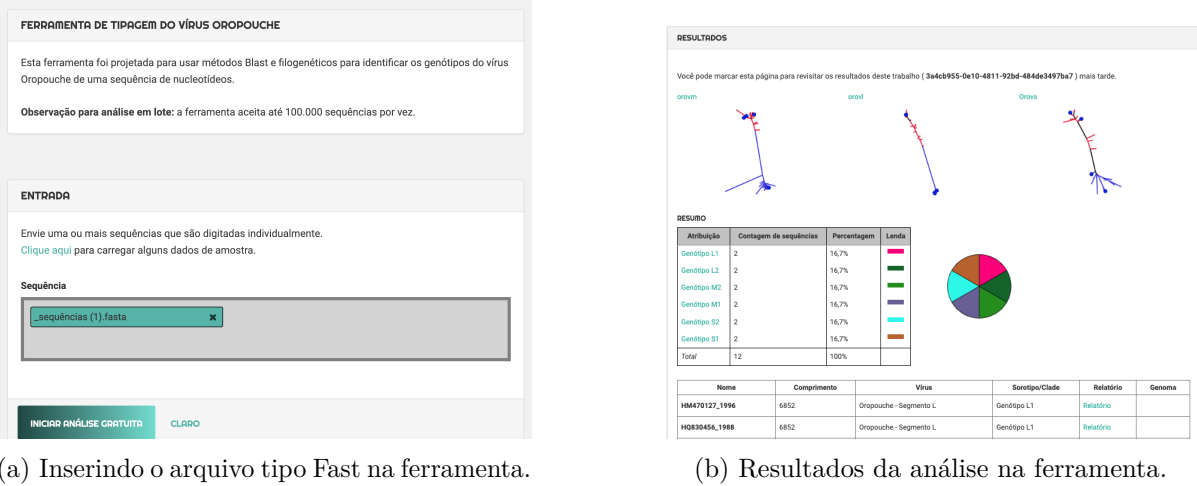
Após o download dessa sequência em um arquivo Fasta ¹, a mesma foi executada na ferramenta, servindo como padrão de referência para avaliar a precisão da ferramenta desenvolvida.

Após a seleção das sequências realizada no processo anterior, a ferramenta foi executada com essas amostras específicas, o principal objetivo foi que ferramenta classificou

¹ Formato baseado em texto utilizado para representar sequências de nucleotídeos ou peptídeos, em que os nucleotídeos ou aminoácidos são representados por códigos de uma única letra (Worley, 2003; Korf; Yandell; Bedell, 2003).

corretamente as sequências de vírus Oropouche, atribuindo-lhes a espécie e linhagem correta com base nas informações contidas nos dados de entrada conforme a Figura 20.

Figura 20 – Processo de verificação após a seleção da sequência.



(a) Inserindo o arquivo tipo Fast na ferramenta.

(b) Resultados da análise na ferramenta.

Fonte: Elaborada pela autora, 2024.

Em seguida foi possível visualizar os resultados obtidos, podendo ser analisados e sua eficácia e a precisão do *pipeline* desenvolvido avaliados, essa etapa é importante, afim de confirmar que a ferramenta atende ao objetivo proposto, classificando de forma correta a sequência de acordo com a linhagem e características do vírus. A concordância dos resultados da ferramenta e a classificações previamente documentada no Gbank reforça a precisão do *pipeline*, gerando uma confiança para a análises genômicas do vírus Oropouche.

Diante disso, a análise com dados de referência é um passo importante afim de assegurar que a ferramenta seja confiável, permitindo que a sua aplicação gere confiança em estudos futuros e na classificação de variantes do vírus Oropouche. Para validar a plataforma e verificar como a mesma se comporta diante das variações nos dados de entrada, foi realizada uma abordagem eficaz para essa avaliação é a introdução de ruído ou variações artificiais nas sequências genéticas, como inserções, deleções ou substituições, dessa forma a simulação de condições de dados imperfeitos ou ruídos comuns em experimentos biológicos, permitindo verificar se a ferramenta consegue lidar com esses desafios sem comprometer sua precisão e eficácia.

Para o processo de comprovação foi necessário realizar por meio de introdução de variações artificiais nas sequências genéticas utilizadas no processo de classificação, sendo eles:

- a) adição de nucleotídeos extras em locais aleatórios da sequência;
- b) remoção de nucleotídeos de posições específicas, criando lacunas ou regiões faltantes nas sequências;

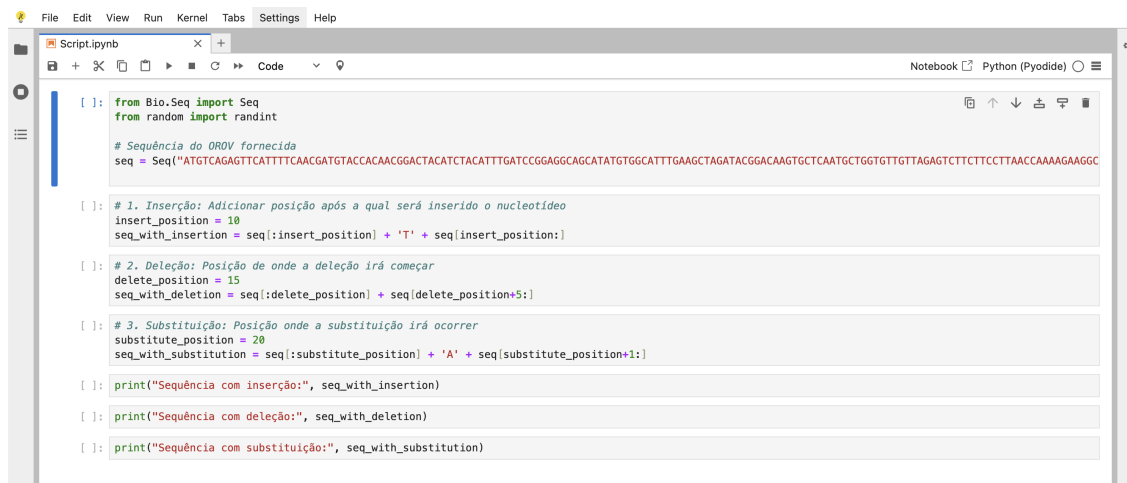
- c) alteração de nucleotídeos em posições específicas da sequência, simulando mutações genéticas naturais.

A ferramenta foi aplicada novamente para classificar a sequência modificada, o seu desempenho foi analisado em relação à sua capacidade de manter a precisão da classificação, apesar das alterações nos dados de entrada.

A criação de um *script*² utilizando a biblioteca Biopython (Cock *et al.*, 2009) para introduzir inserções, deleções e substituições em uma sequência genética e a biblioteca Random para gerar posições aleatórias para as inserções e deleções (Arnold *et al.*, 1995), teve como objetivo manipular as sequências de forma controlada, gerando variações artificiais que simulem erros ou mutações genéticas, para que o teste de robustez e a capacidade da ferramenta.

Para execução do *script* foi usado o Jupyter Notebook (What... , 2015) e a linguagem de programação Python (Borges, 2014), conforme a Figura 21 é possível visualizar o *script* criado para o objetivo desta etapa da pesquisa, a criação desse *script* foi uma maneira eficaz de para variações artificiais em sequências genômicas afim de testar a robustez da ferramenta, a partir da de inserções, deleções e substituições, foi possível simular dados imperfeitos e observar como os algoritmos de análise lidam com essas modificações.

Figura 21 – Script para inserção, deleção e substituição.



```
[ ]: from Bio.Seq import Seq
from random import randint

# Sequência do ORV fornecida
seq = Seq("ATGTCAGAGTTCTTTTCAACGATGTACCAACGGACTACTACATTTGATCGGAGGCAGCATATGTGGCATTGGAAGCTAGATACGGACAAGTGCTCAATGCTGGTGTGTAGAGTCTTCTCTTCAACCAAGGAAGGC")

[ ]: # 1. Inserção: Adicionar posição após a qual será inserido o nucleotídeo
insert_position = 10
seq_with_insertion = seq[:insert_position] + 'T' + seq[insert_position:]

[ ]: # 2. Deleção: Posição de onde a deleção irá começar
delete_position = 15
seq_with_deletion = seq[:delete_position] + seq[delete_position+5:]

[ ]: # 3. Substituição: Posição onde a substituição irá ocorrer
substitute_position = 20
seq_with_substitution = seq[:substitute_position] + 'A' + seq[substitute_position+1:]

[ ]: print("Sequência com inserção:", seq_with_insertion)

[ ]: print("Sequência com deleção:", seq_with_deletion)

[ ]: print("Sequência com substituição:", seq_with_substitution)
```

Fonte: Elaborada pela autora, 2024.

Esse processo é essencial para garantir que a ferramenta seja capazes de lidar com a variabilidade genética encontrada em dados reais, após a execução do *script*, a Figura 22 e suas saídas equivalentes foi possível inserir cada sequência modificada, seja por inserção, deleção e substituição, na plataforma, afim de observar o comportamento da mesma.

² Texto com uma série de instruções escritas.

quantitativas específicas e suas respectivas fórmulas. Durante o processo, foram coletados dados que permitiram avaliar e validar o desempenho da ferramenta em termos de precisão, sensibilidade, especificidade e acurácia. Esse procedimento garantiu que a ferramenta atendesse aos requisitos técnicos estipulados e se mantivesse competitiva no contexto em que será aplicada.

As métricas utilizadas estão descritas a seguir:

- a) **Precisão**: Mede a proporção de resultados positivos identificados pela ferramenta que são verdadeiramente positivos:

$$\text{Precisão} = \frac{VP}{VP + FP}$$

Onde:

- **VP** (*Verdadeiros Positivos*): Casos corretamente identificados como positivos.
 - **FP** (*Falsos Positivos*): Casos incorretamente identificados como positivos.
- b) **Sensibilidade (SENS)**: Avalia a capacidade da ferramenta de identificar corretamente todos os casos positivos, independentemente de erros de classificação:

$$\text{SENS} = \frac{VP}{VP + FN}$$

Onde:

- **FN** (*Falsos Negativos*): Casos positivos que não foram identificados pela ferramenta.
- c) **Especificidade (ESPEC)**: Mede a capacidade da ferramenta de evitar falsos alarmes, identificando corretamente os casos negativos:

$$\text{ESPEC} = \frac{VN}{VN + FP}$$

Onde:

- **VN** (*Verdadeiros Negativos*): Casos corretamente classificados como negativos.
- d) **Acurácia (ACUR)**: Representa a proporção total de classificações corretas realizadas pela ferramenta, abrangendo casos positivos e negativos:

$$\text{ACUR} = \frac{VP + VN}{VP + FN + VN + FP}$$

3.2 Da Ideia ao Código: Os Estágios Essenciais do Desenvolvimento da Oropouche Virus Typing Tool

O grupo de profissionais e pesquisadores envolvidos no desenvolvimento da ferramenta Oropouche Virus Typing Tool é composto por diversos colaboradores que trabalham de forma cooperativa com o objetivo de alcançar um propósito comum. Para o desenvolvimento dessa ferramenta, foi necessário seguir alguns passos, compreender e aforas algumas técnicas.

Para a construção do *pipeline*, foram considerados dois elementos de classificação. O primeiro refere-se à permissão das espécies, a pesquisa de classificação submete uma sequência de consulta à análise BLAST em relação ao conjunto de sequências de referência (Strimmer; Haeseler, 1997), uma consulta é encarregado ao BLAST gerando uma correspondência com uma pontuação que ultrapassa um limite predefinido. O segundo componente de classificação diz respeito à construção de uma árvore filogenética por meio do método NJ, o que possibilitou a atribuições em nível de genótipo e/ou subtipo.

Posteriormente o alinhamento do perfil foi executado através do software ClustalW (Thompson; Higgins; Gibson, 1994), de maneira que a sequência de consulta seja entrelaçada ao alinhamento já existente de sequências de referência, é construído uma árvore filogenética utilizando o método NJ e com a técnica de *bootstrap*, foi construída a árvore com a utilização do software PAUP* (Salemi; Vandamme, 2003).

No momento da segunda classificação, a sequência é atribuída a um genótipo específico, caso a sequência se agrupe monofileticamente com o clado de genótipo que apresenta suporte *bootstrap* superior a 70%, ela é considerada válida, de o suporte *bootstrap* for inferior a 70%, o genótipo será classificado como não válido.

O código-fonte da ferramenta foi desenvolvido na linguagem de programação Java, utilizando o *framework* de arquitetura de software JWT [<https://www.webtoolkit.eu/jwt>] (Web Toolkit, 2024). Possui um conjunto de bibliotecas escritas em linguagem Java que visam aumentar a produtividade da construção de aplicações como por exemplo, garantindo a proteção contra o uso indevido, evitar ataques dentre outros proveitos que proporciona uma interface de desenvolvimento robusta e eficiente, juntamente com a ferramenta Regagenotype³ aliadas a funcionalidades em Java. Essa abordagem resulta em uma ferramenta suficientemente genérica para realizar análises em uma ampla variedade de organismos, capaz de atender às demandas específicas do projeto.

³ Realiza a reconstrução filogenética por meio da integração de diversas ferramentas, como ClustalW, BLAST e PAUP*

4 RESULTADOS E DISCUSSÕES

A ferramenta de tipagem do vírus Oropouche demonstrou-se eficaz na identificação de genótipos específicos a partir de sequências de nucleotídeos. Sua interface web intuitiva permite aos usuários realizar a classificação e visualizar resultados de forma prática. Além disso, a ferramenta suporta a análise simultânea de até 2.000 sequências, otimizando o processamento em estudos de larga escala.

A validação da ferramenta foi realizada por meio da comparação de seus resultados com sequências de referência de bases de dados consolidadas. Esse processo demonstrou a consistência dos agrupamentos filogenéticos gerados e a exatidão na identificação dos genótipos, conforme descrito na 3.1. Essa abordagem confirma a capacidade da ferramenta de replicar padrões filogenéticos previamente descritos na literatura, conferindo maior credibilidade ao método (Altschul *et al.*, 1997)

A validação utilizando bancos de dados conhecidos também possibilita o ajuste contínuo da ferramenta. Isso inclui o refinamento de algoritmos e parâmetros de classificação, garantindo maior precisão e ampliando sua aplicabilidade a diferentes tipos de sequências e condições epidemiológicas.

Para avaliar a aplicabilidade da ferramenta em diferentes contextos globais, é essencial considerar a representatividade das sequências em relação à diversidade geográfica, temporal e genética. O OROV, membro do sorogrupo Simbu, é caracterizado por sua grande diversidade genética e ampla distribuição geográfica (Pimentel, 2019). A diversidade geográfica, por exemplo, permite que a ferramenta responda a diferentes localidades, capturando variações regionais do vírus MARTINS e PEDROSO (2021). Uma amostragem representativa facilita a inclusão de genótipos endêmicos, garantindo maior precisão na identificação de variantes locais.

A diversidade temporal é igualmente relevante, pois o vírus pode sofrer alterações ao longo do tempo, seja por mutações naturais ou pressões seletivas ambientais e epidemiológicas, uma amostragem que inclua sequências com diferentes períodos permite que a ferramenta detecte e identifique possíveis mudanças evolutivas do vírus, assegurando que a mesma mantenha sua precisão ao longo do tempo.

Por último, a diversidade genética, uma vez que permite que a ferramenta detecte variações dentro de genótipos do vírus Oropouche, com uma representatividade genética ampla possibilita que a ferramenta capte uma gama maior de variações, fornecendo maior sensibilidade na tipagem viral, esse contraste genético robusta entre as sequências de entrada assegura que a ferramenta seja aplicável a diferentes cepas do vírus, independentemente das mutações ou variantes regionais.

A representatividade geográfica, temporal e genética das sequências é indispensável para garantir que a ferramenta consiga genotipar o vírus Oropouche em diferentes cenários epidemiológicos com alta precisão, com essa representatividade só a resposta eficaz a surtos locais não será a única, mas irá contribuir para uma vigilância global, atestando que a ferramenta mantenha sua relevância em contextos de rápida evolução viral.

No desenvolvimento da ferramenta, foram coletadas 2.400 sequências de nucleotídeos a partir de bancos de dados públicos, abrangendo diferentes estruturas genéticas e representações geográficas do vírus Oropouche. Contudo, apenas 31 sequências do segmento S, 35 do segmento M e 33 do segmento L foram utilizadas, representando aproximadamente 4,2% do total. Essas sequências foram escolhidas por apresentarem uma diversidade genética capaz de representar o conjunto completo. No entanto, embora esse conjunto de dados seja diversificado, ele pode não incluir adequadamente genótipos raros ou variantes altamente divergentes, o que pode comprometer a capacidade de classificação em surtos que envolvam essas variantes menos comuns.

Além do mais, os resultados dos testes detalhados em 3.1 indicaram como a ferramenta responde a diferentes tipos de ruído nos dados de entrada, o que é de suma importância para a confiabilidade da análise. Observamos que a ferramenta conseguiu identificar e processar corretamente as variações introduzidas, o que reforça sua capacidade de realizar classificações precisas mesmo em condições adversas. Em resumo, esses testes demonstram que a ferramenta possui uma robustez satisfatória. A validação da ferramenta com sequências genômicas contendo erros controlados é um passo fundamental para garantir sua aplicabilidade em estudos com dados reais, assegurando que ela oferece resultados confiáveis e consistentes no estudo do vírus Oropouche.

A análise crítica dos resultados obtidos permite uma avaliação detalhada da precisão e das limitações da ferramenta de tipagem desenvolvida, embora a ferramenta tenha demonstrado eficácia na identificação dos genótipos predominantes do vírus Oropouche, em relação à capacidade de classificação de genótipos raros ou altamente divergentes. Essas dificuldades refletem uma limitação comum em ferramentas de tipagem baseadas em bancos de dados de sequências, que podem estar sujeitos a um viés de representatividade.

Além disso, pode ocorrer da precisão da classificação pode ser influenciada por variações na qualidade das sequências de entrada, isso reforça a importância de utilizar sequências de alta qualidade e de ampliar continuamente os bancos de dados com novas amostras genéticas, o que permitiria uma representação mais robusta e precisa de genótipos raros ou divergentes.

Uma interface web foi desenvolvida utilizando a tecnologia Java, com o objetivo de atribuir genótipos às sequências fornecidas pelos usuários. A ferramenta realiza uma análise detalhada das sequências, iniciando com a identificação das mesmas por meio do programa BLAST (McGinnis; Madden, 2004). Posteriormente, as sequências são submetidas a um

alinhamento utilizando o programa ClustalW (Thompson; Higgins; Gibson, 1994). Por fim, a reconstrução filogenética é efetuada através do programa PAUP* (Salemi; Vandamme, 2003). A classificação genotípica é determinada quando as sequências do usuário são agrupadas filogeneticamente com um valor de *bootstrap* igual ou superior a 70%.

O método *bootstrap* é usado frequentemente para gerar valores de suporte em análises filogenéticas, oferece um valor probabilístico de confiança para cada agrupamento. Segundo (Felsenstein, 1985) esses valores de entre 70-90% são geralmente considerados moderadamente confiáveis, enquanto valores que estão acima de 90% indicam agrupamentos altamente confiáveis. Diante disso, a escolha de 70% como limite mínimo é apropriada para identificar ramos com suporte moderado e, ao mesmo tempo, evitar a exclusão de agrupamentos informativos, mas que não alcançam um nível de suporte máximo.

Os valores permitem que a análise capture os agrupamentos mais robustos quanto aqueles que ainda apresentam significado biológico, mesmo com suporte moderados metodológicos como sugere Felsenstein e Kishino (1993), sugerem que valores de *bootstrap* inferiores a 70% têm maior probabilidade de indicar agrupamentos instáveis, dessa forma os agrupamentos que podem não ser recuperados consistentemente em análises repetidas ou em diferentes subconjuntos de dados.

A escolha de 70% como um valor de corte apropriado para evitar falsos positivos e garantir que os agrupamentos selecionados possuam relevância estatística. Dessa forma, o uso do limite de 70% em estudos de filogenia tem se estabelecido como prática comum e é amplamente aceito como um indicador de robustez para o método NJ.

4.1 Interface e Usuário

Nesta seção, foram apresentadas as principais características da ferramenta de tipagem do vírus Oropuche, com enfoque em aspectos relacionados ao design de sua interface gráfica, aos mecanismos de visualização e às funcionalidades oferecidas para uso por pesquisadores.

A interface web da ferramenta foi projetada com o objetivo de proporcionar uma experiência de forma intuitiva e acessível, permitindo que os usuários possam realizar a análise filogenética de forma eficiente. De acesso livre, proporciona aos usuários uma plataforma para realizar a classificação de suas sequências e visualizar os resultados, conforme ilustrado na Figura 23.

Figura 23 – A interface web da ferramenta de genotipagem do vírus do Oropouche.

The screenshot displays the web interface of the 'OROPOUCHE VIRUS TYPING TOOL'. At the top, there is a navigation bar with links: 'ENVIAR TRABALHO', 'MONITORAR TRABALHO', 'DOCUMENTAÇÃO', 'PRÊMIO', and a 'CONECTE-SE' button. Below this, the tool's name is prominently displayed. The main content area is titled 'FERRAMENTA DE TIPAGEM DO VÍRUS OROPOUCHE' and includes a version indicator 'Versão 1'. A descriptive text explains the tool's purpose: using BLAST and phylogenetic methods to identify Oropouche virus genotypes from nucleotide sequences. It also notes a capacity of up to 100,000 sequences per batch. The 'ENTRADA' (Input) section prompts users to upload sequences individually or via a file, with a 'CLIQUE AQUI' link for sample data. A large text area is provided for pasting FASTA sequences. Below the input section, there are buttons for 'INICIAR ANÁLISE GRATUITA' and 'CLARO', along with a link to register for premium features. The bottom section, 'REVISE OS RESULTADOS DE UMA EXECUÇÃO ANTERIOR', features an input field for the 'ID do trabalho' and a corresponding 'ID' button.

Fonte: Elaborada pela autora, 2024.

Para obter um fluxo de navegação contínuo, a interface foi estruturada para ser simples e direta, de forma que a organização facilite a interação do usuário, desde o carregamento dos dados até a interpretação dos resultados. O primeiro passo do fluxo é o processo de carregamento de dados, sendo uma das funcionalidades centrais da interface.

Ao acessar a ferramenta web, o usuário se depara com a página onde pode facilmente carregar suas sequências, a ferramenta aceita arquivos no formato FASTA na Figura 24a, que é amplamente utilizado para sequências genéticas, e fornece um botão de "iniciar análise gratuita", permite selecionar o arquivo diretamente do dispositivo, a interface também oferece a opção de arrastar e soltar os arquivos, proporcionando uma interação mais prática, caso o usuário não tenha nenhuma sequência, é possível selecionar dados de amostra na própria ferramenta na Figura 24b.

A ferramenta oferece diferentes formatos de exportação a fim de garantir que

Figura 24 – Processo inserção de dados.

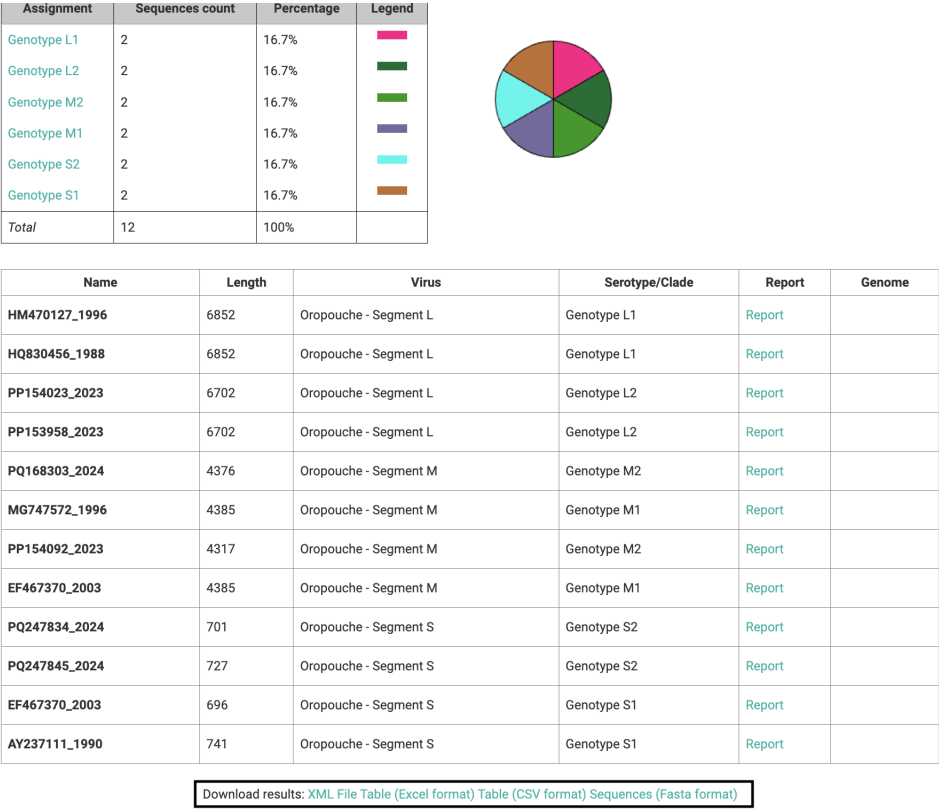


(a) Inserindo o arquivo na ferramenta. (b) Inserindo dados de amostra.

Fonte: Elaborada pela autora, 2024.

os resultados da análise sejam acessíveis a uma ampla gama de usuários, permitindo flexibilidade e adaptabilidade no uso dos dados gerados. Esses formatos incluem XML, tabelas como Excel e CSV, e sequências, em formato FASTA, os resultados podem ser exportados em diversos formatos de arquivo conforme ilustrado na Figura 25.

Figura 25 – Formatos de exportações disponíveis.



Fonte: Elaborada pela autora, 2024.

Por exemplo na exportação do arquivo em Excel na Figura 26 apresenta informações detalhadas, incluindo o nome da sequência, seu comprimento, o vírus e o genótipo atribuídos, uma representação gráfica da posição da sequência no genoma viral, e uma seção dedicada à análise filogenética.

Figura 26 – Exemplo de exportação em formato Excel na interface web da ferramenta de genotipagem do vírus do Oropouche.

PQ247845_2024	727	Oropouche - Segment S	Genotype S2	Report	
EF467370_2003	696	Oropouche - Segment S	Genotype S1	Report	
AY237111_1990	741	Oropouche - Segment S	Genotype S1	Report	

Download results: [XML File Table \(Excel format\)](#) [Table \(CSV format\)](#) [Sequences \(Fasta format\)](#)

name	assignment	length	begin	end	species	species-score	type	type support	subtype	subtype support
HM470127_1996		6852	0	6852	Oropouche - Segment L		Genotype L1	100.0		
HQ830456_1988		6852	0	6852	Oropouche - Segment L		Genotype L1	100.0		
PP154023_2023		6702	0	6702	Oropouche - Segment L		Genotype L2	100.0		
PP153958_2023		6702	0	6702	Oropouche - Segment L		Genotype L2	100.0		
PQ168303_2024		4376	-4	4372	Oropouche - Segment M		Genotype M2	100.0		
MG747572_1996		4385	0	4385	Oropouche - Segment M		Genotype M1	100.0		
PP154092_2023		4317	0	4317	Oropouche - Segment M		Genotype M2	100.0		
EF467370_2003		4385	0	4385	Oropouche - Segment M		Genotype M1	100.0		
PQ247834_2024		701	35	736	Oropouche - Segment S		Genotype S2	100.0		
PQ247845_2024		727	9	736	Oropouche - Segment S		Genotype S2	100.0		
EF467370_2003		696	44	740	Oropouche - Segment S		Genotype S1	100.0		
AY237111_1990		741	0	741	Oropouche - Segment S		Genotype S1	100.0		

Fonte: Elaborada pela autora, 2024.

Os formatos de tabela, como Excel e CSV Figura 26, são altamente acessíveis e amplamente utilizados possui uma visualização clara e organizada dos dados, já formato FASTA é indispensável para pesquisadores e profissionais envolvidos com a análise detalhada e a construção de novas análises a partir das sequências processadas, reutilização dos dados em contextos de pesquisa avançada.

O formato XML Figura 27 é amplamente utilizado para o intercâmbio de dados entre sistemas e pode ser particularmente útil para pesquisadores que desejam integrar os resultados da análise com outras ferramentas ou bancos de dados, por ser um formato estruturado e hierárquico, o XML permite que os resultados sejam facilmente manipulados e processados por programas de análise de dados, como linguagens de programação (Python, R) ou sistemas de gerenciamento de bancos de dados.

Figura 27 – Exemplo de exportação em formato XML na interface web da ferramenta de genotipagem do vírus do Oropouche.

```

rov 1
.TGTCAGAGTTCAATTTTCAACGATGTACCAACACGGACTACATCTACATTTGATCCGGAGGAGCAGCATATGTGGCAITTTGAAGCTAGATACGGACAAGTGCTCAATGCTGGTGTGTAGAGTCTTCTTCCT
696 yes 1.0014389 orov Oropouche Virus 1104.0 1104.0 100.14388 100.14388 OROV false orov Oropouche Virus OROV -1 Unknown 0 MG747503_G3 1 Genotype 1 0.0 r9944254.nex r3364196.log r3908958.tre
/assigned Could not assign Not supported by phylogenetic analysis
.TGTCAGAGTTCAATTTTCAACGATGTACCAACACGGACTACATCTACATTTGATCCGGAGGAGCAGCATATGTGGCAITTTGAAGCTAGATACGGACAAGTGCTCAATGCTGGTGTGTAGAGTCTTCTTCCT
696 yes 1.0014389 orov Oropouche Virus 1104.0 1104.0 100.14388 100.14388 OROV false orov Oropouche Virus OROV -1 Unknown 0 MG747509_G4 4 Genotype IV 100.0 4 Genotype IV 100.0 r5917256.nex
5139161.log r7661708.tre 4 Genotype IV 100.0 Supported with phylogenetic analysis and bootstrap 100.0 (>= 70.0)
.TGTCAGAGTTCAATTTTCAACGATGTACCAACACGGACTACATCTACATTTGATCCGGAGGAGCAGCATATGTGGCAITTTGAAGCTAGATACGGACAAGTGCTCAATGCTGGTGTGTAGAGTCTTCTTCCT
696 yes 1.0014389 orov Oropouche Virus 1104.0 1104.0 100.14388 100.14388 OROV false orov Oropouche Virus OROV -1 Unknown 0 MG747515_G1 1 Genotype 1 99.0 1 Genotype 1 99.0 r7331702.nex r3833497
9368665.tre 1 Genotype 1 99.0 Supported with phylogenetic analysis and bootstrap 99.0 (>= 70.0)
.TGTCAGAGTTCTTTTCAATGATGTACCAACACGGACTACATCTACATTTGATCCGGAGGAGCAGCATATGTGGCAITTTGAAGCTAGATACGGACAAGTGCTCAATGCTGGTGTGTAGAGTCTTCTTCCT
696 yes 1.0014389 orov Oropouche Virus 1104.0 1104.0 100.14388 100.14388 OROV false orov Oropouche Virus OROV -1 Unknown 0 MG747551_G2 1 Genotype 1 0.0 r5388912.nex r8480822.log r9612944.tre
/assigned Could not assign Not supported by phylogenetic analysis

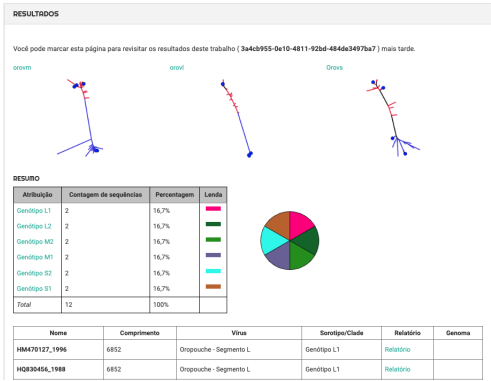
```

Fonte: Elaborada pela autora, 2024.

A flexibilidade oferecida pelos diferentes formatos de exportação é fundamental para garantir que os resultados da ferramenta de tipagem do vírus Oropouche sejam acessíveis e utilizáveis por diversos tipos de usuários, desde aqueles mais técnicos, como bioinformatas, até profissionais de saúde pública e pesquisadores de outras áreas. Cada formato atende a necessidades específicas, permitindo que os usuários escolham o mais adequado ao seu nível de experiência e às suas necessidades analíticas

O relatório de tipagem, mostrado na Figura 28 apresenta informações abrangentes, incluindo o nome da sequência de consulta, o comprimento em nucleotídeos, a localização da sequência no genoma viral, a atribuição de espécie e a designação de genótipo. Adicionalmente, é fornecido um relatório detalhado referente à análise filogenética que fundamentou a classificação realizada.

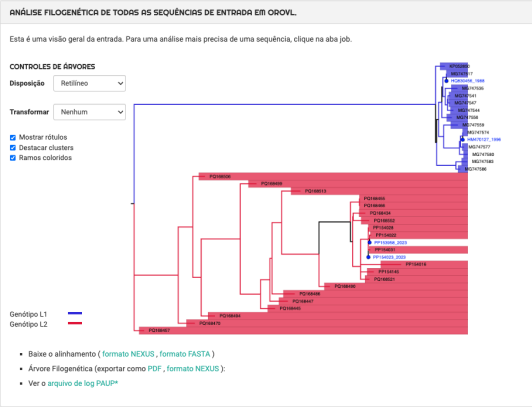
Figura 28 – Exemplo da classificação da sequências e visualização dos resultados na ferramenta de genotipagem do vírus do Oropouche.



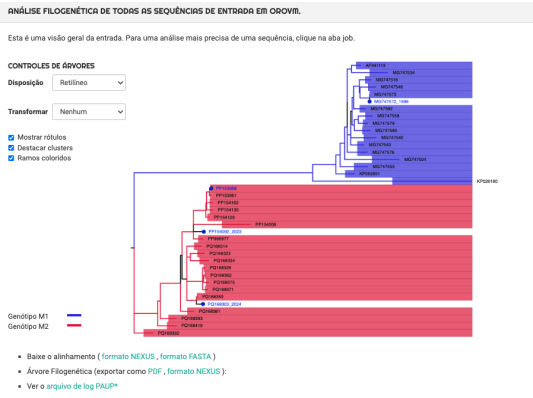
Fonte: Elaborada pela autora, 2024.

Nesse ponto do fluxo exibe o alinhamento e a filogenia construída, com a sequência de consulta posicionada no topo da árvore filogenética para o segmento L Figura 29a, segmento M Figura 29b e segmento S Figura 29c.

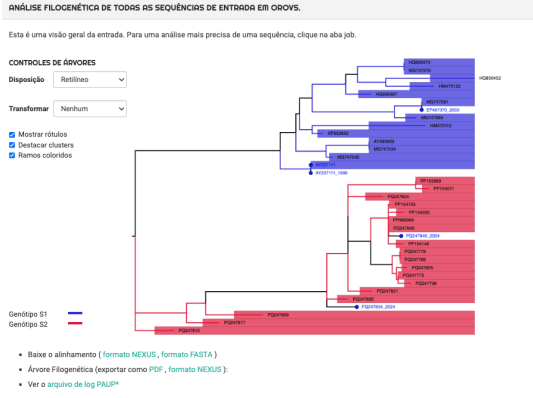
Figura 29 – Análise filogenética de todas as sequências de entrada.



(a) Segmento L.



(b) Segmento M.

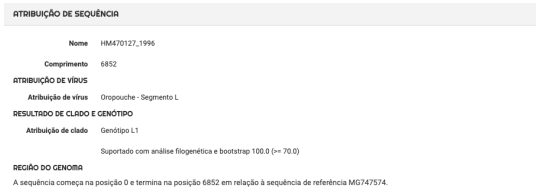


(c) Segmento S.

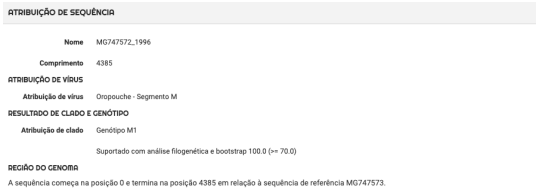
Fonte: Elaborada pela autora, 2024.

Tendo em vista que a construção de árvores filogenéticas é um componente importante para a análise de dados genéticos, permitindo que seja identificado relações evolutivas entre diferentes sequências de nucleotídeos é possível obter uma visualização através das atribuição de sequência para o segmento L Figura 30a, segmento M Figura 30b e segmento S Figura 30c.

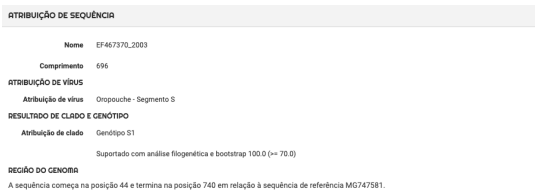
Figura 30 – Atribuição de sequência.



(a) Segmento L.



(b) Segmento M.

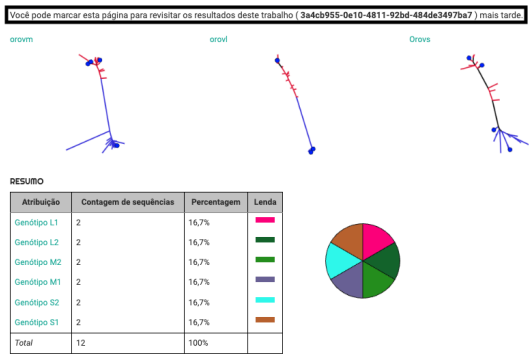


(c) Segmento S.

Fonte: Elaborada pela autora, 2024.

É possível a visualização futura dos resultados na Figura 31b pelo id Figura 31a disponibilizado pela plataforma, a estrutura do banco de dados foi otimizada para acomodar grandes volumes de dados e facilitar a integração de novas sequências à medida que são geradas.

Figura 31 – Processo de recuperar Id e execução de trabalhos anteriores.



(a) Id do trabalho disponibilizado pela plataforma.



(b) Carregando execução anterior.

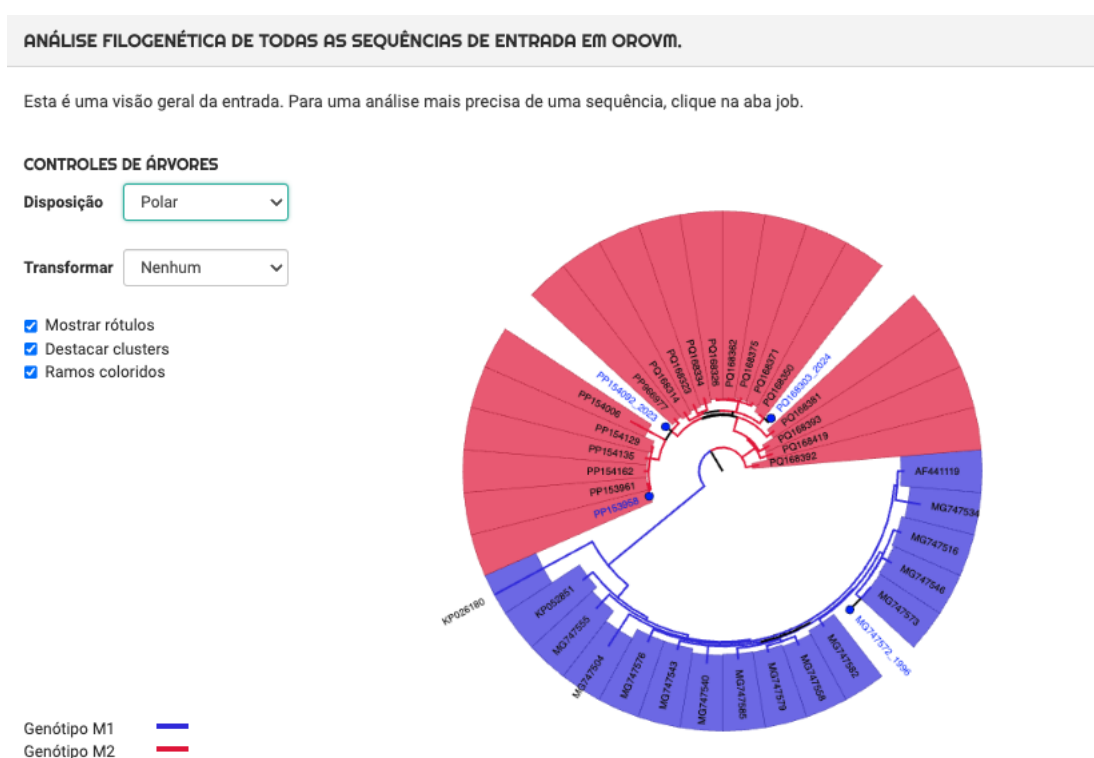
Fonte: Elaborada pela autora, 2024.

A ferramenta utiliza algoritmos de construção de árvores filogenéticas, como o método Neighbor-Joining, Máxima Verossimilhança e Bayesiano, cada um com características que influenciam a precisão, robustez e sensibilidade dos resultados, as vantagens e motivos para utilização dos métodos que foram discutidos na seção 3.

No momento da comparação, é possível perceber que o NJ é o mais rápido e computacionalmente acessível, porém com limitações em termos de robustez quando se lida com dados divergentes ou incompletos, o ML, capaz de lidar com modelos evolutivos complexos, embora mais preciso, apresenta desafios computacionais, particularmente em grandes bases de dados, em outro ponto, o método Bayesiano oferece uma robustez superior ao incorporar a incerteza nas estimativas filogenéticas, contudo é o mais exigente em termos de tempo computacional.

Na ferramenta a interface web permitirá que os usuários realizem as análises do vírus OROV, projetadas para fornecer informações detalhadas e relevantes no contexto do estudo. Entre as análises disponíveis, alguns pontos como a análise filogenética se destaca, será possível a visualização da árvores filogenéticas para investigar as relações evolutivas da sequência do vírus, sendo possível controlar a disposição da árvore podendo ser, retilíneo, polar, demonstrado na Figura 32, ou radial, entre as transformações proporcional, comprimento igual e cladograma.

Figura 32 – Interface web e suas disposições disponíveis.



A sensibilidade da ferramenta ao alinhar sequências altamente divergentes ou incompletas por sequências de nucleotídeos mal alinhadas ou com grandes lacunas podem afetar a qualidade e a robustez das árvores filogenéticas geradas, essas sequências divergentes podem resultar em ruído filogenético, o que dificulta a construção de uma árvore precisa, além desses pontos, as sequências incompletas, com gaps ou nucleotídeos faltantes, podem causar problemas na definição das relações evolutivas, levando a agrupamentos errôneos.

Dessa forma, a quantidade de dados e a qualidade do alinhamento são fatores cruciais para garantir a robustez das árvores filogenéticas, a ferramenta deve incorporar estratégias de qualidade de alinhamento e validação de dados, como a verificação de sequências inconsistentes ou a remoção de sequências com alta taxa de erro, para minimizar o impacto de dados problemáticos na construção das árvores.

4.2 *Benchmarking* da ferramenta

Com base na aplicação do *benchmarking*, destrinchada na 3.1, foi possível validar a ferramenta utilizando as métricas propostas. Os resultados obtidos, destacados na Tabela 1, Tabela 2, Tabela 3, Tabela 4, Tabela 5 e Tabela 6, oferecem uma visão clara sobre o desempenho da ferramenta, permitindo a tomada de decisões informadas quanto a ajustes, melhorias ou futuras implementações. Tal abordagem sistemática garante maior confiança no processo de avaliação e reforça a capacidade da ferramenta de atender aos padrões esperados.

Para essa análise, foram utilizados dados de validação abrangentes, coletados a partir de 2.478 sequências do vírus Oropouche, distribuídas entre os segmentos L (757 sequências), M (789 sequências) e S (932 sequências). Para o segmento L, foram analisadas 758 sequências com o objetivo de avaliar a eficiência da ferramenta em genotipar corretamente o vírus.

Dessas sequências do segmento L, 71 foram classificadas como pertencentes ao genótipo L1 e 673 como pertencentes ao genótipo L2. Outras 9 sequências não puderam ser genotipadas automaticamente pela ferramenta e, portanto, foram classificadas posteriormente por meio de análise filogenética convencional, utilizando o método de máxima verossimilhança, usando o software iqtree na versão 2 com o modelo de substituição GTR+G4+F, conforme resumido na Tabela 1.

A análise foi repetida desconsiderando as 9 sequências do segmento L que não puderam ser genotipadas automaticamente pela ferramenta na análise inicial. Nessa nova abordagem, as 71 sequências previamente classificadas como pertencentes ao genótipo L1 e as 673 sequências como pertencentes ao genótipo L2 foram mantidas, mas sem a inclusão das 9 sequências classificadas manualmente.

Com essa modificação, observou-se que os valores de sensibilidade, especificidade e

Tabela 1 – Segmento L considerando sequências com baixa qualidade

Classe	VP	VN	FP	FN	SENS	ESPEC	ACUR
L2	672	84	0	1	100%	100%	100%
L1	75	672	10	0	100%	99%	99%
Unassigned	0	748	0	9	0%	100%	99%

Fonte: Autora, 2024

acurácia aumentaram para 100%, indicando que a ferramenta apresentou um desempenho ideal para a genotipagem automática quando avaliadas apenas as sequências passíveis de classificação direta. Esses resultados reforçam a robustez da ferramenta no contexto de genotipagem para o segmento L do vírus Oropouche, conforme descrito na Tabela 2.

Tabela 2 – Segmento L removendo sequências com baixa qualidade

Classe	VP	VN	FP	FN	SENS	ESPEC	ACUR
L2	672	75	0	1	100%	100%	100%
L1	75	672	1	0	100%	100%	100%

Fonte: Autora, 2024

Dessas sequências do segmento M, 48 foram classificadas como pertencentes ao genótipo M1 e 675 como pertencentes ao genótipo M2. Outras 66 sequências não puderam ser genotipadas automaticamente pela ferramenta e, portanto, foram classificadas posteriormente por meio de análise filogenética convencional, utilizando o método de máxima verossimilhança com o software IQ-TREE na versão 2 e o modelo de substituição GTR+G4+F, conforme resumido na Tabela 3.

Tabela 3 – Segmento M considerando sequências com baixa qualidade

Classe	VP	VN	FP	FN	SENS	ESPEC	ACUR
M2	671	112	2	4	99%	98%	99%
M1	47	694	47	1	98%	94%	94%
Unassigned	22	723	0	44	33%	100%	94%

Fonte: Autora, 2024.

A análise foi repetida desconsiderando as 66 sequências do segmento M que não puderam ser genotipadas automaticamente pela ferramenta na análise inicial. Nessa nova abordagem, as 48 sequências previamente classificadas como pertencentes ao genótipo M1 e as 675 sequências como pertencentes ao genótipo M2 foram mantidas, mas sem a inclusão das 66 sequências classificadas manualmente.

Com essa modificação, observou-se que os valores de sensibilidade, especificidade e acurácia aumentaram para 99%, indicando que a ferramenta apresentou um desempenho ideal para a genotipagem automática quando avaliadas apenas as sequências passíveis de classificação direta. Esses resultados reforçam a robustez da ferramenta no contexto de genotipagem para o segmento M do vírus Oropouche, conforme descrito na Tabela 4.

Tabela 4 – Segmento M removendo sequências com baixa qualidade

Classe	VP	VN	FP	FN	SENS	ESPEC	ACUR
M2	671	47	1	4	99%	98%	99%
M1	47	671	4	1	98%	99%	99%

Fonte: Autora, 2024.

Dessas sequências do segmento S, 143 foram classificadas como pertencentes ao genótipo S1 e 783 como pertencentes ao genótipo S2. Outras 6 sequências não puderam ser genotipadas automaticamente pela ferramenta e, portanto, foram classificadas posteriormente por meio de análise filogenética convencional, utilizando o método de máxima verossimilhança com o software IQ-TREE na versão 2 e o modelo de substituição GTR+G4+F, conforme resumido na Tabela 5.

Tabela 5 – Segmento S considerando sequências com baixa qualidade

Classe	VP	VN	FP	FN	SENS	ESPEC	ACUR
S1	139	780	9	4	97%	99%	99%
S2	777	146	3	6	99%	98%	99%
Unassigned	2	924	2	4	33%	100%	99%

Fonte: Autora, (2024).

A análise foi repetida desconsiderando as 6 sequências do segmento S que não puderam ser genotipadas automaticamente pela ferramenta na análise inicial. Nessa nova abordagem, as 143 sequências previamente classificadas como pertencentes ao genótipo S1 e as 783 sequências como pertencentes ao genótipo S2 foram mantidas, mas sem a inclusão das 6 sequências classificadas manualmente.

Com essa modificação, observou-se que os valores de sensibilidade, especificidade e acurácia obteve o mesmo desempenho, indicando que a ferramenta apresentou um desempenho ideal para a genotipagem automática quando avaliadas apenas as sequências passíveis de classificação direta. Esses resultados reforçam a robustez da ferramenta no contexto de genotipagem para o segmento S do vírus Oropouche, conforme descrito na Tabela 6.

Tabela 6 – Segmento S removendo sequências com baixa qualidade

Classe	VP	VN	FP	FN	SENS	ESPEC	ACUR
S1	139	778	5	4	97%	99%	99%
S2	777	140	3	6	99%	98%	99%
Unassigned	0	924	2	0	–	100%	100%

Fonte: Autora, 2024.

Dessa forma, a utilização do *benchmarking* com métricas quantitativas ofereceu uma abordagem sólida e sistemática para validar a ferramenta, garantindo que ela não apenas atenda aos requisitos técnicos, mas também seja competitiva em termos de desempenho. Esse processo permitiu a análise detalhada de métricas como precisão, sensibilidade, especificidade e acurácia, proporcionando uma visão ampla do comportamento da ferramenta frente aos segmentos analisados. Com base nesses resultados, decisões sobre ajustes, melhorias ou implementações futuras podem ser feitas com maior confiança.

A validação da ferramenta de genotipagem demonstrou resultados excelentes, evidenciando alta precisão, sensibilidade, especificidade e acurácia nos segmentos analisados do vírus Oropouche. Após a reanálise, desconsiderando sequências de baixa qualidade ou não classificáveis, a ferramenta atingiu índices de desempenho superiores, iguais a 99%, confirmando sua robustez e confiabilidade para a genotipagem automática. Esses resultados não apenas reforçam a eficácia da metodologia implementada, como também posicionam a ferramenta como uma solução eficiente e competitiva para estudos genômicos, contribuindo significativamente para a análise filogenética e a vigilância epidemiológica de vírus emergentes.

4.3 Limitações Técnicas e Impacto nos Resultados

Embora a ferramenta de tipagem do vírus Oropouche tenha sido projetada para fornecer uma análise robusta e eficiente de sequências virais, a presença de algumas limitações conhecidas que podem impactar a precisão dos resultados em cenários específicos, tais limitações podem ser atribuídas a fatores como o tamanho das sequências analisadas, sua divergência genética e as características do algoritmo de reconstrução filogenética utilizado.

Uma das limitações significativas da ferramenta está relacionada à análise de sequências muito curtas, o comprimento das sequências tem um impacto direto na qualidade da reconstrução filogenética, pois sequências curtas contêm menos informações genéticas, o que pode resultar em agrupamentos filogenéticos imprecisos, sequências com um número reduzido de nucleotídeos podem não refletir adequadamente as relações evolutivas entre os

genótipos, levando a possíveis erros no agrupamento.

Dessa forma, as sequências curtas não fornecem informação suficiente para distinguir com precisão entre diferentes genótipos ou cepas, tornando o processo de genotipagem mais propenso a falhas. A falta de dados genéticos completos pode causar ambiguidades nos resultados, gerando impacto nos resultados.

Outra limitação importante está associada à análise de sequências altamente divergentes, quando as sequências em questão apresentam uma grande diversidade genética. A construção de árvores filogenéticas torna-se mais desafiadora, especialmente quando a divergência é tal que as sequências podem representar diferentes grupos evolutivos ou cepas do vírus. Dessa forma, a ferramenta pode, nesse caso, ter dificuldades em resolver relações filogenéticas precisas, levando à formação de agrupamentos equivocados ou a dificuldades em discernir a posição de algumas sequências na árvore filogenética.

Além das sequências curtas e divergentes, outra limitação potencial está associada à presença de gaps nas sequências, como sequências incompletas ou com áreas ausentes de dados. Os dados incompletos podem dificultar a construção de árvores filogenéticas precisas, já que as lacunas podem ser interpretadas de forma inadequada pelos algoritmos, levando a uma distorção nos agrupamentos filogenéticos.

As limitações da ferramenta refletem os desafios comuns enfrentados em análises filogenéticas, particularmente no que se refere ao tratamento de sequências curtas, divergentes ou incompletas, e às limitações dos algoritmos de reconstrução filogenética utilizados. Tais limitações podem afetar a precisão dos resultados e a confiabilidade dos agrupamentos filogenéticos, especialmente em cenários de dados ruidosos ou altamente divergentes.

4.4 Direções Futuras e Oportunidades de Melhoria

Atualmente, a ferramenta utiliza um conjunto de métodos clássicos de análise filogenética, como os algoritmos NJ, Máxima ML e Bayesiano. No entanto, há uma crescente necessidade de incorporar métodos mais avançados e específicos que possam melhorar a precisão e a robustez das análises, especialmente em cenários com sequências altamente divergentes ou incompletas.

Sendo assim, para melhorar a ferramenta futuramente, seria vantajoso incorporar modelos evolutivos mais flexíveis, estratégias de validação de dados e algoritmos de alinhamento mais robustos, além de fornecer recomendações claras sobre a qualidade dos dados necessários para garantir resultados precisos e confiáveis. O uso de modelos de aprendizado de máquina pode ser explorado para melhorar a capacidade da ferramenta de identificar padrões em grandes volumes de dados e classificar sequências com maior precisão discutida em 2.5. Esses métodos poderiam ser treinados para reconhecer relações

evolutivas complexas e melhorar a qualidade da tipagem em cenários mais desafiadores.

A utilização de técnicas de processamento paralelo e otimização de algoritmos, como a implementação de algoritmos distribuídos, poderia melhorar significativamente o tempo de execução, permitindo que a ferramenta lidasse com grandes volumes de dados de forma mais eficiente, gerando uma otimização computacional. Outro aspecto importante é melhorar a qualidade da análise de dados de baixa qualidade, sendo uma das limitações, como sequências com gaps, erros de sequenciamento ou baixa cobertura, a implementação de algoritmos mais robustos para o tratamento de dados ruidosos poderia permitir a análise de um maior número de sequências, minimizando os impactos de dados com qualidade inferior.

A expansão e aprimoramento da ferramenta de tipagem do vírus Oropouche podem torná-la um recurso ainda mais valioso para a comunidade científica. A inclusão de novos métodos de análise, como algoritmos de aprendizado de máquina ou redes filogenéticas, são passos fundamentais para aumentar a flexibilidade, precisão e acessibilidade da ferramenta. Além disso, a otimização da capacidade de processamento e a integração com outras plataformas de colaboração científica podem tornar a ferramenta mais útil em estudos virais, epidemiológicos e evolutivos, consolidando-a como uma ferramenta indispensável no cenário global de pesquisa em virologia.

5 CONCLUSÃO

Este trabalho teve como objetivo desenvolver uma ferramenta de bioinformática para a classificação genômica do vírus Oropouche, focando na automação, precisão e acessibilidade para profissionais da saúde pública.

Os resultados alcançados demonstraram que a ferramenta é altamente eficaz, com desempenho superior a 99% em sensibilidade, especificidade e acurácia, evidenciando seu alinhamento aos objetivos de automação e classificação precisa. A integração de pipelines automatizados e modelos filogenéticos permitiu a análise rápida e confiável de linhagens virais, enquanto a interface gráfica intuitiva garantiu acessibilidade e facilidade de uso, atendendo às necessidades de pesquisadores e profissionais de saúde.

Além disso, a ferramenta destacou-se por sua capacidade de realizar análises em larga escala, contribuindo significativamente para a vigilância genômica e o controle de surtos virais. Por exemplo, sua aplicação poderia ter auxiliado no monitoramento de surtos recentes, como o ocorrido na Bahia, oferecendo suporte em tempo real para decisões estratégicas das autoridades de saúde.

Apesar de seu sucesso, limitações foram identificadas, como a baixa qualidade de algumas sequências e desafios de escalabilidade. Para enfrentar esses obstáculos, estratégias como a integração de algoritmos mais eficientes, uso de aprendizado de máquina e adoção de tecnologias de computação em nuvem foram propostas, garantindo a evolução contínua da ferramenta.

O progresso contínuo na bioinformática será essencial para que a ferramenta se mantenha relevante diante das rápidas mudanças genéticas do OROV e de outras doenças infecciosas emergentes. Com a incorporação de novas tecnologias e a atualização de seus algoritmos, a plataforma continuará sendo uma solução indispensável para a vigilância genômica e epidemiológica no futuro.

REFERÊNCIAS

- ABADI, S. *et al.* Modelteller: model selection for optimal phylogenetic reconstruction using machine learning. **Molecular biology and evolution**, Oxford University Press, v. 37, n. 11, p. 3338–3352, 2020. Citado na página 40.
- ABDALLA, B. **Replicação Viral**. 2020. <https://jornal.usp.br/ciencias/covid-19-como-o-virus-saltou-de-morcegos-para-humanos/>. Citado na página 23.
- ABUÍN, J. M. *et al.* Bigbwa: approaching the burrows–wheeler aligner to big data technologies. **Bioinformatics**, Oxford University Press, v. 31, n. 24, p. 4003–4005, 2015. Citado na página 38.
- ACRANI, G. O. *et al.* Establishment of a minigenome system for oropouche virus reveals the s genome segment to be significantly longer than reported previously. **Journal of General Virology**, Microbiology Society, v. 96, n. 3, p. 513–523, 2015. Citado nas páginas 26 e 48.
- ALBRECHT, M. *et al.* **Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies**. 2012. Citado na página 41.
- ALHARBI, W. S.; RASHID, M. A review of deep learning applications in human genomics using next-generation sequencing data. **Human Genomics**, Springer, v. 16, n. 1, p. 26, 2022. Citado na página 18.
- ALTSCHUL, S. F. *et al.* Basic local alignment search tool. **Journal of molecular biology**, Elsevier, v. 215, n. 3, p. 403–410, 1990. Citado nas páginas 35 e 52.
- ALTSCHUL, S. F. *et al.* Gapped blast and psi-blast: a new generation of protein database search programs. **Nucleic acids research**, Oxford University Press, v. 25, n. 17, p. 3389–3402, 1997. Citado na página 64.
- AMBRÓSIO, P. E. Biologia computacional no apoio à saúde. **Journal of Health Informatics**, v. 13, n. 3, 2021. Citado na página 33.
- ANDERSON, D. J. **Kanban: successful evolutionary change for your technology business**. [S.l.]: Blue Hole Press, 2010. Citado na página 56.
- ARNOLD, K. The java programming language. **Addition Wesley google schola**, v. 2, 2005. Citado na página 57.
- ARNOLD, L. *et al.* **Random dynamical systems**. [S.l.]: Springer, 1995. Citado na página 60.
- ARONSON, S. J.; REHM, H. L. Building the foundation for genomics in precision medicine. **Nature**, Nature Publishing Group UK London, v. 526, n. 7573, p. 336–342, 2015. Citado na página 33.
- AZOURI, D. *et al.* Harnessing machine learning to guide phylogenetic-tree search algorithms. **Nature communications**, Nature Publishing Group UK London, v. 12, n. 1, p. 1983, 2021. Citado na página 40.

- BANDEIRA, A. C. *et al.* Fatal oropouche virus infections in nonendemic region, brazil, 2024. **Emerging Infectious Diseases**, v. 30, n. 11, p. 2370, 2024. Citado na página 15.
- BARO, E. *et al.* Toward a literature-driven definition of big data in healthcare. **BioMed research international**, Wiley Online Library, v. 2015, n. 1, p. 639021, 2015. Citado na página 37.
- BAXEVANIS, A. D.; BADER, G. D.; WISHART, D. S. **Bioinformatics**. [S.l.]: John Wiley & Sons, 2020. Citado na página 16.
- BEATY, B. J.; TRENT, D. W.; ROEHRIG, J. T. Virus variation and evolution: mechanisms and epidemiological significance. In: **The Arboviruses**. [S.l.]: CRC Press, 2020. p. 59–86. Citado na página 32.
- Benegrip. **O que são imunidade inata e adaptativa? Para que servem?** 2024. Disponível em: <https://www.benegrip.com.br/saude/prevencao-gripe/o-que-sao-imunidade-inata-e-adaptativa-para-que-servem>. Citado na página 24.
- Biomedicina Padrão. **O que são Polimorfismos de Nucleotídeo Único (SNPs)?** 2020. Disponível em: <https://www.biomedicinapadrao.com.br/2020/02/o-que-sao-polimorfismos-de-nucleotideo.html>. Citado na página 43.
- BLOMBERG, S. P.; JR, T. G.; IVES, A. R. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. **Evolution**, Wiley Online Library, v. 57, n. 4, p. 717–745, 2003. Citado na página 54.
- BORGES, L. E. **Python para desenvolvedores: aborda Python 3.3**. [S.l.]: Novatec Editora, 2014. Citado na página 60.
- BRASIL. Constituição da república federativa do brasil de 1988. **Presidência da República** — Art. 225, Presidência da República, 1988. Título VIII: Da Ordem Social. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Citado na página 45.
- BROMBERG, Y.; CAPRIOTTI, E. Thoughts from snp-sig 2012: future challenges in the annotation of genetic variations. **BMC genomics**, Springer, v. 14, n. Suppl 3, p. S1, 2013. Citado na página 42.
- BROOKES, A. J. The essence of snps. **Gene**, Elsevier, v. 234, n. 2, p. 177–186, 1999. Citado na página 43.
- BULT, C. J. Bioinformatics: A practical guide to the analysis of genes and proteins. **Science**, American Association for the Advancement of Science, v. 282, n. 5389, p. 635–636, 1998. Citado na página 17.
- CALLEGARI-JACQUES, S. M. **Bioestatística: princípios e aplicações**. [S.l.]: Artmed Editora, 2009. Citado na página 56.
- CAPURRO, R. Privacy. an intercultural perspective. **Ethics and information technology**, Springer, v. 7, p. 37–47, 2005. Citado na página 44.
- CARRIÇO, J. Visualização e algoritmos de clustering para análise filogenética. **Citado**, v. 3, p. 11, 2015. Citado na página 55.

CASSEB, A. R. *et al.* Prevalência de anticorpos contra arbovírus da família bunyaviridae em búfalos de água. **Ciência Animal Brasileira**, SciELO Brasil, v. 16, n. 3, p. 428–436, 2015. Citado na página 37.

CHERRY, J. *et al.* **Feigin and Cherry's Textbook of Pediatric Infectious Diseases E-Book: 2-Volume Set**. [S.l.]: Elsevier Health Sciences, 2013. Citado na página 30.

CHICCO, D. Ten quick tips for machine learning in computational biology. **BioData mining**, Springer, v. 10, n. 1, p. 35, 2017. Citado na página 33.

CHOR, B.; TULLER, T. Maximum likelihood of evolutionary trees: hardness and approximation. **Bioinformatics**, Oxford University Press, v. 21, n. suppl_1, p. i97–i106, 2005. Citado na página 40.

CHUTE, C. G. *et al.* Some experiences and opportunities for big data in translational research. **Genetics in Medicine**, Elsevier, v. 15, n. 10, p. 802–809, 2013. Citado na página 38.

CLARKE, D. H.; CASALS, J. Techniques for hemagglutination and hemagglutination-inhibition with arthropod-borne viruses. **Am J Trop Med Hyg**, ASTMH, v. 7, n. 5, p. 561–573, 1958. Citado na página 32.

CLEEMPOT, S. *et al.* Genome detective coronavirus typing tool for rapid identification and characterization of novel coronavirus genomes. **Bioinformatics**, Oxford University Press, v. 36, n. 11, p. 3552–3555, 2020. Citado na página 50.

CLOTET, J. Bioética como ética aplicada e genética. **Revista Bioética**, v. 5, n. 2, 1997. Citado na página 45.

COCK, P. J. *et al.* Biopython: freely available python tools for computational molecular biology and bioinformatics. **Bioinformatics**, Oxford University Press, v. 25, n. 11, p. 1422, 2009. Citado na página 60.

COIMBRA, R. C. M. Ferramenta de visualização interativa de comparação entre múltiplos genomas para a identificação de sentenças. 2013. Citado nas páginas 46 e 47.

CROSBIE, D. **Protection of genetic information: an international comparison**. [S.l.]: Department of Health, 2000. Citado na página 45.

CRUVINEL, W. d. M. *et al.* Sistema imunitário: Parte i. fundamentos da imunidade inata com ênfase nos mecanismos moleculares e celulares da resposta inflamatória. **Revista Brasileira de Reumatologia**, SciELO Brasil, v. 50, p. 434–447, 2010. Citado nas páginas 23, 24 e 25.

DANTAS, L. E. C.; ANDRADE, G. S. *et al.* Proteção de dados genéticos: repercussões jurídicas e éticas na ausência de legislação não discriminatória no campo das seguradoras. **Revista Jurídica (FURB)**, v. 26, p. e10127–e10127, 2022. Citado na página 45.

D'ANTONIO, M. *et al.* Wep: a high-performance analysis pipeline for whole-exome data. **BMC bioinformatics**, Springer, v. 14, p. 1–11, 2013. Citado na página 42.

DARCY, A. M.; LOUIE, A. K.; ROBERTS, L. W. Machine learning and the profession of medicine. **Jama**, American Medical Association, v. 315, n. 6, p. 551–552, 2016. Citado na página 39.

- DARLING, A. C. *et al.* Mauve: multiple alignment of conserved genomic sequence with rearrangements. **Genome research**, Cold Spring Harbor Lab, v. 14, n. 7, p. 1394–1403, 2004. Citado na página 46.
- DASA. **Imunidade inata e adquirida**. 2024. Disponível em: <https://nav.dasa.com.br/blog/imunidade-inata-e-adquirida>. Citado na página 24.
- DAWKINS, R. Arresting evidence. **The Sciences**, Wiley Online Library, v. 38, n. 6, p. 20–25, 1998. Citado na página 45.
- DEELEN, P. *et al.* Calling genotypes from public rna-sequencing data enables identification of genetic variants that affect gene-expression levels. **Genome medicine**, Springer, v. 7, p. 1–13, 2015. Citado na página 34.
- DEGRAVE, W.; MACHADO, C. J. S. Bioinformática e saúde: uma visão geral. **Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, v. 1, n. 2, 2007. Citado na página 17.
- DELCHER, A. L. *et al.* Alignment of whole genomes. **Nucleic acids research**, Oxford University Press, v. 27, n. 11, p. 2369–2376, 1999. Citado na página 46.
- DELGADO, M. X. T. Aprimoramentos da solução paralela baseada em operações coletivas para o bootstrap da reconstrução de árvores filogenéticas no phyml 3.0. In: **Colloquium Exactarum. ISSN: 2178-8332**. [S.l.: s.n.], 2020. v. 12, n. 3, p. 39–52. Citado na página 55.
- DEO, R. C. Machine learning in medicine. **Circulation**, Am Heart Assoc, v. 132, n. 20, p. 1920–1930, 2015. Citado nas páginas 38 e 39.
- DICIO. **Significado de Negligenciados**. 2024. Disponível em: <https://www.dicio.com.br/negligenciados/#:~:text=Significado%20de%20Negligenciados,%2C%20esquecidos%2C%20ignorados%2C%20largados>. Citado na página 13.
- DIGIAMPIETRI, L. A.; SETUBAL, J. C.; MEDEIROS, C. B. Gerenciamento de workflows científicos em bioinformática. **SBC**, p. 33, 2008. Citado na página 42.
- DIGITAL, O. Febre oropouche: casos estão se espalhando pelo brasil, alerta ministério da saúde. 2024. Disponível em: <https://olhardigital.com.br/2024/05/15/medicina-e-saude/febre-oropouche-casos-estao-se-espalhando-pelo-brasil-alerta-ministerio-da-saude/>. Citado na página 15.
- DOXSEY, J. R.; RIZ, J. D. Metodologia da pesquisa científica. **ESAB–Escola Superior Aberta do Brasil**, v. 2003, p. 36, 2002. Citado na página 52.
- DUGGAL, P. *et al.* The evolving field of genetic epidemiology: from familial aggregation to genomic sequencing. **American journal of epidemiology**, Oxford University Press, v. 188, n. 12, p. 2069–2077, 2019. Citado na página 34.
- DUTHEIL, J. Y.; FIGUET, E. Optimization of sequence alignments according to the number of sequences vs. number of sites trade-off. **BMC bioinformatics**, Springer, v. 16, p. 1–11, 2015. Citado na página 34.
- EDUCAÇÃO, M. **Vírus**. 2024. <https://mundoeducacao.uol.com.br/biologia/virus.htm>. Citado na página 21.

- FANG, Z. *et al.* Multiple nucleotide polymorphism dna markers for the accurate evaluation of genetic variations. **bioRxiv**, Cold Spring Harbor Laboratory, p. 2021–03, 2021. Citado na página 44.
- FELSENSTEIN, J. Confidence limits on phylogenies: an approach using the bootstrap. **evolution**, Blackwell Publishing Inc Malden, USA, v. 39, n. 4, p. 783–791, 1985. Citado nas páginas 55 e 66.
- FELSENSTEIN, J. Phylip (phylogeny inference package) version. **(No Title)**, v. 3, 2006. Citado na página 54.
- FELSENSTEIN, J.; KISHINO, H. Is there something wrong with the bootstrap on phylogenies? a reply to hillis and bull. **Systematic Biology**, JSTOR, v. 42, n. 2, p. 193–200, 1993. Citado na página 66.
- FERRUCCI, D. *et al.* Building watson: An overview of the deepqa project. **AI magazine**, v. 31, n. 3, p. 59–79, 2010. Citado na página 19.
- FIGUEIREDO, M. L. G. d.; FIGUEIREDO, L. T. M. Emerging alphaviruses in the americas: Chikungunya and mayaro. **Revista da Sociedade Brasileira de Medicina Tropical**, SciELO Brasil, v. 47, p. 677–683, 2014. Citado na página 15.
- FIGUEIREDO, R. *et al.* Arboviroses. 2017. Citado na página 13.
- FOLEY, J. A. *et al.* Amazonia revealed: forest degradation and loss of ecosystem goods and services in the amazon basin. **Frontiers in Ecology and the Environment**, Wiley Online Library, v. 5, n. 1, p. 25–32, 2007. Citado nas páginas 36 e 37.
- FONSECA, N. A. *et al.* Tools for mapping high-throughput sequencing data. **Bioinformatics**, Oxford University Press, v. 28, n. 24, p. 3169–3177, 2012. Citado na página 44.
- FONSECA, V. *et al.* A computational method for the identification of dengue, zika and chikungunya virus species and genotypes. **PLoS neglected tropical diseases**, Public Library of Science San Francisco, CA USA, v. 13, n. 5, p. e0007231, 2019. Citado na página 18.
- FONSECA, V. *et al.* A computational method for the identification of dengue, zika and chikungunya virus species and genotypes. **PLoS neglected tropical diseases**, Public Library of Science San Francisco, CA USA, v. 13, n. 5, p. e0007231, 2019. Citado na página 49.
- FORLEO-NETO, E. *et al.* Influenza. **Revista da Sociedade Brasileira de Medicina Tropical**, Sociedade Brasileira de Medicina Tropical - SBMT, v. 36, n. 2, p. 267–274, Mar 2003. ISSN 0037-8682. Disponível em: <https://doi.org/10.1590/S0037-86822003000200011>. Citado na página 22.
- FRIEDMAN, J. The elements of statistical learning: Data mining, inference, and prediction. **(No Title)**, 2009. Citado na página 38.
- FRITH, M. C.; WAN, R.; HORTON, P. Incorporating sequence quality data into alignment improves dna read mapping. **Nucleic acids research**, Oxford University Press, v. 38, n. 7, p. e100–e100, 2010. Citado na página 34.

- Fundação Oswaldo Cruz. Arboviroses que afetam a saúde pública no Brasil. 2024. Disponível em: https://vigiepidemia.matogrossodosul.fiocruz.br/vigilancia/modulo3/mod3_un1_sec4_pag1.html. Citado na página 31.
- G1 Bahia. **Bahia registra 1041 casos de Febre Oropouche em setembro.** 2024. Disponível em: <https://g1.globo.com/ba/bahia/noticia/2024/09/18/casos-febre-oropouche-setembro.ghtml>. Citado na página 16.
- GARRISON, E.; MARTH, G. Haplotype-based variant detection from short-read sequencing. **arXiv preprint arXiv:1207.3907**, 2012. Citado na página 44.
- GHEDIN, E. *et al.* Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. **Nature**, Nature Publishing Group UK London, v. 437, n. 7062, p. 1162–1166, 2005. Citado na página 47.
- GIOVANETTI, M. *et al.* Genomic epidemiology unveils the dynamics and spatial corridor behind the yellow fever virus outbreak in southern Brazil. **Science Advances**, v. 9, n. 35, p. eadg9204, 2023. Disponível em: <https://www.science.org/doi/abs/10.1126/sciadv.adg9204>. Citado na página 18.
- GOLDSTEIN, D. B. *et al.* Sequencing studies in human genetics: design and interpretation. **Nature Reviews Genetics**, Nature Publishing Group UK London, v. 14, n. 7, p. 460–470, 2013. Citado na página 42.
- GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep learning. book in preparation for MIT Press. **URL: <http://www.deeplearningbook.org>**, v. 1, 2016. Citado na página 39.
- GRIFFITHS, A. J. *et al.* Introdução à genética. In: **Introdução à genética**. [S.l.: s.n.], 2006. p. 743–743. Citado na página 43.
- HAASTER, C. M. Clerx-van *et al.* Nucleotide sequence analyses and predicted coding of bunyavirus genome RNA species. **Journal of Virology**, Am Soc Microbiol, v. 41, n. 1, p. 119–128, 1982. Citado na página 28.
- HARDISON, R. C. Comparative genomics. **PLoS biology**, Public Library of Science San Francisco, USA, v. 1, n. 2, p. e58, 2003. Citado na página 46.
- HASIN, Y.; SELDIN, M.; LUSIS, A. Multi-omics approaches to disease. **Genome biology**, Springer, v. 18, p. 1–15, 2017. Citado na página 17.
- HE, K. Y.; GE, D.; HE, M. M. Big data analytics for genomic medicine. **International journal of molecular sciences**, MDPI, v. 18, n. 2, p. 412, 2017. Citado na página 38.
- HEALTH, N. I. of *et al.* What are single nucleotide polymorphisms (snps). **Genetics Home Reference-NIH. US National Library of Medicine. URL: <https://ghr.nlm.nih.gov/primer/genomeresearch/snp>**, 2019. Citado na página 43.
- HOCH, A. L.; ROBERTS, D. R.; PINHEIRO, F. d. P. Breeding sites of *Culiseta* paraensis and options for control by environmental management. **Bulletin of the Pan American Health Organization (PAHO)**; 20 (3), 1986, 1986. Citado na página 28.

- HOTEZ, P. J. **Forgotten people, forgotten diseases: the neglected tropical diseases and their impact on global health and development**. [S.l.]: John Wiley & Sons, 2021. Citado na página 17.
- HUANG, Q. *et al.* **Innovative testing and measurement solutions for smart grid**. [S.l.]: John Wiley & Sons, 2015. Citado na página 37.
- International Business Machines Corporation. **IBM - United States**. 2024. Disponível em: <https://www.ibm.com/us-en>. Citado na página 17.
- JALILI, V. *et al.* The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. **Nucleic acids research**, Oxford University Press, v. 48, n. W1, p. W395–W402, 2020. Citado na página 41.
- JAYASHANKAR, C. *et al.* Neglected tropical diseases: A comprehensive review. **Cureus**, Cureus, v. 16, n. 2, 2024. Citado na página 13.
- JO, H.; KOH, G. Faster single-end alignment generation utilizing multi-thread for bwa. **Bio-medical materials and engineering**, IOS Press, v. 26, n. s1, p. S1791–S1796, 2015. Citado na página 36.
- JUMA, J. *et al.* Genomic surveillance of rift valley fever virus: from sequencing to lineage assignment. **BMC Genomics**, v. 23, n. 1, p. 520, July 2022. Disponível em: <https://doi.org/10.1186/s12864-022-08764-6>. Citado na página 18.
- JUMA, J. *et al.* Genomic surveillance of rift valley fever virus: from sequencing to lineage assignment. **BMC genomics**, Springer, v. 23, n. 1, p. 520, 2022. Citado na página 50.
- JUNIOR, H. L. R.; OLIVEIRA, R. T. G. d.; CECCATTO, V. M. Bioinformática como recurso pedagógico para o curso de ciências biológicas na universidade estadual do ceará-uece-fortaleza, estado do ceará. **Acta Scientiarum. Education**, v. 34, n. 01, p. 129–138, 2012. Citado na página 33.
- JUNIOR, N. F. de A. Ferramentas para comparação genômica. 2002. Citado na página 46.
- KAHL, G. *et al.* Single nucleotide polymorphisms: detection techniques and their potential for genotyping and genome mapping. **The Handbook of Plant Genome Mapping: Genetic and Physical Mapping**, Wiley Online Library, p. 75–107, 2005. Citado na página 43.
- KARABATSOS, N. American society of tropical medicine and hygiene. **San Antonio, TX**, 1985. Citado na página 31.
- KATOH, K.; ASIMENOS, G.; TOH, H. Multiple alignment of dna sequences with mafft. **Bioinformatics for DNA sequence analysis**, Springer, p. 39–64, 2009. Citado na página 36.
- KATOH, K. *et al.* Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. **Nucleic acids research**, Oxford University Press, v. 30, n. 14, p. 3059–3066, 2002. Citado nas páginas 36 e 53.

- KATOH, K.; STANDLEY, D. M. Mafft multiple sequence alignment software version 7: improvements in performance and usability. **Molecular biology and evolution**, Society for Molecular Biology and Evolution, v. 30, n. 4, p. 772–780, 2013. Citado na página 53.
- KEARSE, M. *et al.* Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. **Bioinformatics**, Oxford University Press, v. 28, n. 12, p. 1647–1649, 2012. Citado na página 55.
- KEEL, B. N.; SNELLING, W. M. Comparison of burrows-wheeler transform-based mapping algorithms used in high-throughput whole-genome sequencing: application to illumina data for livestock genomes. **Frontiers in genetics**, Frontiers Media SA, v. 9, p. 35, 2018. Citado na página 44.
- Khan Academy. **The Immune System Review**. 2017. Disponível em: <https://pl.khanacademy.org/science/high-school-biology/hs-human-body-systems/hs-the-immune-system/a/hs-the-immune-system-review>. Citado na página 25.
- Khan Academy. Sistema imunológico. 2018. Disponível em: <https://pt.khanacademy.org/science/7-ano/sistema-imunologico/corpo-humano-sistema-imunologico/a/sistema-imunologico>. Citado na página 24.
- KNIBERG, H.; SKARIN, M. **Kanban and Scrum-making the most of both**. [S.l.]: Lulu. com, 2010. Citado na página 56.
- KORF, I.; YANDELL, M.; BEDELL, J. **Blast**. [S.l.]: "O'Reilly Media, Inc.", 2003. Citado na página 58.
- KUMAR, S.; BANKS, T. W.; CLOUTIER, S. Snp discovery through next-generation sequencing and its applications. **Crop Breeding**, Apple Academic Press, p. 209–244, 2016. Citado na página 44.
- LAM, H. Y. *et al.* Detecting and annotating genetic variations using the hugeseq pipeline. **Nature biotechnology**, Nature Publishing Group US New York, v. 30, n. 3, p. 226–229, 2012. Citado na página 36.
- LAN, K. *et al.* A survey of data mining and deep learning in bioinformatics. **Journal of medical systems**, Springer, v. 42, p. 1–20, 2018. Citado na página 35.
- LARRANAGA, P. *et al.* Machine learning in bioinformatics. **Briefings in bioinformatics**, Oxford University Press, v. 7, n. 1, p. 86–112, 2006. Citado na página 38.
- LEÃO, R. N. Q. d. Doenças infecciosas e parasitárias: enfoque amazônico. In: **Doenças infecciosas e parasitárias: enfoque amazônico**. [S.l.: s.n.], 1997. p. 886–886. Citado nas páginas 30, 31 e 32.
- LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015. Citado na página 39.
- LEDUC, J. *et al.* Epidemic oropouche virus disease in northern brazil. **Bulletin of the Pan American Health Organization (PAHO)**; **15 (2)**, 1981, 1981. Citado na página 30.
- LEDUC, J. W.; PINHEIRO, F. P. Oropouche fever. In: **The Arboviruses**. [S.l.]: CRC Press, 2019. p. 1–14. Citado na página 29.

- LEE, S. A. *et al.* The impact of climate suitability, urbanisation, and connectivity on the expansion of dengue in 21st century brazil. **PLoS Neglected Tropical Diseases**, Public Library of Science San Francisco, CA USA, v. 15, n. 12, p. e0009773, 2021. Citado na página 13.
- LENGAUER, T. Bioinformatics—from genomes to therapies. **Bioinformatics-From Genomes to Therapies**, Wiley-VCH Verlag GmbH Weinheim, Germany, p. 1–24, 2007. Citado na página 16.
- LI, H. **Burrows-Wheeler Aligner (BWA)**. 2024. Disponível em: <https://bio-bwa.sourceforge.net/>. Citado na página 38.
- LI, H. *et al.* The sequence alignment/map format and samtools. **bioinformatics**, Oxford University Press, v. 25, n. 16, p. 2078–2079, 2009. Citado na página 44.
- LI, W.; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, Oxford University Press, v. 22, n. 13, p. 1658–1659, 2006. Citado na página 35.
- LI, Z. *et al.* The worldwide seroprevalence of dengue, chikv and zikv infection: A systematic review and meta-analysis. **PLOS Neglected Tropical Diseases**, Public Library of Science, v. 15, n. 4, p. 1–17, 04 2021. Disponível em: <https://doi.org/10.1371/journal.pntd.0009337>. Citado na página 14.
- LIM, S. M. *et al.* West nile virus: Immunity and pathogenesis. **Viruses**, v. 3, n. 6, p. 811–828, 2011. ISSN 1999-4915. Disponível em: <https://www.mdpi.com/1999-4915/3/6/811>. Citado na página 14.
- LINLEY, J.; HOCH, A.; PINHEIRO, F. Biting midges (diptera: Ceratopogonidae) and human health. **Journal of Medical Entomology**, Oxford University Press Oxford, UK, v. 20, n. 4, p. 347–364, 1983. Citado na página 28.
- LONG, H.; LI, M.; FU, H. Determination of optimal parameters of mafft program based on balibase3.0 database. **SpringerPlus**, Springer, v. 5, p. 1–9, 2016. Citado na página 54.
- LOPES, F. F. C. H. S. G. B. C. **Vírus**. 2014. Disponível em: <https://www.unijales.edu.br/library/downebook/id:235>. Citado nas páginas 22 e 23.
- LOPES, N.; NOZAWA, C.; LINHARES, R. E. C. Características gerais e epidemiologia dos arbovírus emergentes no brasil. **Revista Pan-Amazônica de Saúde**, v. 5, n. 3, p. 10–10, 2014. Citado nas páginas 13 e 31.
- LORENZ, C. *et al.* Impact of environmental factors on neglected emerging arboviral diseases. **PLoS neglected tropical diseases**, Public Library of Science San Francisco, CA USA, v. 11, n. 9, p. e0005959, 2017. Citado na página 37.
- LOWE, R. *et al.* The zika virus epidemic in brazil: from discovery to future implications. **International journal of environmental research and public health**, MDPI, v. 15, n. 1, p. 96, 2018. Citado na página 13.
- LUIGI. **Luigi: A Python Package for Building Complex Pipelines**. 2024. Disponível em: <https://github.com/spotify/luigi>. Citado na página 41.

MADIGAN, M. *et al.* **Brock biology of microorganisms 15th edition (2019)**(pp. 261–427). [S.l.]: Pearson, 2019. Citado na página 22.

MANUALS, M. **Considerações Gerais sobre Infecções Virais**. 2023. MSD Manuals. Disponível em: <https://www.msdmanuals.com/pt-br/casa/infec%C3%A7%C3%B5es/considera%C3%A7%C3%B5es-gerais-sobre-infec%C3%A7%C3%B5es-virais/considera%C3%A7%C3%B5es-gerais-sobre-infec%C3%A7%C3%B5es-virais>. Citado na página 22.

MARCHI, S. *et al.* Emerging and re-emerging arboviral diseases as a global health problem. **Public Health: Emerging and Re-emerging Issues. United Kindgom: IntechOpen**, p. 25–46, 2018. Citado na página 14.

MARIZ, F. **Covid-19: como o vírus saltou de morcegos para humanos**. 2020. <https://jornal.usp.br/ciencias/covid-19-como-o-virus-saltou-de-morcegos-para-humanos/>. Citado na página 23.

MARTINS, F. P.; PEDROSO, L. B. espaço geográfico. 2021. Citado na página 64.

MARX, V. The big challenges of big data. **Nature**, Nature Publishing Group, v. 498, n. 7453, p. 255–260, 2013. Citado na página 42.

MCGINNIS, S.; MADDEN, T. L. Blast: at the core of a powerful and diverse set of sequence analysis tools. **Nucleic acids research**, Oxford University Press, v. 32, n. suppl_2, p. W20–W25, 2004. Citado na página 65.

MCKENNA, A. *et al.* The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. **Genome research**, Cold Spring Harbor Lab, v. 20, n. 9, p. 1297–1303, 2010. Citado nas páginas 41 e 44.

MENEGUETTI, V. A. F. Dionatas Ulises de O. **Vírus ser vivo ou não? Eis a questão!** 2024. Disponível em: <https://www.redalyc.org/pdf/5704/570463830001.pdf>. Citado na página 21.

METRO1. Sobe para 307 número de casos da febre oropouche na bahia. 2024. Disponível em: <https://www.metro1.com.br/noticias/bahia/149673,sobe-para-307-numero-de-casos-da-febre-oropouche-na-bahia>. Citado na página 16.

MICROBIOLOGANDO. **Você sabe o que é um vírus?** 2024. <https://www.ufrgs.br/microbiologando/voce-sabe-o-que-e-um-virus/>. Citado na página 21.

MIN, S.; LEE, B.; YOON, S. Deep learning in bioinformatics. **Briefings in bioinformatics**, Oxford University Press, v. 18, n. 5, p. 851–869, 2017. Citado na página 39.

Ministério da Saúde. **Número de exames detectáveis para o vírus Oropouche por semana epidemiológica, Brasil, 2023 e 2024**. 2024. <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/o/oropouche/painel-epidemiologico>. Citado na página 15.

Ministério da Saúde. **Programa Educacional em Vigilância e Cuidado em Saúde no Enfrentamento da COVID-19 e de outras Doenças Virais**. 2024. https://vigiepidemia.matogrossodosul.fiocruz.br/vigilancia/modulo3/mod3_un1_subA_sec4_pag1.html. Citado na página 32.

- MITCHELL, T. M. **The discipline of machine learning**. [S.l.]: Carnegie Mellon University, School of Computer Science, Machine Learning ..., 2006. v. 9. Citado na página 38.
- MÖLDER, F. *et al.* Sustainable data analysis with snakemake. **F1000Research**, Faculty of 1000 Ltd, v. 10, 2021. Citado na página 41.
- MOLYNEUX, D. H.; SAVIOLI, L.; ENGELS, D. Neglected tropical diseases: progress towards addressing the chronic pandemic. **The Lancet**, Elsevier, v. 389, n. 10066, p. 312–325, 2017. Citado na página 17.
- MONTESINOS-LÓPEZ, O. A. *et al.* A review of deep learning applications for genomic selection. **BMC genomics**, Springer, v. 22, p. 1–23, 2021. Citado na página 18.
- MOTA, M. T. d. O. *et al.* Mosquito-transmitted viruses-the great brazilian challenge. **brazilian journal of microbiology**, SciELO Brasil, v. 47, p. 38–50, 2016. Citado na página 14.
- MOUNT, D. W. Using the basic local alignment search tool (blast). **Cold spring harbor Protocols**, Cold Spring Harbor Laboratory Press, v. 2007, n. 7, p. pdb-top17, 2007. Citado na página 35.
- MSD Manual Consumer Version. **Imunidade Adquirida**. 2017. <https://www.msdmanuals.com/pt-br/casa/doen%C3%A7as-imunol%C3%B3gicas/biologia-do-sistema-imunol%C3%B3gico/imunidade-adquirida>. Citado na página 25.
- MURPHY, F. A. *et al.* **Virus taxonomy: classification and nomenclature of viruses**. [S.l.]: Springer Science & Business Media, 2012. v. 10. Citado nas páginas 26 e 28.
- MURRAY, P.; ROSENTHAL, K.; PFALLER, M. **Togavírus e Flavivírus. Microbiologia médica**. [S.l.]: Rio de Janeiro: Guanabara Koogan, 2006. Citado na página 31.
- National Geographic Brasil. **Como um vírus infecta uma pessoa**. 2023. Disponível em: <https://www.nationalgeographicbrasil.com/ciencia/2023/05/como-um-virus-infecta-uma-pessoa#:~:text=De%20acordo%20com%20o%20manual,e%20libera%20seu%20material%20gen%C3%A9tico>. Citado na página 22.
- National Human Genome Research Institute. **Single Nucleotide Polymorphisms (SNPs)**. 2024. Disponível em: <https://www.genome.gov/genetics-glossary/Single-Nucleotide-Polymorphisms>. Citado na página 43.
- NAVECA, F. G. *et al.* Multiplexed reverse transcription real-time polymerase chain reaction for simultaneous detection of mayaro, oropouche, and oropouche-like viruses. **Memorias do Instituto Oswaldo Cruz**, SciELO Brasil, v. 112, p. 510–513, 2017. Citado na página 15.
- NELSON, M. I. *et al.* Stochastic processes are key determinants of short-term evolution in influenza a virus. **PLoS pathogens**, Public Library of Science San Francisco, USA, v. 2, n. 12, p. e125, 2006. Citado na página 47.

- NELSON, M. I. *et al.* Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. **PLoS pathogens**, Public Library of Science San Francisco, USA, v. 3, n. 9, p. e131, 2007. Citado na página 47.
- NETO, A. Z. *et al.* Análise de dados de rna-seq utilizando o galaxy. 2016. Citado na página 41.
- NEXTFLOW. **Nextflow Training: Customizing and Using nf-core Pipelines**. 2024. Disponível em: https://training.nextflow.io/nf_customize/02_nf-core/. Citado na página 41.
- NIELSEN, R. *et al.* Genotype and snp calling from next-generation sequencing data. **Nature Reviews Genetics**, Nature Publishing Group UK London, v. 12, n. 6, p. 443–451, 2011. Citado na página 44.
- NOGUEIRA, R. M. R.; ARAÚJO, J. M. G. d.; SCHATZMAYR, H. G. Dengue viruses in Brazil, 1986-2006. **Revista Panamericana de Salud Publica**, SciELO Public Health, v. 22, p. 358–363, 2007. Citado na página 13.
- NUNES, M. R. T. *et al.* Oropouche virus isolation, southeast Brazil. **Emerging Infectious Diseases**, Centers for Disease Control and Prevention, v. 11, n. 10, p. 1610, 2005. Citado nas páginas 29 e 30.
- OBERMEYER, Z.; EMANUEL, E. J. Predicting the future—big data, machine learning, and clinical medicine. **New England Journal of Medicine**, Mass Medical Soc, v. 375, n. 13, p. 1216–1219, 2016. Citado nas páginas 38 e 39.
- OCEANDROP. **Imunidade Inata e Adaptativa: Entenda a Diferença**. 2024. Disponível em: <https://www.oceandrop.com.br/blog/imunidade-inata-e-adaptativa>. Citado na página 23.
- OEVEREN, J. van; JANSSEN, A. Mining snps from dna sequence data; computational approaches to snp discovery and analysis. **Single Nucleotide Polymorphisms: Methods and Protocols**, Springer, p. 73–91, 2009. Citado na página 43.
- OGDEN, T. H.; ROSENBERG, M. S. Multiple sequence alignment accuracy and phylogenetic inference. **Systematic biology**, Oxford University Press, v. 55, n. 2, p. 314–328, 2006. Citado na página 40.
- OPENSTAX. **OpenStax**. 2021. <https://openstax.org/>. Citado na página 24.
- ORACLE. **Aprendizado Profundo Definido**. 2024. Disponível em: <https://www.oracle.com/br/artificial-intelligence/machine-learning/what-is-deep-learning/>. Citado na página 39.
- PAGEL, M. Inferring the historical patterns of biological evolution. **Nature**, Nature Publishing Group UK London, v. 401, n. 6756, p. 877–884, 1999. Citado na página 54.
- PANDEY, D.; PERUMAL, P. O. A scoping review on deep learning for next-generation rna-seq. data analysis. **Functional & Integrative Genomics**, Springer, v. 23, n. 2, p. 134, 2023. Citado na página 34.

- PAULA, S. O. D.; FONSECA, B. A. L. d. Dengue: a review of the laboratory tests a clinician must know to achieve a correct diagnosis. **Brazilian Journal of Infectious Diseases**, SciELO Brasil, v. 8, p. 390–398, 2004. Citado na página 32.
- PEARSON, W. R. Blast and fasta similarity searching for multiple sequence alignment. **Multiple sequence alignment methods**, Springer, p. 75–101, 2014. Citado na página 36.
- PEREIRA, R.; OLIVEIRA, J.; SOUSA, M. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. **Journal of clinical medicine**, MDPI, v. 9, n. 1, p. 132, 2020. Citado na página 33.
- PEREIRA, R. S. *et al.* Aspectos epidemiológicos del virus oropouche (orthobunyavirus) en américa del sur: una revisión sistemática. **Revista Colombiana de Ciencias Químico-Farmacéuticas**, v. 51, n. 1, p. 166–184, 2022. Citado na página 14.
- PEREIRA, T. N. *et al.* Vector competence of aedes aegypti, aedes albopictus, and culex quinquefasciatus mosquitoes for mayaro virus. **PLOS Neglected Tropical Diseases**, v. 14, n. 4, p. e0007518, 2020. Disponível em: <https://doi.org/10.1371/journal.pntd.0007518>. Citado na página 14.
- PERES, P. S. Alinhamento múltiplo de seqüências através de técnicas de agrupamento. Universidade Federal do Amazonas, 2006. Citado na página 53.
- PICOULT-NEWBERG, L. *et al.* Mining snps from est databases. **Genome research**, Cold Spring Harbor Lab, v. 9, n. 2, p. 167–174, 1999. Citado na página 43.
- PIGNATTI, M. G. Saúde e ambiente: as doenças emergentes no brasil. **Ambiente & sociedade**, SciELO Brasil, v. 7, p. 133–147, 2004. Citado na página 37.
- PIMENTEL, A. L. d. S. **Revisão da análise filogenética dos segmentos L, M e S do vírus Oropouche**. Dissertação (B.S. thesis) — Universidade Federal do Rio Grande do Norte, 2019. Citado na página 64.
- PINEDA-PEÑA, A.-C. *et al.* Automated subtyping of hiv-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new rega version 3 and seven other tools. **Infection, genetics and evolution**, Elsevier, v. 19, p. 337–348, 2013. Citado na página 50.
- PINHEIRO, F. *et al.* Tratado de infectologia. **Tratado de Infectologia. Sao Paulo, Brazil:[sn]**, p. 169–180, 1996. Citado na página 32.
- PINHEIRO, F. P. *et al.* Transmission of oropouche virus from man to hamster by the midge culicoides paraensis. **Science**, American Association for the Advancement of Science, v. 215, n. 4537, p. 1251–1253, 1982. Citado na página 30.
- PINHEIRO, F. P.; ROSA, A. P. Travassos da; ROSA, J. F. Travassos da. Oropouche virus. i. a review of clinical, epidemiological, and ecological findings. 1981. Citado nas páginas 29 e 30.
- RAGHUPATHI, W.; RAGHUPATHI, V. Big data analytics in healthcare: promise and potential. **Health information science and systems**, Springer, v. 2, p. 1–10, 2014. Citado na página 38.

- RIDDER, D. de; RIDDER, J. D.; REINDERS, M. J. Pattern recognition in bioinformatics. **Briefings in bioinformatics**, Oxford University Press, v. 14, n. 5, p. 633–647, 2013. Citado na página 33.
- ROMERO-ALVAREZ, D.; ESCOBAR, L. E. Oropouche fever, an emergent disease from the americas. **Microbes and Infection**, Elsevier, v. 20, n. 3, p. 135–146, 2018. Citado nas páginas 27, 28, 29 e 48.
- ROSA, A. P. A. Travassos da *et al.* Advances in virus research. **Academic Press**, MDPI, v. 98, p. 269–299, 2017. Citado na página 14.
- ROSA, A. P. T. da; VASCONCELOS, P. F.; ROSA, J. F. T. da. **An overview of arbovirology in Brazil and neighbouring countries**. [S.l.]: Citeseer, 1998. Citado nas páginas 28 e 30.
- SAGIROGLU, S.; SINANC, D. Big data: A review. In: IEEE. **2013 international conference on collaboration technologies and systems (CTS)**. [S.l.], 2013. p. 42–47. Citado na página 39.
- 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. Citado na página 55.
- SAKKAS, H. *et al.* Oropouche fever: a review. **Viruses**, MDPI, v. 10, n. 4, p. 175, 2018. Citado nas páginas 14 e 30.
- SALEMI, M.; VANDAMME, A.-M. **The phylogenetic handbook: a practical approach to DNA and protein phylogeny**. [S.l.]: Cambridge University Press, 2003. Citado nas páginas 63 e 66.
- SAMUEL, P. P.; TYAGI, B. Diagnostic methods for detection & isolation of dengue viruses from vector mosquitoes. **Indian journal of medical research**, INDIAN COUNCIL OF MEDICAL RESEARCH, v. 123, n. 5, p. 615, 2006. Citado na página 32.
- SanarMed. **Linfócitos T**. 2023. <https://sanarmed.com/linfocitos-t/#:~:text=As%20c%C3%A9lulas%20T%20que%20medeiam,parte%20da%20resposta%20imunol%C3%B3gica%20humoral>. Citado na página 26.
- SANTOS, F. D. M. Falta de saneamento básico é considerado como um dos vilões da proliferação de aedes aegypti. 2016. Disponível em: <https://www.gov.br/fundacentro/pt-br/comunicacao/noticias/noticias/2016/4/falta-de-saneamento-basico-e-considerado-como-um-dos-viloes-da-proliferacao-de-aedes-aegypti>. Citado na página 13.
- SANTOS, N. S. d. O.; ROMANOS, M. T. V.; WIGG, M. D. **Virologia humana**. [S.l.]: Grupo Gen-Guanabara Koogan, 2015. Citado na página 30.
- SANTOS, V. S. dos. **Multiplicação dos Vírus**. 2024. <https://www.preparaenem.com/biologia/multiplicacao-dos-virus.htm>. Citado nas páginas 22 e 23.
- SCHADT, E. E. *et al.* Computational solutions to large-scale data management and analysis. **Nature reviews genetics**, Nature Publishing Group UK London, v. 11, n. 9, p. 647–657, 2010. Citado na página 17.

SCHATZMAYR, H. G. Viroses emergentes e reemergentes. **Cadernos de Saúde Pública**, SciELO Brasil, v. 17, p. S209–S213, 2001. Citado na página 36.

SCHMIDT, H. A. *et al.* Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. **Bioinformatics**, Oxford University Press, v. 18, n. 3, p. 502–504, 2002. Citado na página 55.

SCHMIDT, N. J. Diagnostic procedures for viral, rickettsial and chlamydial infections. **Enteroviruses and reoviruses**, American Public Health Association, p. 513–579, 1995. Citado na página 32.

Sensi Saúde. **O que é Infecção Viral?** 2024. Disponível em: <https://sensisaude.com.br/glossario/o-que-e-infeccao-viral/>. Citado na página 21.

SHOPE, R. E. The discovery of arbovirus diseases a. **Annals of the New York Academy of Sciences**, Wiley Online Library, v. 740, n. 1, p. 138–145, 1994. Citado na página 27.

SILVA, A. E. A. da. **Uma abordagem multi-objetivo e multimodal para reconstrução de Árvores filogenéticas**. Tese (Doutorado) — UNIVERSIDADE ESTADUAL DE CAMPINAS, 2007. Citado na página 55.

SILVA, D. M. F. D. *et al.* Detection of arboviruses in *Aedes aegypti* through transovarian analysis: A study in Goiânia, Goiás. **Revista da Sociedade Brasileira de Medicina Tropical**, Sociedade Brasileira de Medicina Tropical - SBMT, v. 57, p. e00400–2023, 2024. ISSN 0037-8682. Disponível em: <https://doi.org/10.1590/0037-8682-0280-2023>. Citado na página 14.

SILVA-NUNES, M. d. *et al.* The acre project: the epidemiology of malaria and arthropod-borne virus infections in a rural amazonian population. **Cadernos de Saúde Pública**, SciELO Public Health, v. 22, p. 1325–1334, 2006. Citado na página 32.

SILVA, R. C. C. da; ALVES, M. C. S. O uso de ferramentas de bioinformática para análise de dados genéticos: uma revisão. **Scientific Electronic Archives**, v. 17, n. 1, 2024. Citado nas páginas 34 e 35.

SIMIÃO, A. R. *et al.* A major chikungunya epidemic with high mortality in northeastern Brazil. **Revista da Sociedade Brasileira de Medicina Tropical**, SciELO Brasil, v. 52, p. e20190266, 2019. Citado nas páginas 13 e 26.

SOARES, T. M.; MENDONÇA, M. C. M. Construção de um modelo de regressão hierárquico para os dados do Simave-2000. **Pesquisa Operacional**, SciELO Brasil, v. 23, p. 421–441, 2003. Citado na página 56.

SOUZA, C. Wanderley de. **Doenças negligenciadas**. Rio de Janeiro: Academia Brasileira de Ciências, 2010. (Ciência e tecnologia para o desenvolvimento nacional. Estudos estratégicos). ISBN 978-85-85761-30-1. Citado na página 13.

SOUZA, E.; MINAYO, M. d. S.; MINAYO, M. d. S. **Os muitos Brasis: saúde e população na década de 80**. 1995. Citado na página 37.

SOUZA, G. B. S. *et al.* Novas abordagens para integração de bancos de dados e desenvolvimento de ferramentas bioinformáticas para estudos de genética de populações. Universidade Federal de Minas Gerais, 2014. Citado na página 42.

- SPENCER, D. H. *et al.* Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. **The Journal of Molecular Diagnostics**, Elsevier, v. 16, n. 1, p. 75–88, 2014. Citado na página 36.
- STEIGER, D. B. M.; RITCHIE, S. A.; LAURANCE, S. G. Mosquito communities and disease risk influenced by land use change and seasonality in the australian tropics. **Parasites & vectors**, Springer, v. 9, p. 1–13, 2016. Citado na página 36.
- STRIMMER, K.; HAESELER, A. V. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 94, n. 13, p. 6815–6819, 1997. Citado nas páginas 55 e 63.
- SUTHERLAND, J.; SCHWABER, K. The scrum guide. **The definitive guide to scrum: The rules of the game. Scrum. org**, v. 268, p. 19, 2013. Citado na página 57.
- TAJUDEEN, Y. A. *et al.* Emerging arboviruses of public health concern in africa: priorities for future research and control strategies. **Challenges**, MDPI, v. 13, n. 2, p. 60, 2022. Citado na página 14.
- THOMPSON, I.; COX, A. Don't imitate, innovate. **Supply Management**, v. 2, n. 22, p. 40–42, 1997. Citado na página 61.
- THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic acids research**, Oxford university press, v. 22, n. 22, p. 4673–4680, 1994. Citado nas páginas 53, 63 e 66.
- THORISSON, G. A.; MUILU, J.; BROOKES, A. J. Genotype–phenotype databases: challenges and solutions for the post-genomic era. **Nature Reviews Genetics**, Nature Publishing Group UK London, v. 10, n. 1, p. 9–18, 2009. Citado na página 42.
- THORNE, J. L. Models of protein sequence evolution and their applications. **Current opinion in genetics & development**, Elsevier, v. 10, n. 6, p. 602–605, 2000. Citado na página 40.
- TIBSHIRANI, R. J.; EFRON, B. An introduction to the bootstrap. **Monographs on statistics and applied probability**, v. 57, n. 1, p. 1–436, 1993. Citado na página 57.
- Toda Matéria. **Vírus**. 2024. <https://www.todamateria.com.br/virus/>. Citado na página 21.
- TOMMASO, P. D. *et al.* Nextflow enables reproducible computational workflows. **Nature biotechnology**, Nature Publishing Group, v. 35, n. 4, p. 316–319, 2017. Citado na página 41.
- TOUCHMAN, J. Comparative genomics. **Nature Education Knowledge**, v. 3, n. 10, p. 13, 2010. Citado na página 46.
- TREES, R. P. The neighbor-joining method: a new method for. **Mol Biol Evol**, v. 4, n. 4, p. 406–425, 1987. Citado na página 40.

- Tribunal de Justiça do Distrito Federal e Territórios. **Gripe H1N1: Sintomas e Prevenção**. 2024. Disponível em: [https://www.tjdft.jus.br/informacoes/programas-projetos-e-acoes/pro-vida/dicas-de-saude/pilulas-de-saude/gripe-h1n1-sintomas-e-prevencao#:~:text=De%20acordo%20com%20o%20Minist%C3%A9rio,H1N1\)%2C%20sendo%20764%20%C3%B3bitos](https://www.tjdft.jus.br/informacoes/programas-projetos-e-acoes/pro-vida/dicas-de-saude/pilulas-de-saude/gripe-h1n1-sintomas-e-prevencao#:~:text=De%20acordo%20com%20o%20Minist%C3%A9rio,H1N1)%2C%20sendo%20764%20%C3%B3bitos). Citado na página 21.
- VASCONCELOS, H. B. *et al.* Oropouche fever epidemic in northern brazil: epidemiology and molecular characterization of isolates. **Journal of Clinical Virology**, Elsevier, v. 44, n. 2, p. 129–133, 2009. Citado na página 14.
- VASCONCELOS, P. F. *et al.* Arboviruses pathogenic for man in brazil. In: **An overview of arbovirology in Brazil and neighbouring countries**. [S.l.: s.n.], 1998. p. 72–99. Citado na página 32.
- VASCONCELOS, P. F. *et al.* Inadequate management of natural ecosystem in the brazilian amazon region results in the emergence and reemergence of arboviruses. **Cadernos de Saúde Pública**, SciELO Public Health, v. 17, p. S155–S164, 2001. Citado na página 32.
- VASCONCELOS, S.; KOVALESKI, F.; JUNIOR, C. T. Doenças negligenciadas: revisão da literatura sobre as intervenções propostas. **Saúde & Transformação Social / Health & Social Change**, v. 6, n. 2, p. 114–131, 2016. Disponível em: <https://incubadora.periodicos.ufsc.br/index.php/saudeetransformacao/article/view/3714>. Citado na página 13.
- VIANA, C. J. M. Aspectos de genômica comparativa. 2006. Citado na página 47.
- VIEIRA, S. **Introdução à bioestatística**. [S.l.]: Elsevier Brasil, 1997. Citado na página 56.
- WALJEE, A. K.; HIGGINS, P. D. Machine learning in medicine: a primer for physicians. **Official journal of the American College of Gastroenterology | ACG**, LWW, v. 105, n. 6, p. 1224–1226, 2010. Citado na página 39.
- WALSH, C. E.; ROBERT, M. A.; CHRISTOFFERSON, R. C. Observational characterization of the ecological and environmental features associated with the presence of oropouche virus and the primary vector *Culiseta paraensis*: Data synthesis and systematic review. **Tropical Medicine and Infectious Disease**, MDPI, v. 6, n. 3, p. 143, 2021. Citado na página 37.
- Web Toolkit. **Java Web Toolkit (JWT)**. 2024. Disponível em: <https://www.webtoolkit.eu/jwt/>. Citado na página 63.
- WEBB, C. O. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. **The American Naturalist**, The University of Chicago Press, v. 156, n. 2, p. 145–155, 2000. Citado na página 54.
- WHAT is Jupyter? — Jupyter Documentation 4.1.1 alpha documentation. [S.l.: s.n.], 2015. https://docs.jupyter.org/en/latest/what_is_jupyter.html. Citado na página 60.
- WORLEY, K. Sequence analysis in a nutshell: A guide to common tools and databases. **The American Journal of Human Genetics**, Elsevier, v. 73, n. 5, p. 1213, 2003. Citado na página 58.

- YANG, H.-c. *et al.* Map-reduce-merge: simplified relational data processing on large clusters. In: **Proceedings of the 2007 ACM SIGMOD international conference on Management of data**. [S.l.: s.n.], 2007. p. 1029–1040. Citado na página 44.
- YAO, Z. *et al.* Evaluation of variant calling tools for large plant genome re-sequencing. **BMC bioinformatics**, Springer, v. 21, p. 1–16, 2020. Citado na página 36.
- ZHANG, Y. *et al.* Oropouche virus: A neglected global arboviral threat. **Virus Research**, Elsevier, v. 341, p. 199318, 2024. Citado na página 15.
- ZHOU, L. *et al.* Machine learning on big data: Opportunities and challenges. **Neurocomputing**, Elsevier, v. 237, p. 350–361, 2017. Citado na página 39.
- ZOLET, A. C. T. *et al.* Marcadores moleculares na era genômica: metodologias e aplicações. Sociedade Brasileira de Genética, 2017. Citado na página 44.