FIOCRUZ

FUNDAÇÃO OSWALDO CRUZ
INSTITUTO GONÇALO MONIZ

Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa

TESE DE DOUTORADO

SAGA: SISTEMA DE ANÁLISE GENÔMICA DOS ARBOVÍRUS *ZIKA VIRUS, DENGUE VIRUS, CHIKUNGUNYA VIRUS E YELLOW FEVER VIRUS*.

JOSÉ IRAHE KASPRZYKOWSKI GONÇALVES

Salvador - Bahia
2020

FUNDAÇÃO OSWALDO CRUZ
INSTITUTO GONÇALO MONIZ

Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa

SAGA: SISTEMA DE ANÁLISE GENÔMICA DOS ARBOVÍRUS *ZIKA VIRUS, DENGUE VIRUS, CHIKUNGUNYA VIRUS E YELLOW FEVER VIRUS*.

JOSÉ IRAHE KASPRZYKOWSKI GONÇALVES

Tese apresentada ao Curso de Pós-Graduação em Biotecnologia em Saúde e Medicina Investigativa para a obtenção do grau de Doutor.

Orientador: Prof. Dr. Artur Trancoso Lopo de Queiroz

Salvador - Bahia
2020

SAGA: SISTEMA DE ANÁLISE GENÔMICA DOS ARBOVÍRUS *ZIKA VIRUS, DENGUE VIRUS, CHIKUNGUNYA VIRUS E YELLOW FEVER VIRUS*.

JOSÉ IRAHE KASPRZYKOWSKI GONÇALVES

Folha de Aprovação

Comissão Examinadora

_____	_____
Dr. Artur Trancoso L. de Queiroz	Dr. Luciano Kalabric Silva
FIOCRUZ/CPqGM	FIOCRUZ/IGM
_____	_____
Dr. Pablo Ivan Pereira Ramos	Dra. Thessika Hialla Almeida Araújo
CIDACS	EBMSP

FONTES DE FINANCIAMENTO

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

AGRADECIMENTOS

Inicialmente ao Grande Arquiteto do Universo por continuamente iluminar e guiar os caminhos de todos nós em busca da justiça e perfeição, nos inspirando com coragem, justiça temperança e prudência para vencer os diversos percalços no caminho. Não teria como compor um agradecimento sem agradecer ao meu PAI Tuca que me acompanha desde menino e que formou em mim um profissional mais responsável, um educador exigente, um formador inspirado, um pai mais dedicado, sem dúvidas um cidadão consciente ~~e espero eu, que um Doutor~~. Também não tenho como não agradecer à minha esposa Carol que está comigo a cada dia, estimulando, apoiando, ajudando e na medida do possível, incentivando para que fossem vencidos os percalços do caminho. Preciso agradecer ainda ao meu filho Davi, que mesmo sem saber, é a maior inspiração da minha vida. À minha mãe, que sempre sabe o que dizer, mesmo em silêncio (mães tem esse super poder). Á Biblioteca de Ciências Biomédias Eurydice Pires de Sant'Anna e mais precisamente à Ana Maria Fiscina Vaz Sampaio pelo apoio nos momentos finais da construção deste. Seria injusto citar mais nomes ou menos nomes aqui, pois não caberiam todos os que ajudaram e contribuíram de maneira direta ou indireta para a conclusão deste trabalho. Assim, deixo aqui o meu Muito Obrigado.

O sucesso nasce do querer, da determinação e persistência em se chegar a um objetivo. Mesmo não atingindo o alvo, quem busca e vence obstáculos, no mínimo fará coisas admiráveis.

José de Alencar

GONÇALVES, José Irahe Kasprzykowski. SAGA: Sistema de análise genômica dos arbovírus *Zika Virus, Dengue Virus, Chikungunya Virus e Yellow Fever Virus*. 2020. 91 f. Tese (Doutorado em Biotecnologia em Saúde e Medicina Investigativa) – Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, 2020.

RESUMO

INTRODUÇÃO: Doenças causadas por arbovírus (*ARthropod-BOrne VIRUS*) representam um grave problema de saúde pública. Com a expansão do vetor *Aedes aegypti* há o agravante crescimento dos casos de arboviroses no país nos últimos 10 anos. Para melhor abordar esse problema, é preciso obter o máximo de informações sobre os agentes etiológicos associados. Informações como classificação, mapeamento genômico e resposta imune. **OBJETIVO:** Assim, o objetivo deste trabalho é desenvolver uma plataforma de coleta e análise automáticas e contínuas de informações genômicas dos arbovírus *Zika Virus, Dengue Virus, Chikungunya Virus e Yellow Fever Virus.* **MATERIAL E MÉTODOS**: Para tal, foram importadas sequências nucleotídicas dos bancos de dados internacionais (*GenBank, DDBJ* e *EMBL).* Estas sequências foram classificadas quanto aos subtipos disponíveis na literatura para cada vírus, e mapeada ainda quanto a sua região de correspondência com a anotação do genoma completo do organismo. A partir daí, foram importados todos os epítopos do *IEDB*(*Immune Epitope Database*) e mapeados nas regiões codificantes das sequências. Foi então construido uma interface de acesso WEB onde é possível verificar secções semanais de dados. **RESULTADOS**: Do corte inicial foi possível obter informações como frequência de subtipo. Onde os mais frequentes são o subtipo tanzaniano, com presença de 73,10% nas sequências de *Chikungunya virus*; o Subtipo 1 com 34,53% de frequência nas sequências de Dengue vírus; o subtipo peruano com 59,18% nas sequências do vírus da febre amarela e o subtipo do Camboja com 51,35% das sequências do *Zika virus*. Em relação ao mapeamento genômico os dados mostram que a região mais bem representada no conjunto de dados é aquela que codifica a proteína do envelope viral. O mapeamento dos epítopos mostraram que o epítopo "AMTDTTPFGQQRVFK" é o mais frequente no vírus da dengue enquanto o "STKDNFNVYKATRPYLAHC" é mais frequente no *Chikungunya*. Já no vírus da zika, o epítopo mais frequentemente observado é o "DQRGSGQVVTYALNT" e, para o Vírus da Febre Amarela, o epítopo "DRDFIEGVHGGTWVS". CONCLUSÃO: Assim, a plataforma desenvolvida representa uma importante ferramenta de suporte a estudos sobre os agentes etiológicos em questão. Uma vez que gera continuamente informações relevantes, além de organizar, armazenar e disponibilizar

secções personalizadas de dados. Estas informações reduzem o tempo necessário para realizar estudos sobre estes organismos, acelerando assim os processos de desenvolvimento de vacinas e técnicas diagnósticas.
Palavras-chaves: Virologia; Genômica; Bioinformática; Zika; Chikungunya; Dengue; Febre Amarela

GONÇALVES, José Irahe Kasprzykowski. SAGA:Genomic Analysis System for *Zika Virus, Dengue Virus, Chikungunya Virus e Yellow Fever Virus*. 2020. 91 f. Tese (Doutorado em Biotecnologia em Saúde e Medicina Investigativa) – Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, 2020.

ABSTRACT

INTRODUCTION: Diseases caused by arboviruses (ARthropod-BOrne VIRUS) represent a serious public health problem. With the expansion of the Aedes aegypti vector, there has been an aggravating growth in cases of arboviruses in the country in the last 10 years. To better address this problem, it is necessary to obtain as much information about the associated etiological agents. Information such as classification, genomic mapping and immune response. **OBJECTIVE:** Thus, the objective of this work is to develop a platform for automatic and continuous collection and analysis of genomic information of the arboviruses Zika Virus, Dengue Virus, Chikungunya Virus e Yellow Fever Virus. **MATERIAL AND METHODS**: To this end, genomic sequences were imported from international databases (GenBank, DDBJ and EMBL). These sequences were classified according to the subtypes available in the literature for each virus, and also mapped according to their region of correspondence with the annotation of the organism complete genome. From there, all epitopes from the IEDB (Immune Epitope Database) were imported and mapped in the coding regions of the sequences. A WEB interface was then built where it is possible to check weekly sections of data. **RESULTS:** From the initial cut it was possible to obtain information such as subtype frequency. Where the most frequent are the Tanzanian subtype, with a presence of 73.10% in the sequences of Chikungunya virus; Subtype 1 with a frequency of 34.53% in Dengue virus sequences; the Peruvian subtype with 59.18% of the yellow fever virus sequences and the Cambodia subtype with 51.35% of the Zika virus sequences. Regarding genomic mapping, the data show that the region best represented in the dataset is the one that encodes the viral envelope protein. Epitope mapping showed that the "AMTDTTPFGQQRVFK" epitope is the most frequent in the dengue virus while the "STKDNFNVYKATRPYLAHC" is the most frequent in Chikungunya. In the Zika virus, the most frequently observed epitope is "DQRGSGQVVTYALNT" and for Yellow Fever Virus, the epitope "DRDFIEGVHGGTWVS". **CONCLUSION**: Thus, the developed platform represents an important support tool for studies on the etiological agents in question. As it continuously generates relevant information, in addition to organizing, storing and making available personalized sections of data. This information reduces the time needed to carry out studies on these organisms, thus speeding up the vaccine development processes and diagnostic techniques.

Keywords: Virology; Genômics; Bioinformathics; Zika; Chikungunya; Dengue; Yellow Fever

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Arbovirus	*ARthropod-BOrne VIRUS*
DDBJ	Banco de Dados de DNA do Japão
EMBL	European Molecular Biology Laboratory
IEDB	*Immune Epitope Database*
WEB	World Wide *Web*
CKV	*Chikungunya Virus*
DENV	*Dengue Virus*
ZKV	*Zika Virus*
YFV	*Yellow Fever Virus*
HIV	*Human Immunodeficiency Virus*
HCV	*Hepatitis C Virus*
HFV	*Hemorragic Fever Virus*
SAGA	Sistema de Análise Genômica de Arbovírus
CD8	Linfócitos T Citotóxicos
CD4	Linfócitos T Auxiliares
INSDB	*International Sequence Database*
OMS	Organização Mundial da Saúde
ORF	*Open Read Frame*
ECSA	*East–Central–South–African*
WA	*West African*
IOL	*Indian Ocean Lineage*
Indels	*Insertions and/or deletions*
HMM	*Hidden Markov Models*
API	*Application Programming Interface*
HTTP	*Hyper Text Transfer Protocol*
HTTPS	*Secure Hyper Text Transfer Protocol*
HTML	*Hypertext Markup Language*
REST	*Representational State Transfer*
JSON	*Javascript Object Notation*
VSDBM	*Viral Sequence Database Manager*
GI	*Global Identificator*

DER	Diagrama de Entidades e Relacionamentos
NCBI	*National Center for Biotechnology Information*
XML	Extensible Markup Language
FTP	File Transfer Protocol

SUMÁRIO

1 INTRODUÇÃO

Com o recente aumento dos casos de doenças virais transmitidas pelo mosquito *Aedes aegypti*, é reforçada a necessidade de ferramentas que auxiliem no combate às doenças (MARCONDES, 2016; DONATELI, 2019). O vírus causador da dengue, bem como outros arbovírus, vêm se espalhando em escalas sem precedentes no mundo (DONATELI, 2019; TAMI, 2016).

No Brasil, de 2007 a 2009 foram reportados 2.756.622 casos de dengue (WHO,2009). Segundo o boletim epidemiológico do Ministério da Saúde, em 2015 foram registrados mais de um milhão e meio de casos da doença (BRASIL. MINISTÉRIO DA SAÚDE, 2016). Apenas no estado do Rio de Janeiro foram reportados 69.000 casos, o que representa uma incidência de 417 pessoas infectadas a cada 100.000 habitantes (WILSON, 2016). Esses dados demonstram o aumento expressivo de casos dessa doença em relação a anos anteriores.

Além dessa doença, suspeitas de casos da febre chikungunya vêm sendo notificados no país desde 2014 (BRASIL. MINISTÉRIO DA SAÚDE, 2016), com 2.772 casos confirmados (HONÓRIO, 2015). Apenas no primeiro trimestre do ano de 2015, foram confirmados 1.513 casos de infecção pelo *Chikungunya virus* (CKV) (HONÓRIO, 2015). Ainda em 2015, foi observada a introdução do *Zika virus* (ZKV) cocirculando com outros arbovírus, como o *Dengue vírus* (DENV) e CKV, em diversos estados do nordeste brasileiro (GUABIRABA, 2014). Já em 2017, foi observado um surto nos casos de febre amarela, causada pelo *Yellow Fever Virus* (YFV), em território nacional (VALESKA ROSSETTO, 2017). Esta situação é reportada pelo ministério da saúde como crítica, com 901 suspeitas de casos da doença. Destes, mais de 70% ainda estão em investigação, assim como 60% das mortes (BRASIL. MINISTÉRIO DA SAÚDE, 2017).

O ZKV, DENV, CKV e YFV são transmitidos pelo mesmo vetor, o *Aedes aegypti* (OLSON, 1981; CAMPOS,2015; SUAYA,2009; HOLMES, 2003; SAM, 2012; SIMPSON, 1694) que no Brasil apresenta uma larga distribuição, ocasionada pela sua adaptação ao ambiente urbano e ao difícil controle. Esta dificuldade está principalmente associada à dispersão de lixo nas áreas urbanas, sendo um dos maiores desafios do controle do vetor.

Estas quatro arboviroses transmitidas pelo mosquito *Aedes* representam um sério problema de saúde pública, além de um grave problema econômico (GUABIRABA, 2014), uma vez que os quadros sintomáticos geralmente causam a incapacitação profissional do paciente (SISSOKO, 2008). Além disso, estima-se que nas Américas o custo do combate apenas à Dengue, por exemplo, chega à 580 milhões de dólares por ano (SUAYA, 2009).

Estes vírus além de serem facilmente transmissíveis, também foram associados com patologias crônicas como a microcefalia (EICKMANN, 2016), problemas oftalmológicos (WEAVER, 2016), ósseos e síndrome de Guillain-Barré (LEBRUN, 2009).

Devido a indistinguibilidade dos sintomas clínicos da dengue e da febre chikungunya (REZZA, 2019), os métodos de diagnóstico molecular são considerados mais eficazes que os convencionais para o diagnóstico das doenças (ÁLVAREZ-ARGÜELLES, 2019) principalmente nos casos de coinfecção. Na Coreia do Sul, por exemplo, pacientes foram diagnosticados com CKV, quando anteriormente se acreditava possuir DENV (CHA, 2013). Estas observações ressaltam a necessidade do de métodos diagnóstico mais adequados para estas etiologias no Brasil, onde estes arbovírus são prevalentes.

O desenvolvimento de métodos de diagnóstico molecular só é possível devido à identificação dos genomas virais (CHA, 2013). Para tal, são necessárias uma série de informações sobre as cepas contidas nos conjuntos de dados que auxiliam na identificação das características funcionais dos genomas, além da identificação das variações genotípicas (KUNO, 2007; SETTE, 2007) extremamente importantes para desenvolvimento de métodos de diagnóstico acurados. Além disso, esta identificação é essencial para o desenvolvimento de métodos diagnósticos e vacinas, uma vez que subtipos diferentes podem apresentar quadro clínico e resposta imune diferente (CUNHA, 2017; HASAN, 2016).

A distribuição das variações desses agentes etiológicos, além de sua heterogeneidade genotípica, ressalta a necessidade da criação e aprimoramento de ferramentas de análise de sua estrutura genômica e pós traducional buscando o aprimoramento da vigilância epidemiológica (CAMPOS, 2015).

Este tipo de abordagem já é utilizado para os agentes etiológicos responsáveis pelas principais epidemias e pandemias. O HIV (Vírus da Imunodeficiência Humada, do inglês *Human Immunodeficiency Virus*), por exemplo, possui uma plataforma específica para análise de sequências, genômicas e sorotipos conhecida como *HIV Databases* (LEITNER, 2017). Assim ocorre com o HCV (Vírus da Hepatite C do inglês *Hepatitis C virus*), Ebola, HFV (Vírus da Febre Hemorrágica do inglês *Hemorragic Fever virus)* e outros vírus (KUIKEN, 2005; KUIKEN, 2012).

Entretanto, estas ferramentas possuem uma limitação em comum: ausência de recursividade na coleta e análise dos dados. O *HCV Databases* por exemplo não é atualizado desde 2009 (KUIKEN, 2005), o que atrasa consideravelmente o processo de desenvolvimento de vacinas e evolução de tratamentos. Além disso, estes bancos de dados especializados

trabalham de forma isolada entre si, não compartilhando dados, tolhendo a gama de estudos e inferências possíveis.

Já os bancos de dados biológicos primários não possuem dados de mapeamento de epítopos, enquanto os especializados possuem estas informações para apenas uma fração do conjunto de dados. Essas características limitam o potencial de geração de informações relevantes por parte destes.

Apesar da latente necessidade de uma plataforma unificada de gerenciamento de informações genômicas dos arbovírus supracitados, esta não está ainda disponível. Consequentemente a maioria dos dados referentes a estes agentes etiológicos está disponível apenas em bancos de dados primários, de maneira esparsa, não organizada e em sua maioria, não curada. Além disso, os bancos de dados primários não oferecem informações sobre subtipagem, mapeamento nas regiões do genoma completo ou mesmo mapeamento de epítopos.

Essas informações auxiliam diretamente para o desenvolvimento de métodos preventivos e diagnósticos para as etiologias associadas (PRATHEEK, 2015). Outrossim, são importantes para a identificação e rastreamento de mutações que podem levar à falha diagnóstica (CHA, 2013). Desta forma, estes dados podem ainda auxiliar no controle da progressão das pandemias, além de apoiar o desenvolvimento de vacinas.

Este trabalho consistiu no desenvolvimento do Sistema de Análise Genômica de Arbovírus (SAGA). Durante o processo, foi necessário o desenvolvimento de algoritmos de comparação e mapeamento de sequências nucleotídicas, para identificar e classificar os dados relacionados com esses vírus. Assim, foi possível organizar e classificar as informações genéticas destes organismos, fornecendo uma plataforma robusta e continuamente atualizada de informações úteis para os processos de desenvolvimento de diagnósticos e vacinas (DUFFY, 2009; FILIPE, 1973).

Também foram desenvolvidos protocolos e algoritmos para coleta e mapeamento de epítopos imunogênicos nos conjuntos de dados, possibilitando identificar suas frequências, assim como identificação de epítopos exclusivos entre os organismos/subtipos.

O desenvolvimento visou fornecer à ferramenta, a capacidade de realizar estas coletas e análises de forma recursiva e disponibilizar os dados resultantes sob demanda. Isso auxilia no entendimento dessas infecções e da interação com o hospedeiro, além de fornecer importantes informações de base para estudos sobre estes organismos.

Desta forma, a ferramenta pode apoiar o desenvolvimento de novas estratégias de diagnóstico e prevenção para as doenças associadas. Além disso, a plataforma busca, importa

e mapeia epítopos no conjunto de dados. Estes epítopos são importados recursivamente diretamente do *Immune Epitope Database* (IEDB) (VITA, 2015) e constituem, até então em mais de 790 mil epítopos importados e posteriormente mapeados.

Apesar de computacionalmente complexo, o processo de busca e mapeamento de epítopos contribui significativamente para a compreensão da resposta do vírus ao sistema imune (GIMENEZ, 2016), facilitando o desenvolvimento de vacinas e métodos diagnósticos eficazes inclusive nos casos de coinfecção.

Além desses, os resultados do mapeamento de epítopos auxiliam em estudos baseados em resposta imune. Esse tipo de dado vem sendo utilizados para a identificação e posterior predição de epítopos imunogênicos que estimulem a resposta à células T CD8+. No entanto, apesar de auxiliar nos trabalhos de desenvolvimento, mais estudos são necessários para a avaliação desta resposta (PRATHEEK, 2015). Desta forma, a plataforma pode servir como repositório dessas informações com uma interface de acesso rápido e facilitado.

Portanto, esperamos facilitar o acesso a informações dessas arboviroses com essa plataforma, centralizando o processo de vigilância epidemiológica. Dessa maneira a plataforma atuaria de forma dinâmica no combate a etiologias causadas por arbovírus, também influenciando diretamente no processo de desenvolvimento de tratamentos e vacinas mais eficazes.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Desenvolver uma plataforma unificada para análise genômica e armazenamento de informações sobre *Zika virus*, *Dengue virus*, *Chikungunya virus* e *Yellow Fever Virus*.

2.2 OBJETIVOS ESPECÍFICOS

- Desenvolver um modelo de coleta e armazenamento contínuo de sequências nucleotídicas dos agentes etiológicos em questão a partir do *INSDB(International Sequence Database);*
- Desenvolver processo de Alinhamento e mapeamento das sequências obtidas contra o genoma completo do organismo;
- Desenvolver processo de classificação de sequências de arbovírus segundo os genótipos conhecidos;
- Desenvolver um modelo de coleta e armazenamento contínuo de epítopos a partir do *IEDB;*
- Implementar algoritmo de busca de epítopos em regiões codificantes das sequências obtidas;
- Desenvolver uma interface de acesso WEB para acesso aos resultados de análises e ao conjunto de dados propriamente dito.

3 REVISÃO DA LITERATURA

3.1 O VETOR

Os arbovírus DENV, CKV, ZKV e YFV representam importante problema de saúde pública e são transmitidos pelo vetor *Aedes aegypti* (POWELL, 2013). Este vetor pertencente à família *Culidae* e é o principal responsável pelas epidemias e arbovírus nas Américas, sendo considerado o mais eficiente para transmissão de arbovírus (COFFEY, 2014). Este vetor, é prevalente em regiões tropicais e subtropicais do globo (LEDERMANN, 2014; KRAEMER, 2015; GAFFIGAN, 2016).

Estes insetos estão bem adaptados ao convívio humano e ao ambiente urbano, o que representa um dos mais importantes fenômenos na entomologia médica (FUKUTANI, 2018). Este fato se dá também pela dificuldade no controle da reprodução desse vetor, ausência de políticas eficazes de saneamento, tratamento de resíduos e saúde coletiva (SUAYA, 2009). Esta dificuldade no controle populacional do vetor leva à sua alta prevalência, facilitando a manutenção do ciclo transmissor urbano das arboviroses (FUKUTANI, 2018). No ambiente urbano, esta prevalência gera uma séria preocupação, pois pode representar um novo cenário no processo evolutivo das arboviroses. Uma vez que foi observado que este vetor pode transmitir simultaneamente uma série de arbovírus (VALDERRAMA, 2017; ROTH, 2014). Apesar de o completo impacto da cotransmissão e coinfecção pelos agentes etiológicos responsáveis pelas arboviroses ainda não ser definido (RÜCKERT, 2017; GUANCHE GARCELL, 2020), é claro o desafio de processar, compilar e analisar essas informações (ROTH, 2018).

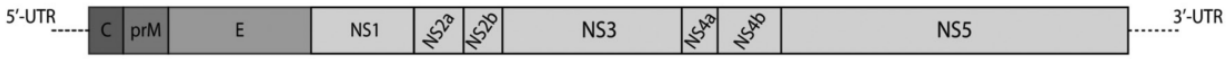
Considerando a indistinguibilidade sintomática apresentados por algumas das arboviroses (GUANCHE GARCELL, 2020), aliado a possibilidade de o vetor transmitir simultaneamente mais de um agente etiológico, a dificuldade no diagnóstico das infecções é elevada (RÜCKERT, 2017; GUANCHE GARCELL, 2020) principalmente nos casos de coinfecção. Além disso, a interação do hospedeiro com a coinfecção de mais de um agente etiológico pode gerar modificações nos genomas, aumento de carga viral, alterações na patogênese e dificultando cada vez mais o controle das epidemias (RÜCKERT, 2017).

3.2 DENGUE VIRUS

Flavivírus da família *Flaviviridae,* o vírus causador da Dengue é um arbovírus endêmico nas regiões onde os vetores *Aedes* são mais prevalentes (regiões tropicais e subtropicais). Este agente etiológico apresenta 4 principais sorotipos genômicamente

definidos (DEN-1, DEN-2, DEN-3 e DEN-4) (LINARES, 2013; HALSTEAD, 1988; KURANE, 2007) que apresentam manifestações clínicas similares. Além destes, foi identificado um quinto subtipo deste agente etiológico, o DEN-5 (MUSTAFA, 2015). Entretanto, este subtipo ainda não apresenta relevante impacto na epidemia vigente, apesar de poder influenciar no desenvolvimento de vacinas (MUSTAFA, 2015). Dos 4 sorotipos mais prevalentes, o DEN-4 está associado mais predominantemente com a forma severa da doença, ou seja, acometimentos hemorrágicos (KUMARIA, 2010).

O genoma viral é composto por aproximadamente 10 mil pares de base, que são traduzidos em uma única poliproteína. Essa é clivada em três proteínas estruturais (Capsídeo (C), Membrana/Pré Membrana(prM/M e envelope(E)) e sete não estruturais (NS1, NS2A, NS2B, NS3, NS4A, NS4B e NS5) (DIAMOND, 2015). A organização destas estruturas no genoma viral pode ser visto na Figura 1.



Fonte: (SANTOS, 2015)

Figura 1 - Organização das estruturas gênicas no genoma do *DENV*

Tomando como base os acometimentos clínicos, a Organização Mundial da Saúde (OMS) divide a Dengue em dois principais tipos, descomplicado e severo (WHITEHORN, 2010; SORIA SEGARRA, 2015). Casos severos são associados diretamente à perda plasmática, hemorragia e danos aos órgãos internos, enquanto outros casos sem intercorrência destas complicações são considerados descomplicados (SORIA SEGARRA, 2018).

A infecção pelo vírus da Dengue apresenta-se com quadro clínico diverso, que varia de acometimento assintomático até o acometimento severo da febre hemorrágica (GUABIRABA, 2014; WHO, 2012) que levam à debilitação do paciente, gerando desconforto e até a morte. Estima-se que em 30% dos pacientes há um acometimento na mucosa oral, que está mais associado ao tipo severo da doença que ao tipo descomplicado (THOMAS, 2007).

Estima-se que mais de dois bilhões e meio de pessoas vivam em áreas endêmicas deste agente etiológico (GUZMAN, 2010). Desta forma, este vírus foi considerado pela OMS como um dos principais problemas de saúde pública nas regiões tropicais e subtropicais (WHO, 2007). Nas últimas décadas o número de casos de dengue cresceu 30%, associado

principalmente com o crescimento populacional exagerado, aquecimento global e controle ineficiente do vetor (WHO, 2007). Anualmente são observados cerca de 400 milhões de infecções com índices de mortalidade chegando a 20% em algumas áreas (LINARES, 2013).

A patogênese da dengue se inicia pela entrada do agente infeccioso no organismo do hospedeiro, através da picada do vetor. Depois há progressão infecciosa e por vezes sintomáticas onde ocorre resposta imune humoral, celular e inata. Após a fase aguda, a carga viral é reduzida no hospedeiro, ainda que este desenvolva a forma severa da etiologia (WHITEHORN, 2011). Os fenótipos mais perigosos da doença, se apresentam paradoxalmente durante esta fase de redução da carga viral (WHITEHORN, 2011) indicando que os fenótipos estão associados diretamente à resposta imune entre o hospedeiro e o agente infeccioso (GREEN, 2006).

Diversos estudos têm demonstrado que uma segunda infecção por um genótipo diferente deste agente infeccioso pode representar um aumento no risco para complicações clínicas (KURANE, 1994; KURANE, 2007; KYLE, 2008; SANGKAWIBHA, 1984). Por tanto, o processo de diagnóstico e classificação genotípica do agente infeccioso em questão influencia diretamente no curso de tratamento ideal.

Além de influenciar diretamente no processo de diagnóstico e no desenvolvimento de vacinas, a variabilidade genotípica apresentada por este arbovírus modifica a dinâmica da infecção e consequentemente seu quadro sintomático e suas potenciais sequelas (OLSON, 1981; GUABIRABA, 2014; WEAVER, 2012; VENTURA, 2016).

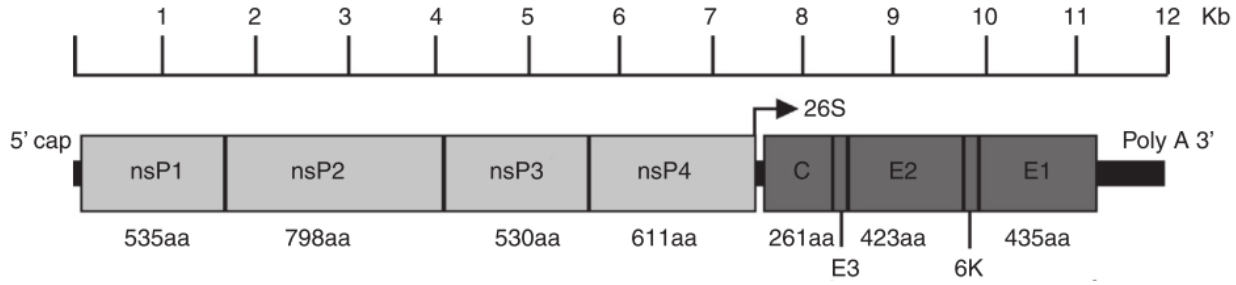
Apesar dos extensos esforços e custos (SUAYA, 2009) para o controle da infecção e tratamento da doença, não existe hoje uma vacina polivalente para os diversos subtipos do DENV (VAUGHN, 1996; BHAMARAPRAVATI, 2000; HADDOW, 2012). Isso se dá também pela heterogeneidade genotípica apresentada pelos subtipos deste vírus.

Desta forma, é importante aprimorar o conhecimento sobre o genoma viral. Com isso, será possível compreender melhor seus subtipos, patogênese, interação com o hospedeiro dentre outros aspectos.

3.3 *CHIKUNGUNYA VIRUS*

Descrito inicialmente em 1952, durante uma epidemia ocorrida em *Makonde* na Tanzânia (ROBINSON, 1957), o *Chikungunya Virus*(*CKV*) é um arbovírus transmitido pelos mosquitos *Aedes* (SETTE, 2017). Desde sua descrição inicial, este arbovírus infectou milhões de pessoas nos continentes africano, índico, europeu, asiático e americano (FARIA, 2016).

Este agente etiológico é da família *Togaviridae*, do gênero *Alphavirus* e pertence ao complexo antigênico de *Semliki Forest* (CLETON, 2012). Possui aproximadamente 12 mil pares de base em seu genoma completo, com dois suportes abertos de leitura 3'ORF(*Open Read Frame*) e 5'ORF, que codificam respectivamente as proteínas estruturais (C, E1, E2 e 6K) e proteínas não estruturais(nsP1, nsP2, nsP3 e nsP4) (VOLK, 2010; SCHWARTZ, 2010). A organização do genoma deste agente etiológico pode ser visto sumarizado na Figura 2.



Fonte: (GALÁN-HUERTA, 2015)

Figura 2 - Organização das estruturas gênicas no genoma do *CKV*

Diferentes genótipos deste organismo vêm sendo reportados durante as infecções nas diversas partes do globo(FARIA, 2016). Estes genótipos podem ser descritos como: *East–Central–South–African* (*ECSA*), correspondendo às cepas isoladas nas regiões leste, central e sul do continente africano; *West African* (*WA*), correspondendo a cepas isoladas no oeste africano, *Asian*, correspondendo às amostras isoladas no continente asiático; e *Indian Ocean Lineage* (*IOL*) correspondendo à cepas isoladas em 2004 durante a epidemia nesta região, como sendo descendente do *ECSA* (SAM, 2012; CAVALCANTI, 2017; NUNES, 2015; TSETSARKIN, 2007).

A infecção pelo agente etiológico apresenta inicialmente a reprodução no tecido epitelial, seguido pelo espalhamento através da corrente sanguínea para o fígado, juntas, músculos, linfonodos, baço e cérebro. O período de viremia em hospedeiros vertebrados pode durar de dois dias a até 10 dias após a infecção (KAM, 2009; PANNING, 2008). Durante e após este período, o paciente pode apresenta acometimento clínico assintomático, ou apresentar quadros que variam de formas mais suaves a condições severas e debilitantes (CUNHA, 2017). A quantidade relativa de pacientes que apresentam ou não cometimento assintomático ainda é desconhecida. Esta variação no quadro clínico pode ser associada à interação entre o agente etiológico e o sistema imune do hospedeiro (CUNHA, 2017).

O acometimento sintomático da infecção pelo CKV pode ser dividido em 3 fases: aguda, pós-aguda e crônica. Indivíduos na fase aguda, apresentam principalmente poliartralgia/poliartrite, mialgia intensa e febre alta. Estes sintomas são comumente acompanhados por dores, fotofobia e erupções cutâneas (CUNHA, 2017).

Geralmente, indivíduos que apresentam sintomas no período pós-crônico (após o vigésimo primeiro dia da manifestação clínica durando até 3 meses) (SIMON, 2015) demonstram melhoras transitórias, com recaídas após um breve período de melhoras. Nestes pacientes persistem os sintomas clínicos anteriores com aumento de intensidade, principalmente a poliartralgia ou poliartrite (SAM, 2015).

A doença associada a este vírus pode se desdobrar em acometimentos considerados crônicos quando a artralgia persiste por um período maior que três meses, atingindo cerca de 80% dos pacientes (KUNO, 2007). Podem ainda haver diferenças significativas de espectro e frequência dos sintomas clínicos, a artralgia e a artrite, por exemplo, tendem a ser bilaterais, simétricas e migratórias.

O impacto da febre chikungunya é extensivo, principalmente quando considerada a qualidade de vida do paciente, uma vez que é uma patologia severamente debilitante e potencialmente crônica (GATHERER, 2016). O que impacta ainda economicamente por retirar o paciente do ambiente de trabalho e requerer auxílio especializado.

O *Chikungunya virus* tem causado diversas epidemias localizadas nos últimos anos. Durante a epidemia ocorrida no continente africano em 2004, diversos casos foram reportados em regiões tropicais e subtropicais do planeta (PETERSEN, 2016). De 2005 a 2007, uma epidemia localizada na região do Oceano Índico chegou a infectar até 34% da população da ilha de *La Reunion* (LO PRESTI, 2014). Ainda em 2006, diversos casos foram reportados no continente europeu, incluindo Itália, França Suíça, Alemanha, Bélgica e Inglaterra (POWERS, 2007).

Nas américas, casos foram reportados a partir de 2013 (OPS/OMS, 2013) e em 2014 foi observado o primeiro caso de transmissão de *CKV* no Brasil (CARDOSO, 2017). O sequenciamento do genoma viral mostrou que a epidemia foi causada pelo genótipo asiático (CARDOSO, 2017). No mesmo período, foram reportados em outro estado da federação, novos casos de infecção por este organismo associados com o *ECSA* (NAVECA, 2019). No final do período de levantamento de 2014 o Ministério da Saúde já reportava quase 4 mil casos da doença (BRASIL. MINISTÉRIO DA SAÚDE, 2015). Com a presença confirmada de mais de um genótipo circulando no país, estima-se que 94% da população encontra-se em risco (BRASIL. MINISTÉRIO DA SAÚDE, 2015).

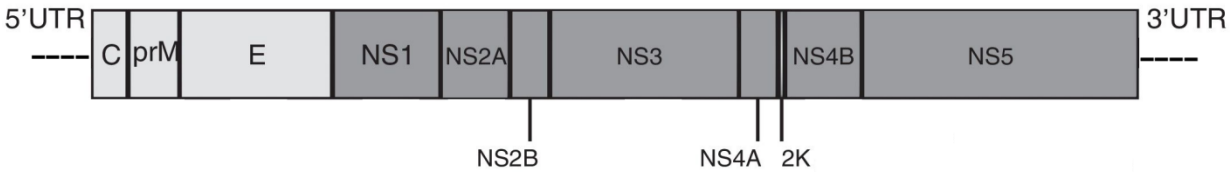
Apesar do elevado risco de infecção e do consequente prejuízo tanto financeiro quanto de qualidade de vida, não existem tratamentos específicos e eficazes para infecção aguda e crônica de *CKV* (GOUPIL, 2016). A maioria dos cursos de tratamentos aplicados no momento consistem em cuidados sintomáticos para reduzir o impacto gerado pelos sintomas na qualidade de vida do indivíduo (ALI OU ALLA, 2011).

Além da ausência de tratamentos eficazes para os sintomas gerados pela doença, destaca-se a ausência de vacinas comercialmente disponíveis (GOUPIL, 2016). Nos últimos anos, diversas tentativas de desenvolvimento de vacinas vêm sendo executadas (THIBERVILLE, 2013, AHOLA, 2015; MORRISON, 2016). Entretanto, nenhuma dessas tentativas obteve sucesso confirmado em modelos humanos. A ausência de tratamentos e vacinas eficazes, demonstra a necessidade de mais estudos sobre o genoma viral. Informações sobre resposta imune e interação entre o agente etiológico em questão e o hospedeiro são essenciais para o processo de desenvolvimento de vacinas e tratamentos mais eficazes.

3.4 *ZIKA VIRUS*

O *Zika virus (ZKV)* foi identificado inicialmente em 1947, em macacos na Uganda e, em 1952, foram reportados casos de infecção em humanos na Uganda, Tanzânia e Nigéria. Posteriormente, foram confirmados casos na África, Ásia, Pacífico e Américas (DOLMA, 2016). Em 2015 o reaparecimento do agente etiológico no Brasil levou à rápida dispersão pelas Américas (PLOURDE, 2016).

Este organismo é um Flavivírus (da família *Flaviviridae*) (LANCIOTTI, 2016). Seu genoma apresenta cerca de 10.700 pares de base, codificando mais de 3.400 aminoácidos (KUNO, 2007). Assim como outros Flavivírus, este agente etiológico é composto por duas regiões não codificantes que flanqueiam um quadro aberto de leitura (*ORF*) que, por sua vez, codifica uma poliproteína com clivagens de função estrutural (Capsídeo (C), Membrana/Pré Membrana (prM/M) e Envelope (E)) e 8 proteínas não estruturais (NS1, NS2A, NS2B, NS3, NS4A, 2K, NS4B, NS5) (KUNO, 2007) como pode ser visto na Figura 3.



Fonte: (GALÁN-HUERTA, 2016)

Figura 3 - Organização das estruturas gênicas no genoma do *ZKV*

Análises filogenéticas sugerem que este organismo pode ser classificado em duas linhagens principais: Africana e Asiática, ambas com origem no leste africano entre 1800 e 1900 (GATHERER, 2016). Análise de cepas coletadas de 4 países no leste africano de 1947 a 2007 revelaram uma forte pressão seletiva no genoma viral (FAYE, 2014), o que sugere a possibilidade de recombinação, rara em flavivírus (FAYE, 2014). Essas observações ressaltam a necessidade de uma avaliação mais complexa do perfil do genoma viral.

A patogênese da infecção por ZKV inicia-se pela entrada do *virion* através do tecido epitelial, por meio de receptores celulares. Diversos fatores facilitam a infecção e autofagia celular, necessária para a replicação de flavivírus. A partir da replicação viral, ocorre a migração para os linfonodos e consequentemente para a corrente sanguínea (HAMEL, 2015).

Assim como outros arbovírus, o *Zika virus* é transmitido principalmente por artrópodes como o *Aedes Aegypti.* Entretanto, foram observadas outras formas de contágio, como: congênita (OLIVEIRA, 2016), perinatal (BESNARD, 2014) e sexual (FOY, 2011; MUSSO, 2015) além de suspeitas de contágio por mordidas de animais (LEUNG, 2015) e transfusão sanguínea (MUSSO, 2014; FDA, 2016).

Clinicamente, a infecção por este organismo passa inicialmente por um período de 3 a 12 dias, correspondente ao período de incubação viral. Após esse período, aproximadamente 80% dos casos é assintomático e auto-limitante (IOOS, 2014). Quando a infecção é sintomática, são observados sintomas similares aos apresentados por outras arboviroses, como dengue e febre chikungunya o que pode confundir e dificultar o diagnóstico (PLOURDE, 2016). Os sintomas mais comuns incluem febre, artralgia, mialgia, fadiga, dor de cabeça, conjuntivite e erupção cutânea (PLOURDE, 2016).

A despeito da natureza auto-limitante dos sintomas, a infecção por este agente etiológico ainda está associada à severas sequelas clínicas. A etiologia foi associada, por exemplo, com o nascimento de bebês com microcefalia, posteriormente chamada de síndrome congênita do vírus da zika (EICKMANN, 2016). Recentemente, esta etiologia foi associada também à síndrome degenerativa de Guillain-Barré, cuja evolução pode levar à morte (STATES, 2016).

Apesar do risco à saúde pública, a facilidade de transmissão e o risco de sequelas crônicas, não existe até o momento uma vacina eficaz para o ZKV (DOLMA, 2016). Além disso, a cocirculação e coinfecção com outros arbovírus, juntamente com a similaridade de sintomas entre as infeções dificulta significativamente o processo de diagnóstico e posterior

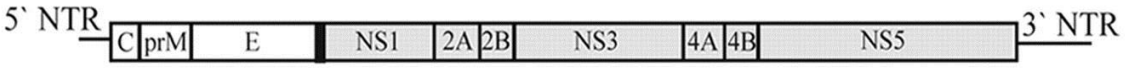
tratamento (DOLMA, 2016). Dado o recente aumento na relevância médica deste agente etiológico e à patologia associada, os esforços para o desenvolvimento de vacinas e e métodos diagnósticos tem aumentado.

A ausência de vacinas e tratamento eficazes reforça a necessidade do estudo aprofundado sobre o genoma viral, interação do agente etiológico com o hospedeiro, resposta imune. Além disso, a dificuldade diagnóstica enaltece o propósito de estudos sobre tipos recombinantes, cocirculação e coinfecção por outros arbovírus.

3.5 *YELLOW FEVER VIRUS*

O agente etiológico apontado como causador da febre amarela, o *Yellow Fever Virus* (*YFV*) foi isolado inicialmente na África em 1927. Entretanto, a primeira epidemia associada ao *YFV*, foi documentada em 1648 no México (CARTER, 1931). Estudos contemporâneos mostram que este agente etiológico ainda causa epidemias localizadas e periódicas nos continentes Africano e Americano (SAAVEDRA, 2017; MONATH, 2015).

Este arbovírus, pertence a família dos Flavivírus cujo nome da família: "*Flaviviridae"* foi dado em sua homenagem, uma vez que *flavus* significa amarelo em latim (BARRETT, 2007). Seu genoma de fita simples possui de 10.500 a 11.000 nucleotídeos. Este genoma possui 3 proteínas estruturais que codificam o Capsídeo (C), Membrana/Pré Membrana (M) e Envelope (E). O genoma codifica ainda 7 proteínas não estruturais (NS1, NS2a, NS2b, NS3, NS4a, NS4b e NS5) (LINDENBACH, 2013; KAUFMANN, 2011; RICE, 1985). A montagem inicial deste genoma pode ser vista na Figura 4.



Fonte: (PATKAR, 2009)

Figura 4 - organização das estruturas gênicas no genoma do *ZKV*

Os *taxons* deste organismo podem ser classificados em três linhagens diferentes: africana do centro-leste, africana do Oeste e sul americana (SAMUEL, 2016; MUTEBI, 2001; CHANG, 1995). Entretanto, existe a possibilidade de recombinação em alguns genótipos de linhagens diferentes (CHANG, 1995) tornando possível desta forma, classificar as cepas provenientes destes organismos em três grupos principais: I, IIA e IIB (CHANG, 1995).

A febre amarela é uma doença debilitante que pode levar até a morte (MONATH, 2015). Os sintomas comumente reportados são: náusea, vômito, dores epigástricas, hepatite com icterícia, falência renal, hemorragia e choque. Podendo chegar ate a morte em uma taxa de 20% a 60% dos casos (TUBOI, 2007).

Considerando o tempo em que o agente etiológico vem gerando epidemias pelo globo, diversos esforços para o desenvolvimento de vacinas foram feitos (MONATH, 2013). Não sem razão, existem vacinas contra o vírus. Esta vacina é considerada muito segura, e raramente causa reações adversas (MONATH, 2013). Entre as reações está a síndrome de *Guillain-Barré* (VITAL, 2002), paralisia bulbar, paralisia do sino e inflamação do nervo óptico (VOIGT, 2001).

Desta forma, é reforçada a necessidade de um estudo mais profundo do genoma viral, e principalmente da interação entre o agente etiológico e o hospedeiro. Desta forma será possível compreender melhor a patogênese, o desenvolvimento e recombinação viral. Além de tornar possível o desenvolvimento de tratamentos e vacinas mais eficazes.

3.6 BANCOS DE DADOS BIOLÓGICOS

Considerando a crescente redução do custo do sequenciamento de alta demanda, os conjuntos de dados biológicos crescem exponencialmente em tamanho e complexidade. O desenvolvimento de plataformas para análise e gerenciamento de grandes quantidades de dados biológicos é considerado fundamental em bioinformática (ZOU, 2015). Apenas em 2014, foi reportada a existência de 1552 bases de dados biológicas de acesso público (FERNÁNDEZ-SUÁREZ, 2014). Esses conjuntos de dados biológicos podem ser classificados de três formas principais: Escopo da cobertura dos dados, tipo de dados gerenciados e método de curadoria. A respeito deste processo, os bancos de dados biológicos podem ser classificados como primários, secundários e especializados.

Os bancos de dados considerados primários contêm dados brutos, geralmente não curados, como o *GenBank*. Já os bancos de dados secundários possuem um certo nível de curadoria. Os bancos de dados especializados, possuem por sua vez dados de um determinado organismo, com um certo nível de curadoria (ZOU, 2015).

Existem alguns conjuntos de dados atualmente, cuja proposta é organizar e analisar dados de agentes etiológicos. Por exemplo, banco de dados específico para o *Ebola*, *HFV* (KUIKEN, 2012), *HCV* (KUIKEN, 2005), *HIV* (LEITNER, 2017), *Influenza virus* (BAO, 2008) e outros vírus.

Esta abordagem é amplamente utilizada, a maioria dos bancos analisaram somente um pequeno conjunto de dados dos bancos de dados primários (LEITNER, 2017). Outrossim, com o crescimento acelerado dos conjuntos de dados primários, os bancos de dados especializados não conseguem acompanhar a escala de informações.

Além do crescimento das bases primárias, existe o desafio da manutenção da curadoria manual dos dados. Uma vez que este é um processo demorado, demanda equipe qualificada e de proporções equivalente ao tamanho e complexidade do conjunto de dados. Este tipo de processo eventualmente se torna um impedimento para o crescimento do conjunto de dados. Por ser um processo custoso, algumas organizações mantenedoras de conjuntos de dados, cancelaram a manutenção e eventualmente foram descontinuados como, por exemplo, o *HCV Databases* (KUIKEN, 2005).

Desta forma, é enaltecida a necessidade de um sistema automatizado de curadoria. Este sistema deve ainda, de maneira recursiva, alimentar o conjunto de dados com os novos dados disponibilizados nos bancos primários, garantindo a manutenção do crescimento e qualidade dos dados. Entretanto, um conjunto de dados desta natureza ainda não está estabelecido na literatura.

3.7 ALINHAMENTO DE SEQUÊNCIAS

Uma das formas de analisar os dados de sequências nucleotídicas é o processo de alinhamento de sequências. Este processo trata-se da comparação de duas ou mais sequências nucleotídicas levando em consideração a similaridade entre regiões das sequências, assim como *INDELs*, que são inserções ou deleções de nucleotídeos nas mesmas. O termo é originado, pois não é possível determinar, sem análises filogenéticas, se houve inserção ou deleção no sítio em questão.

Assumindo a comparação entre duas sequências, é possível representá-las em duas dimensões, para, desta forma identificar os possíveis pareamentos entre as sequências em questão. O resultado do alinhamento trata-se de um coeficiente de similaridade entre as duas sequências utilizado, além de uma sequência que representa o pareamento entre as regiões que é utilizada principalmente para o mapeamento de regiões genômicas (DESHMUKH, 2015).

O processo de alinhamento pode ser classificado quanto ao modelo funcional em duas categorias principais: local e global. O alinhamento global, pressupõe que ambas as sequencias possuem ancestrais comuns, ou seja, são homólogas. Assim, um alinhamento parte da suposição de que as sequências deveriam ter o mesmo tamanho. Enquanto o

alinhamento local, busca a região de maior nível de similaridade (DESHMUKH, 2015). O algoritmo global de alinhamento baseia-se nas dimensões das sequências, considerando que o sítio 1 da sequência A, trata-se do sítio 1 da sequência B, enquanto o sítio n da sequência A refere-se aos sítios n da sequência B.

É possível obter um alinhamento, seja ele global ou local, utilizando duas metodologias principais. A metodologia de alinhamento ótimo, que tem como resultado encontrar o melhor alinhamento possível entre as sequências, utilizando algoritmos exaustivos (ou exatos) de alinhamento. Esta metodologia é bastante utilizada no processo de obtenção de informações de sequências nucleotídicas, pois se trata do melhor alinhamento possível, levando em consideração a ponderação das deleções e inserções de nucleotídeos no genoma. Além da progressão de tais inserções ou deleções (CHAKRABORTY, 2013).

Uma outra maneira é a utilização de abordagens heurísticas no processo de alinhamento. Este processo surgiu pela necessidade de tratamento de extensas bases de dados, uma vez que a estratégia utilizada no encadeamento de algoritmos exatos não satisfazia as necessidades de otimização de recursos. Entretanto, ao utilizar técnicas heurísticas, parte da acurácia do alinhamento é perdido, tendo este como resultado apenas um dos melhores alinhamentos (CHAKRABORTY, 2013).

Os alinhamentos de sequências considerados ótimos, como os propostos por Needleman e Wunsch, e por Smith e Waterman, (NEEDLEMAN, 139; SMITH, 1981) possuem uma ordem de complexidade computacional proporcional ao tamanho das sequências trabalhadas. Desta forma a ordem de complexidade se torna quadrática, em relação ao tempo e recurso de máquina utilizados. Isso se dá porque ambos utilizam matrizes de computação dinâmica, onde cada possibilidade de alinhamento é calculada. Posteriormente, cada possibilidade anotada é considerada, para que este tipo de algoritmo possa garantir sempre o melhor alinhamento possível entre estas duas sequências.

O algoritmo proposto por Needleman e Wunsch (NEEDLEMAN, 1970) é um bom exemplo de alinhamento global exaustivo. Este algoritmo utiliza técnicas de programação dinâmica para alocar, construir, preencher e percorrer uma matriz, que se refere à relação entre os nucleotídeos das sequências em questão. Durante o processo de construção da matriz (Figura 5 a), o algoritmo toma uma série de decisões, que podem ser uma das representadas na Figura 5 b

Cada decisão tomada em uma posição da matriz gera um determinado estado, construindo assim uma matriz de estados, onde o próximo estado é definido pelo estado atual, dadas as posições nas sequências e a decisão atual. Esta matriz de estados é construída como

definido na Figura 5 c, e posteriormente percorrida construindo a sequência final de alinhamento.

O coeficiente de similaridade entre as sequências, resultado final do alinhamento, é calculado de forma cumulativa à matriz de estados atuais. Para tal, é levado em consideração o valor resultante de cada pareamento realizado, e em seu somatório é obtido o escore final de similaridade entre as sequências como sintetizado na seguinte fórmula:

$$S = \max_{u1,u2,\ldots,ul} \sum_{k=1}^{L} f(xk+1, uk)$$

Onde *S* representa o coeficiente de similaridade e é dados pelo somatório das maximizações locais de cada pareamento *u* gerado(*L*) pela função de maximização de *scores*.

a.

	S_2	a	c	g	t	g	a	g	a	g	t
S_1											
t		-1	-1	-1	3	-1	-1	-1	-1	-1	3
t		-1	-1	-1	3	-1	-1	-1	-1	-1	3
c		-1	3	-1	-1	-1	-1	-1	-1	-1	-1
g		-1	-1	3	-1	3	-1	3	-1	3	-1
g		-1	-1	3	-1	3	-1	3	-1	3	-1
a		3	-1	-1	-1	-1	3	-1	3	-1	-1

b.

$$u_k = \begin{cases} \searrow & \text{diagonally right and down,} \\ \rightarrow & \text{right,} \\ \downarrow & \text{down.} \end{cases}$$

c.

	S_2	a	c	g	t	g	a	g	a	g	t
S_1											
t											
t											
c											
g											
g											
a											STOP

Fonte: (POLANSKI, 2007)

Figura 5 - a) Matriz de scores gerados a partir da maximização das possibilidades de pareamento entre as sequências em questão. b) Sumarização de decisões do algoritmo de maximização onde *uk* representa o universo de decisões possíveis para o algoritmo. E cada seta representa uma decisão e por tanto uma movimentação na matriz de decisões. c) Matriz de decisões geradas a partir dos scores utilizando a notação vista em *c*

O algoritmo proposto por Smith e Waterman (SMITH, 1981) é outro exemplo de algoritmo de alinhamento exato, utilizado para análise de sequências nucleotídicas e proteicas. Este algoritmo não pressupõe homologia entre as sequências, buscando apenas a região de melhor pareamento entre as mesmas. Desta forma, para realizar o alinhamento local entre as sequências, uma matriz é criada, posteriormente preenchida, e finalmente percorrida. Isto gera assim, o alinhamento entre as regiões com maior similaridade (DESHMUKH, 2015).

O algoritmo realiza a maximização destas similaridades para identificar regiões de maior similaridade entre as sequências (DESHMUKH, 2015). Para tal, são atribuídas pontuações para cada pareamento de nucleotídeos realizado durante o alinhamento. Estas pontuações variam do tipo de par criado. Caso seja um correspondência, ou não correspondência, o algoritmo credita ou debita uma pontuação fixada. Outrossim, o algoritmo atribui uma penalidade às áreas não pareadas, às áreas com *gaps* (*indels*) e para as extensões dos *gaps*. Estas penalidades permitem que o alinhamento final não apresente necessariamente o mesmo tamanho das sequências alinhadas (DESHMUKH, 2015).

Outras técnicas de alinhamento de sequências largamente utilizadas são as consideradas heurísticas (CHAKRABORTY, 2013; PRUESSE, 2012). Estas técnicas não trabalham com todas as possibilidades de alinhamento do par de sequências para inferir o melhor alinhamento. Este tipo de algoritmo determina dinamicamente os alinhamentos menos prováveis e os elimina do processo, considerando apenas aquelas com maior probabilidade de acerto.

Apesar de eficazes em relação ao tempo e ao custo computacional, a utilização de técnicas heurísticas como Modelos Ocultos de Markov do inglês *Hidden Markov Models* (*HMM*), alinhamento progressivo, técnicas iterativas determinísticas e computação evolutiva (Algoritmos Genéticos) podem reduzir a acurácia do alinhamento final. Além de depender diretamente da forma como os dados estão dispostos inicialmente (CHAKRABORTY, 2013).

Assim, o tratamento heurístico em bancos de dados biológicos dinâmicos não seria recomendável, uma vez que dados são inseridos constantemente e modificam a estrutura do conjunto de dados. Considerando que as análise heurísticas são extremamente parametrizáveis em relação ao conjunto de dados, a dinâmica das análises modifica constantemente, o que impactará no desempenho e resultado das análises.

3.8 MAPEAMENTO DE SEQUÊNCIAS NO GENOMA COMPLETO

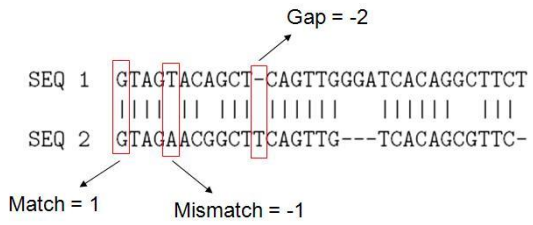
O procedimento de mapeamento de sequências nucleotídicas consiste na identificação dos limites (início e fim) do pareamento entre a sequência em questão, e um mapa de características anotadas no genoma completo de um determinado organismo. Este procedimento é considerado essencial para a análise de dados provenientes de sequenciamento de alto rendimento. Com isso, é possível determinar quais genes uma determinada sequência compreende. Outrossim, no caso da ocorrência de mutações, é possível identificar o gene em que a mesma ocorreu e assim fornecer suporte a estudos de técnicas diagnósticas e vacinas (COMBE, 2014).

Para realizar o processo de mapeamento, é necessário partir da montagem de um mapa de características disponíveis no genoma completo do organismo. Este mapa é criado a partir da sequência de referência que representa o genoma completo, onde as anotações de informações funcionais e posicionais são levadas em consideração e estão disponíveis para consulta. Desta forma, a partir do alinhamento entre as sequências (em questão e referência) é possível inferir o posicionamento de um determinado fragmento em relação ao genoma completo e assim, é possível identificar quais características são comuns. Este processo possui uma ordem de complexidade unitária proporcional ao tamanho das sequências alinhadas e complexidade agrupada da ordem do tamanho do conjunto de dado, crescendo linearmente. Entretanto é considerado um crescimento de complexidade unidirecional, uma vez que o tamanho da sequência de referência é constante (KASPRZYKOWSKI, 2016).

Este processo de alinhamento indicará quais regiões do genoma completo são “cobertas” pela sequência em questão e a partir desta “cobertura” é possível identificar quais características presentes no genoma completo são completamente cobertas na sequência em questão, e quais são apenas parcialmente cobertas. Além disso, com as informações de cobertura das sequências presentes em um conjunto de dados, é possível classificar o conjunto quanto à cobertura. Isso possibilita a geração de sub conjuntos de dados separados em uma determinada região, ou que codifiquem determinado gene. Desta forma aumentando significativamente a qualidade dos dados disponíveis nos conjuntos de dados (KASPRZYKOWSKI, 2016).

A “cobertura” de uma determinada sequência, é determinada pelos limites do alinhamento. Identifica-se o início de um alinhamento observando a primeira posição pareada em relação ao genoma completo. Enquanto a última posição pareada do alinhamento corresponde ao final da cobertura como visto no exemplo da Figura 6. Além dos limites, é preciso identificar se a cobertura contém regiões onde o alinhamento identificou inserções ou deleções. Com essa informação, é possível classificar a cobertura em total ou parcial. Assim é

possível identificar se a sequência em questão codificar totalmente ou parcialmente a característica anotada (KASPRZYKOWSKI, 2016).



Fonte: (POLANSKI, 2007)

Figura 6 - Exemplo de alinhamento global de duas sequências onde há pareamento parcial entre as regiões

Com as informações sobre o mapa de características de cada sequência no conjunto de dados, é possível identificar subconjuntos de dados alvo para determinados estudos. É possível assim selecionar sequências que codificam regiões específicas, como por exemplo, regiões que codificam epítopos (estruturas reconhecidas pelo sistema imune). Desta forma, o processo de mapeamento contribui diretamente para o desenvolvimento de novas vacinas, tratamentos, e qualificação dos existentes.

3.9 CLASSIFICAÇÃO GENOTÍPICA DAS SEQUÊNCIAS

Para melhor compreender a dinâmica viral, seu comportamento mutacional e a interação dos mesmos com os hospedeiros é essencial compreender a classificação entre grupos genotípicos presentes no organismo. É possível identificar tal classificação a partir de um perfil já descrito de seus grupos, do alinhamento entre as sequências do conjunto de dados e as referências destes grupos (CHAN, 2014).

Em resumo, o processo de classificação de uma determinada sequência é o ato de selecionar entre os subtipos já identificados do organismo, aquele com o qual a sequência apresenta maior semelhança. É necessário então, determinar todas as sequências de referência (que representam suas características) do subtipo.

Para realizar a classificação, pode ser utilizada a metodologia de alinhamento local para alinhar cada sequência de referência de cada grupo classificatório, com a sequência que se quer classificar. Este processo de alinhamento gera um escore (coeficiente de similaridade) para cada alinhamento. Considerando este escore como a distância entre as características das sequências, é possível identificar o nível de similaridade da sequência *query* com os subtipos.

Desta forma, é possível identificar o subtipo com o qual a sequência em questão apresenta maior similaridade.

Considerando que, novos subtipos e recombinantes podem ser identificados no futuro, a quantidade de subtipos disponíveis por organismo é variável e dinâmica. Observa-se ainda que a complexidade desta técnica é diretamente proporcional ao número de subtipos e em relação aos tamanhos das sequências de referência dos subtipos já identificados. Ademais, determinados subtipos podem possuir mais de uma sequência de referência identificada, demandando assim que a subtipagem ocorra em ambas as sequências, permanecendo apenas o maior escore. Desta forma aumentando exponencialmente a complexidade do processo.

A utilização de metodologias exaustivas ou ótimas é de ímpar importância para a investigação dos subtipos (PINEDA-PEÑA, 2013). Desta forma, faz-se necessário a avaliação e cálculo de todos os alinhamentos possíveis entre as sequências. Assim, é possível definir matematicamente o melhor alinhamento possível entre as sequências. Este processo é computacionalmente custoso. Entretanto, com os seus resultados, será possível definir as frequências em que determinados subtipos estão presentes no conjunto de dados (CHAN, 2014).

3.10 MAPEAMENTO DE EPÍTOPOS

O processo de identificação e mapeamento de epítopos imunogênicos é considerado crucial para o desenvolvimento de vacinas (LI PIRA, 2010; KAST, 2010). Este processo consiste em identificar a presença de sequências de aminoácidos, capazes de compor peptídeos que por sua vez podem gerar resposta imune, ou seja, imunodominantes (epítopos) (LI PIRA, 2010). Este processo de busca e mapeamento de epítopos contribui significativamente também para a compreensão da patogênese, e para o desenvolvimento de novos tratamentos e vacinas (GIMENEZ, 2016).

Apesar de importante para o processo de desenvolvimento de vacinas, a informação sobre frequência de epítopos não estão disponíveis nos bancos de dados primários. Isto se dá principalmente porque o processo de geração destas informações é complexo, uma vez que os conjuntos de dados são grandes, assim como as bibliotecas de epítopos imunogênicos (GERSHONI, 1981).

Uma das bibliotecas de epítopos mais utilizadas é a *Immune Epitope Database* (IEDB) (VITA, 2015). Este conjunto de dados conta com mais de um milhão e 200 mil epítopos, destes 450 mil apenas em 2018 (VITA, 2015). Além dos dados dos epítopos a base também tem dados sobre os organismos, alelo, agente etiológico entre outras informações.

Para que ocorra o processo de mapeamento de epítopos, é preciso realizar a importação desta biblioteca. Com estas informações coletadas, é preciso identificar as áreas de codificação das sequências disponíveis no conjunto de dados. Desta forma, é possível localizar e referenciar cada ocorrência de um determinado epítopo no conjunto de dados.

Com as informações sobre frequência dos epítopos no conjunto de dados, é possível inferir novos candidatos para estudos relacionados à resposta imune nos agentes etiológicos. É possível ainda identificar regiões por subtipos que tem maior chance de servir como regiões alvo para fármacos, facilitando o processo de desenvolvimento de vacinas e tratamentos eficazes (GERSHONI, 2007).

3.11 INTERFACE DE SOFTWARE

Uma API (*Application Programming Interface*) é um conjunto de funções padronizadas por uma entidade de código que garante a compatibilidade entre aplicações. Este conjunto de funções permite que aplicações desconectadas possam trocar informações, objetos e estados num único padrão. Ao padronizar a interface de uma determinada aplicação, é possível reduzir o tempo e complexidade de desenvolvimento de aplicações com funcionalidades similares (KAMARULZAMAN, 2019).

Assim, uma API é um conjunto de funcionalidades implementada e disponibilizada por uma determinada aplicação à outra aplicação, evitando o retrabalho de desenvolvimento e distribuindo corretamente os recursos computacionais disponíveis. Desta forma, é observado que há a necessidade de padronização do processo de integração do software com as plataformas existentes. Assim, possibilitando o acesso ao conjunto de dados em profundidade.

Uma interface de *software* é um contrato de utilização entre dois sistemas onde é definido um determinado padrão para acesso a uma funcionalidade. Desta forma, é possível realizar uma tarefa genérica mesmo fora do escopo de uma determinada aplicação. Na prática a interface serve como modelo de transferência de estados de objetos entre duas aplicações, organizando a comunicação entre as mesmas.

3.12 INTERFACE WEB

No caso da interface *WEB*, é uma interface de acesso entre um *software* e seus usuários. Neste caso não são trocadas informações sobre estados de objetos. São transmitidas informações através de protocolos de comunicação como HTTP (*Hyper Text Transfer Protocol)* ou HTTPS (*Hyper Text Transfer Protocol Secure)* (ARCURI, 2019). Estes

protocolos são responsáveis pela correta comunicação entre o navegador do usuário e o servidor da aplicação existente. O cliente então recebe os dados em formato HTML (*Hypertext Markup Language*) que essêncialmente é uma página de texto marcado. Este arquivo em HTML é interpretado pelo *browser* cliente mais rapidamente que páginas compiláveis. Assim, de maneira rápida o cliente tem acesso à informações disponíveis no servidor.

Entretanto, nem toda a informação de interface de uso está disponível no HTML, uma vez que o usuário pode interagir ainda no escopo de cliente (navegador) com a página. Assim, para garantir que esta interação gere resultado, são implementadas funções em javascript que são executadas pelo próprio navegador e permitem manipulação de dados e de sua exibição, dinamizando assim a interface.

3.13 REST API

Representational State Transfer (REST) é um padrão de abstração para o desenvolvimento de arquiteturas de software distribuídos, onde é gerada uma arquitetura de microsserviços que trafegam dados, funções e estados de objetos usando protocolo HTTP e HTTPS (ARCURI, 2019).

Este tipo de abstração de API é responsável por criar um alfabeto de transações de dados que permitem que o usuário ou outra aplicação acessem os dados disponíveis em uma API de maneira direta e organizada.

Para tal, são utilizados os verbetes GET, POST, PUT, DEL para respectivamente coletar dados, salvar novos dados, atualizar dados e apagar dados. Para todos estes métodos, são transmitidos dados tanto no próprio link quanto no corpo da requisição em formato JSON (*Javascript Object Notation*).

A utilização de uma API RESTfull permite o desenvolvimento de uma interface WEB desconectada do serviço de aplicação e independente do seu crescimento, uma vez que a maioria do processamento ocorre apenas na API e o *frontend* tem apenas que, utilizando os verbetes corretamente, realizar requisições ao *backend* para fornecer acesso aos dados.

3.14 MICROSSERVIÇOS

Uma arquitetura de microsserviços, é a divisão das funcionalidades de uma determinada aplicação em serviços independentes apesar de interconectados. Cada serviço possui uma escalabilidade programada e é ou não sincronizado com serviços pares.

Esta arquitetura fornece à aplicação flexibilidade suficiente para servir como base funcional para outros projetos, facilitando e acelerando seu desenvolvimento. Uma arquitetura desta natureza permite ainda a escalabilidade vertical (de *hardware*) e horizontal (entre instâncias) para a aplicação, preparando seus serviços para computação em nuvem e em *grid*.

Além do aumento da manutenibilidade da aplicação, a facilidade de adição de novos serviços sob demanda é uma característica primária da arquitetura de microsserviços. Desta forma é possível atender às mais diversas demandas sem necessitar reconstruir funcionalidades.

4 MATERIAIS E MÉTODOS

4.1 COLETA DE SEQUÊNCIAS

O processo de coleta de sequências nucleotídicas foi realizado através da ferramenta *zeseeker*, disponível no *workbench VSDBM* (KASPRZYKOWSKI, 2019). Previamente desenvolvido pelo autor, esta ferramenta possui conexão direta com os serviços do NCBI. Fazendo uso destes serviços, é possível realizar buscas e downloads de sequências por demanda, de forma rápida e prática.

Inicialmente, o sistema faz conexão com a API disponibilizada pelo NCBI, verificando a existência e disponibilidade dos serviços de busca e coleta de dados, além de verificar a estabilidade da conexão e a consistência dos dados no banco de dados *nucleotide*. Desta forma é possível garantir a estabilidade da coleta em larga escala, além de permitir o cálculo do progresso de forma mais precisa.

A partir da comunicação inicial, é passado para o sistema de busca uma *query* contendo o nome do organismo e identificador do organismo como forma de parâmetro. O sistema então retorna uma lista de chaves primárias ou identificadores globais (*GI's e Acessions*) referentes às sequências disponíveis neste banco de dados que estejam relacionadas com o parâmetro supracitado. Neste ponto, o módulo de coleta realiza um escalonamento de tarefas baseado na largura de banda anteriormente calculada, além da disponibilidade do banco, e quantidade de sequências (chaves) retornadas. Desta forma é possível iniciar o processo de download das sequências.

Entretanto, nem todas as sequências representadas pelas chaves são bons candidatos para download, uma vez que estas podem já estar disponíveis no conjunto de dados local. Desta forma, antes de proceder o processo de download, é realizada uma filtragem no conjunto de chaves pré-existentes, efetivamente eliminando as sequências que já estejam disponíveis localmente.

O procedimento de download das sequências nucleotídicas é separado em tarefas e cada tarefa é responsável por realizar o download de uma determinada quantidade de sequências. Sendo esta quantidade definida a partir da largura da banda disponível e da quantidade de chaves já filtradas e candidatadas para download. Como estas tarefas não são inter-dependentes, é possível realizar a paralelização das mesmas. Isso garante que a totalidade da largura de banda seja utilizada, assim como a exploração máxima dos recursos computacionais locais e serviços oferecidos externamente.

Entretanto, a depender do tamanho do conjunto de dados, é preciso realizar o controle da quantidade de tarefas que realizam download simultaneamente. Para tal, foi implementado um gerenciador de tarefas de download, que calcula a quantidade de sequências sendo baixadas em cada tarefa, assim como a banda utilizada. Desta forma escalonando de forma lógica as tarefas que ainda não foram realizadas, otimizando os recursos disponíveis na máquina servidora, otimizando e acelerando o processo de coleta. Este processo de coleta pode ser visto esquematicamente no diagrama de atividades e estados exibido na Figura 7.

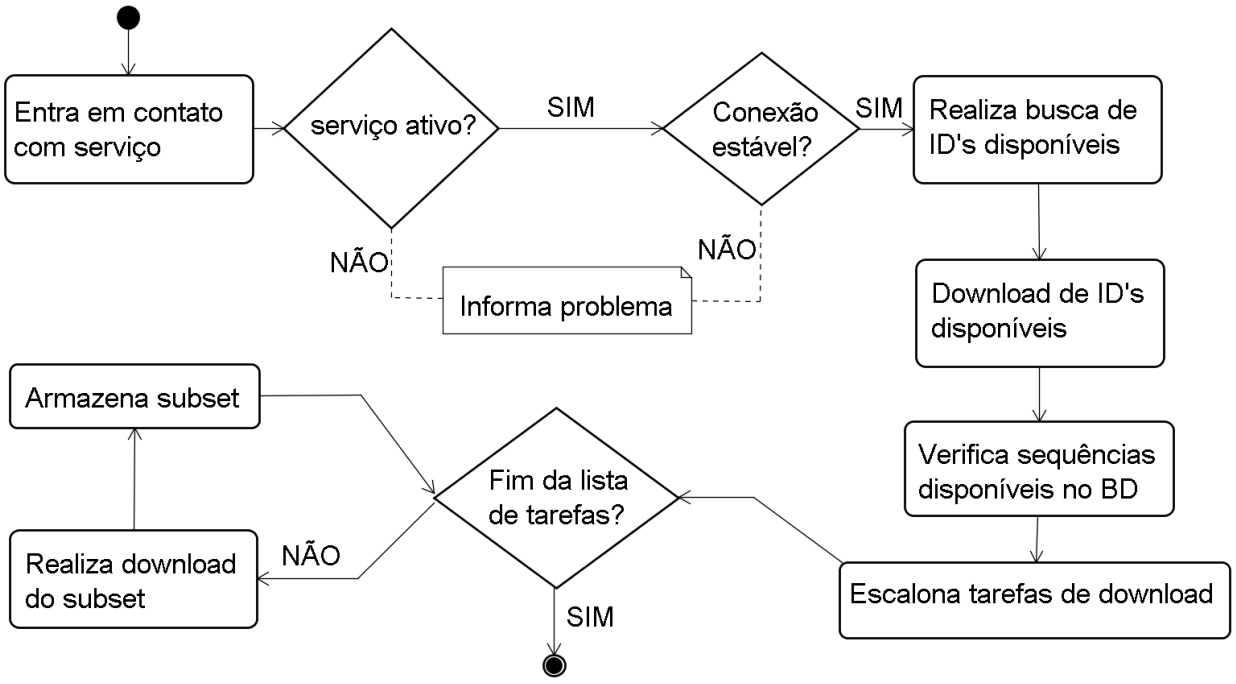


Figura 7 - Diagrama de atividades do processo de coleta de sequências do *GenBank*

Foram coletadas então sequências dos arbovírus ZKV, DENV, CKV e YFV disponíveis no Genbank através de uma junção de buscas utilizando como parâmetros os campos *organism_id* e *organism_name.* Tanto as sequências como os dados de suporte para as mesmas coletadas através de referências cruzadas no próprio conjunto de bancos de dados do Genbank, foram armazenadas no modelo relacional do SAGA Figura 8.

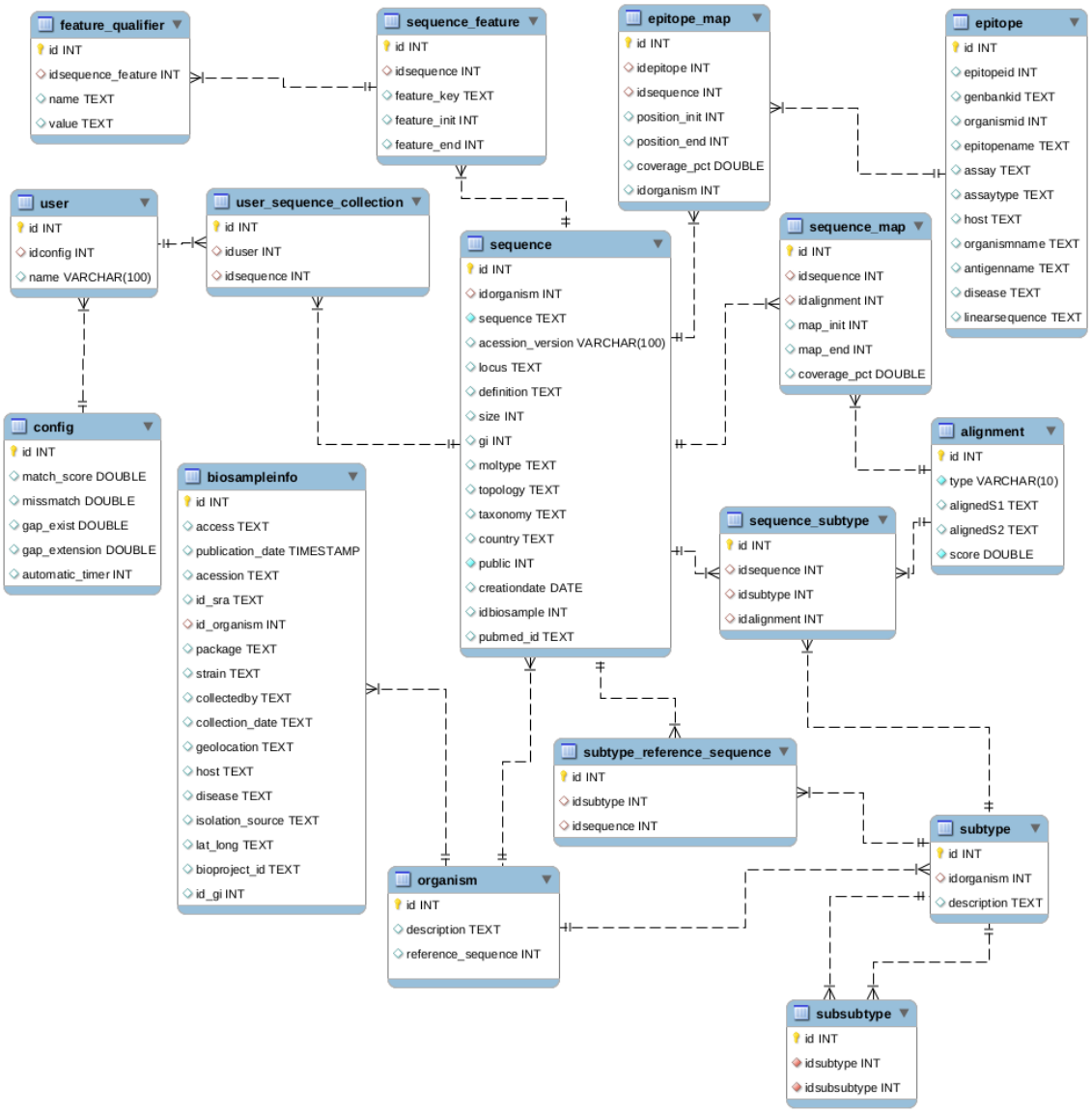


Figura 8 - DER (Diagrama de Entidades e Relacionamentos) do banco de dados compreendendo dados basais coletados de bancos de dados primários e de epítopos bem como dados de suporte à análise e secção de resultados.

Este modelo contempla o armazenamento organizado dos dados desde basais como dados se sequência nucleotídicas, regiões codificantes, referências e organismos, a dados de suporte à análise como dados de subtipagem, mapeamento, alinhamento e dados de mapeamento de epítopos.

Na tabela “sequence” e nas tabelas diretamente associadas temos dados basais das sequências. Na própria tabela apenas dados contidos no arquivo Genbank clássico além de chaves estrangeiras para tabelas suporte. Na linha de tabelas de “feature”, compreendendo “sequence_feature” e “feature_qualifier” temos as facetas representadas pela sequência no XML principal da sequência, tais como regiões codificantes, seus limites e resultados. Na

tabela “biosample” existem dados extras para as sequências baseadas num segundo banco de dados disponibilizado no NCBI. Apesar de estes dados serem mais completos, contendo georreferenciamento, dados de coleta e amostragem, não estão disponíveis para todas as sequências. Existe ainda a tabela “epitope” que armazena as principais informações do XML base de epítopos vinda do IEDB.

Além das tabelas basais, temos as tabelas de configuração como “config”, “organism”, “subtype” que são responsáveis por armazenas referências e configurações a serem utilizadas durante as análises nos processos de alinhamento, classificação e mapeamento das sequências. Além de tabelas de suporte e armazenamento de resultados de análise como “subtype_reference_sequence”, “sequence_subtype”, “sequence_map” e “epitope_map”, que sozinha representa a tabela com maior tráfego do conjunto de dados com 92 milhões de registros.

O processo de coleta, por sua natureza recursiva, é repetido a cada 5 dias corridos (configuráveis através do banco de dados). Após este período, é feita uma verificação no conjunto de dados para garantir integridade, é realizada nova busca, considerando ainda a versão dos identificadores globais (*Acession Version*) sobrepondo informações quando atualizáveis e adicionando informações novas ao conjunto de dados, disparando novas chamadas à análise do conjunto de dados inteiro.

4.2 SUBTIPAGEM

Para o processo de subtipagem, foi preciso criar um mapa classificatório, baseado nas variantes existentes e suas respectivas cepas de referência. Esses dados foram então coletados na literatura base dos agentes etiológicos e então identificados. As sequências de referência para cada grupo foram selecionadas baseado na literatura específica de cada agente etiológico. Desta forma foi feita uma revisão bibliográfica sendo sub-divididos as linhagens genômicas dos organismos em grupos regionais específicos. Para alguns grupos regionais, mais de uma referência foi adicionada, elevando assim a acurácia teórica do processo de classificação. Os grupos genotípicos e suas sequências de referência, assim como a sequência de referência do genoma completo com seus identificadores globais podem ser visos nas Tabelas 1, 2, 3 e 4 para ZKV, DENV, CKV e YFV respectivamente.

Tabela 1 - Linhagens e subtipos com seus respectivos genomas e identificadores globais de ZKV.

Referência do Organismo	NC_012532.1

Linhagem	Subtipo	Referência 1	Referência 2	Referência 3	Referência 4
ASIAN	Micronesia	EU545988.1	KU681082.3	LP918026.1	LP837161.1
ASIAN	Cambodia	JN860885.1	MH063265.1	MH882548.1	NC_035889.1
ASIAN	Malaysia	HQ234499.1	KX601167.1	KX694533.2	KX377336.1
AFRICAN	Senegal	HQ234501.1	KF383118	KF383119.1	KF383116.1
AFRICAN	Nigeria	HQ234500.1	KU963574.2	-	-
AFRICAN	Uganda	HQ234498.1	NC_012532.1	LC002520.1	KF383115.1

Tabela 2 - Linhagens e subtipos com seus respectivos genomas e identificadores globais de DENV.

Referência do Organismo	NC_001477.1
Subtipo	**Referência**
1	KM204119.1
2	KM204118.1
3	KU050695.1
4	KR011349.2

Tabela 3 - Linhagens e subtipos com seus respectivos genomas e identificadores globais de CKV.

Referência do Organismo		NC_004162.2			
Linhagem	**Subtipo**	**Referência 1**	**Referência 2**	**Referência 3**	**Referência 4**
ASIAN	Philipines	AF192895.1	-	-	-
ASIAN	Thailand	AF192896.1	AF192900.1	AF192897.1	AF192899.1
ASIAN	Indonesia	AF192894.1	-	-	-
ASIAN	Malaysia	AF394210.1	-	-	-
ASIAN	India	AF192902.1	AF192901.1	-	-
ECSA	Uganda	AF192907.1	-	-	-
ECSA	Tanzania	L37661.3	-	-	-
ECSA	Central Africa	AF192906.1	-	-	-
ECSA	South Africa	AF192903.1	AF192904.1	-	-
West Africa	Senegal	AF192892.1	AF192891.1	-	-
West Africa	Nigeria	AF192893.1	-	-	-

Tabela 4 - Linhagens e subtipos com seus respectivos genomas e identificadores globais de YFV.

Referência do Organismo		NC_002031.1	
Linhagem	**Genótipo**	**Referência 1**	**Referência 2**
I	Sudan	U23577.1	-
I	Ethiopia	U23576.1	-
I	Uganda	U23578.1	-
I	Central Africa	U23571.1	U23573.1

I	Kenya	U23569.1	U23575.1
IIA	Senegal	L02865.1	U23574.1
IIA	Nigeria	U23572.1	U23567.1
IIA	Trinidad	U23568.1	U23579.1
IIB	Colombia	U23580.1	-
IIB	Brazil	U23570.1	-
IIB	Peru	U23565.1	D14458.1
IIB	Ecuador	U23566.1	-

Com isso, indexou-se quatro subtipos com participação relevante na epidemia associada ao vírus da dengue (LINARES, 2013; KURANE, 2007; MUSTAFA, 2015). Quanto ao vírus causador da febre chikungunya, foram identificadas 3 variações diferentes, baseado nas endemias (SAM, 2012; NUNES, 2015; TSETSARKIN, 2007). Para o *Zika Virus* foram identificadas duas linhagens principais baseado em sua origem de isolamento (GATHERER, 2016; FAYE, 2014). Enquanto foram identificadas 3 linhagens diferentes para o vírus associado com a febre amarela (SAMUEL, 2016; MUTEBI, 2001; CHANG, 1995).

Entretanto, as diversas sequências isoladas destes agentes etiológicos não apontam para uma referência única para cada grupo genotípico. Desta forma, foi preciso considerar cada grupo com um perfil variante de referências. Isso significa que para realizar a classificação de uma sequência, se fez necessária a comparação da mesma com uma série de referências isoladas apenas deste grupo, baseando-se na literatura deste conjunto e só então realizar comparações com outros conjuntos de referência.

Feito este levantamento, foi observado que apesar de a quantidade de subgrupos ser relativamente pequena, a quantidade de cepas a serem comparadas permanece relativamente grande, 47 no total para todos os organismos estudados. Isso significa que o processo de subtipagem sofreu as consequências de problemas de classificação muitos para muitos, esse tipo de problema aumenta a complexidade da classificação, tornando-o mais demorado, custoso e, por vezes, infactível.

Para realizar o processo de subtipagem das sequências disponíveis no conjunto de dados, foi então preciso construir *clusters* de cepas para cada grupo tendo com representantes deste grupo a(s) referências associadas ao mesmo na literatura. Uma vez criados, esses *clusters* são utilizados para identificar o melhor pareamento com a sequência que se deseja classificar. Este pareamento é definido a partir do coeficiente de similaridade obtido a partir

de um alinhamento local entre sequências disponíveis no conjunto de dados e os representantes de cada agrupamento.

Isto posto, para identificar a metodologia a ser utilizada no processo de alinhamento, além da quantidade de cepas de referência identificadas, foi preciso observar que os conjuntos de dados a respeito desse tipo de agente etiológico é dinâmica. Ou seja, se expande e se modifica constantemente, tornando o processo de análise heurística defasado. Considerando a dependência funcional apresentada pelas principais técnicas heurísticas em relação à composição e dinâmica do conjunto de dados.

Desta forma, foi preciso aplicar uma metodologia exata no processo de alinhamento. Assim, cada sequência do conjunto de dados associada a um determinado agente etiológico é alinhada utilizando o alinhamento local ótimo com cada referência de cada grupo genotípico pré cadastrado. Desta forma, utilizando técnicas já validadas pelo nosso grupo, foi possível classificar todas as sequências disponíveis no conjunto de dados em menos de 36 horas.

Apesar da acurácia matemática dos métodos exaustivos de classificação, é preciso considerar a conformação dos dados disponíveis em bancos de dados biológicos primários. Uma vez que estes dados, devido à limitações no processo de sequenciamento, podem representar apenas uma fração do genoma. Assim, sequências muito restritas ou com baixa taxa de informações associadas podem representar um falso-positivo no processo de classificação.

Além de realizar a classificação genotípica, foi preciso ainda registrar todos os melhores alinhamentos para cada sequência. Uma vez que novas variantes dos agentes etiológicos em questão podem ser identificadas. Assim, é possível analisar cada caso sem que todo o processo de classificação precise ser refeito, tornando assim a ferramenta mais estável, e as informações geradas mais acuradas e mais atualizadas.

4.3 MAPEAMENTO GENÔMICO

As fronteiras definidas no processo de anotação destes organismos representam uma informação de relevante importância para o processo de mapeamento do conjunto de dados. Uma vez que são utilizadas para montar um mapa de características estruturais e não estruturais deste organismo. Devidamente mapeadas, as sequências do conjunto de dados podem ser classificadas quanto às estruturas que codificam, e quanto a ou as regiões que representam em relação ao genoma completo dos organismos. Esta informação posicional é utilizada principalmente para definir alvos de estudos e/ou tratamentos.

Para realizar o processo de mapeamento, inicialmente foi preciso construir um "mapa" base para os organismos em questão, a partir das suas sequências de referência. Estas sequências contém uma série de características codificadas por estas, além dos limites traducionais das mesmas. Entretanto, é preciso levar em consideração que determinadas características podem compartilhar área de codificação com outras características.

Assim, foi criado um mapa relacional de características de cada sequência de referência onde a localização e os limites de cada característica são funcionalmente independentes das outras. Sendo criado um acoplamento fraco entre as mesmas. Entretanto mantendo a dependência entre estas e a sequência de referência em questão. Foram utilizadas então *features* (Feições) cuja *key* (chave) é *'CDS'* (do inglês *Coding Sequence*).

A partir da criação dos mapas relacionais de cada organismo, foi possível proceder com o mapeamento das sequências disponíveis no conjunto de dados. Para tal, é necessário realizar um alinhamento global entre as sequências do conjunto de dados e a sequência do mapa base. Desta forma, foi possível identificar os limites resultantes do alinhamento e consequentemente a localização de determinado fragmento no genoma completo de maneira precisa.

O processo de alinhamento global pode gerar uma série de *indels* no início e fim do alinhamento. Isso se dá pelo fato de que as sequências em questão possuem tamanhos diferentes e por tanto essas áreas não podem ser levadas em consideração no mapeamento. Uma vez que se a sequência *querry* não codifica uma região, ela não possui tal característica definida no mapa do organismo para aquela região.

Com o resultado do alinhamento global, foi possível definir a região a qual a sequência mapeada em questão representa em relação ao genoma completo do organismo. Entretanto, por conta das diferentes técnicas e qualidades de sequenciamento utilizadas para coletar as sequências, nem todas alinham completamente com as características mapeadas. Sendo necessário ainda realizar um pós-processamento do resultado do alinhamento, atribuindo uma cobertura para cada característica para cada sequência.

A cobertura de um determinado alinhamento é o quanto uma sequência representa uma determinada região. Sendo que as sequências podem representar completamente ou parcialmente a região traducional de uma determinada feição. Assim, considerando o tamanho da característica e os limites do alinhamento, é possível calcular a cobertura de uma determinada sequência em relação a uma determinada estrutura codificada e anotada no genoma completo. Este cálculo é feito baseado na cobertura completa, na quantidade de inserções e/ou deleções e no tamanho de *gaps* consecutivos. Ou seja: dada uma determinada

região de tamanho x, um alinhamento de tamanho y, terá uma cobertura equivalente à diferença entre x e y, considerando ainda inserções e deleções(V) e seus tamanhos consecutivos (kL) temos a formula a seguir:

$$\text{coverage} \triangleq |(x - y) - (V * kL)|$$

4.4 COLETA DE EPÍTOPOS

Sequências de aminoácidos que formam cada epítopo estão disponíveis nos bancos de dados especializados em epítopos. Um exemplo desse tipo de conjunto de dados é o IEDB, que contém uma série de epítopos identificados em diversos organismos. Outrossim, associados a estes epítopos estão suas sequências, além de outros dados como identificação por epítopo, alelo entre outras informações (VITA, 2015).

Entretanto, este conjunto de dados não possui um serviço recursivo de exportação ou coleta de sequências. Este fato atrapalha na criação de um sistema recursivo de coleta e atualização de dados. Não obstante, as informações contidas nesse conjunto de dados são de real importância, e por tanto foi preciso criar um sistema de exportação dos dados.

Para tal, foi necessário utilizar a exportação do conjunto de dados na íntegra. Esta exportação é oferecida pelos mantenedores do conjunto de dados através do protocolo *FTP* (*File Transfer Protocol*). Esta exportação é realizada através de um arquivo compactado em formato ZIP. Apesar da disponibilidade, não é esclarecido ou informado o período de atualização a que se refere este arquivo.

Assim, foi implementado um módulo de download do conjunto de dados de epítopo. Este módulo utiliza o protocolo FTP para realizar a conexão direta com a localização do arquivo. Através desta conexão, é aberto um *Buffered Input Stream*, onde o arquivo é coletado byte a byte. Durante este processo é calculado o tamanho do *Stream* e, desta forma, é possível verificar se existem adições no conjunto de dados externo. Ou seja, caso o arquivo tenha tamanho superior ao baixado anteriormente, existem adições e é necessário realizar o download novamente.

Uma vez baixado, o arquivo é descompactado no sistema de arquivos temporário do servidor. Após este passo, é obtido um conjunto de arquivos XML. Cada arquivo contém uma série de epítopos diferentes. Assim, foi necessário realizar um processo de *parsing* em cada arquivo, percorrendo cada arquivo caractere por caractere, identificando e coletando informações.

Estas informações são inseridas em instâncias pré criadas e então é feita uma verificação local, onde cada instância é comparada ao conjunto de dados já disponível. Assim, apenas os novos epítopos são salvos no conjunto de dados, evitando duplicidade ou ausência de informações, mantendo o conjunto de dados mais conciso e confiável.

As instâncias candidatas à inserção são então separadas em conjuntos transacionais, chamados de *chunks*, estes conjuntos são inseridos em uma única transação, reduzindo assim o tempo total de armazenamento. O processo esquemático da coleta recursiva de epítopos pode ser visualizada na Figura 9.

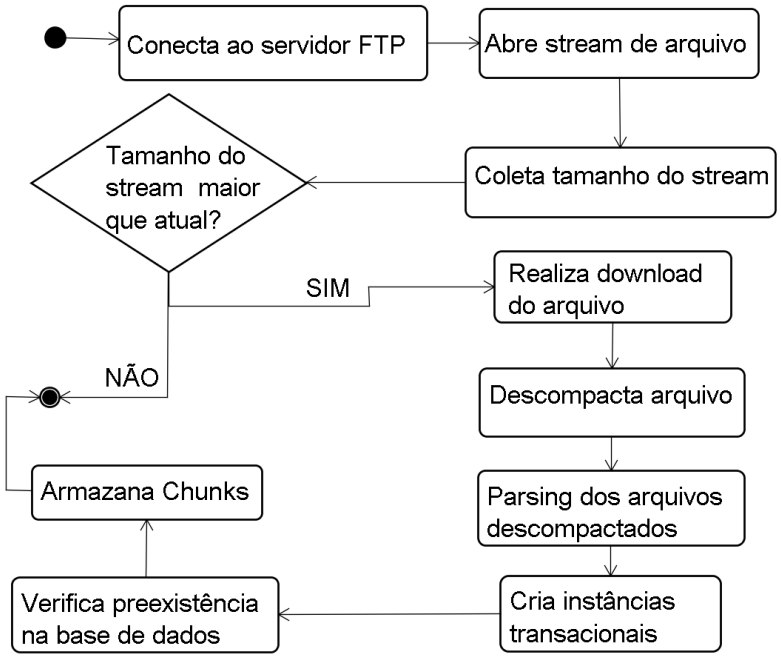


Figura 9 - Diagrama de atividades representando a coleta de epítopos do *IEDB*

4.5 MAPEAMENTO DE EPÍTOPOS

O processo de mapeamento de epítopos é feito através do mapa de características previamente criado, aliado ao conjunto de epítopos base disponível no IEDB e coletado para a base local. Este mapeamento fornecerá informações sobre frequência de epítopos no conjunto de dados do organismo. Desta forma, fornecendo candidatos para investigações sobre resposta imune. Este tipo de investigação determina bons candidatos para vacinas e fármacos mais eficazes para as etiologias associadas.

Assim, as sequências dos epítopos coletadas anteriormente foram comparados a este conjunto base, todos as traduções disponíveis nas feições das sequências do conjunto de dados. Ou seja, foram utilizadas informações de regiões codificantes já disponíveis nas sequências provenientes do NCBI. Estas informações foram então comparadas com aquelas disponibilizadas pelo IEDB.

Como resultante, foi possível observar a frequência de pareamento de cada epítopo no conjunto de dados. Eliminando assim aqueles que não geram pares ou não representam uma parcela relevante da codificação disponível nas sequências. Os epítopos remanescentes foram salvos na base de dados com cobertura completa. Ou seja, todos os aminoácidos da sequência de base do epítopo, estão dispostos na mesma ordem na feição da sequência do organismo.

Entretanto, o processo de mapeamento é computacionalmente complexo. Isto significa que um elevado poder computacional foi necessário para realizar este procedimento. Isso se dá pelo fato de haver aproximadamente um milhão e 300 mil sequências de epítopos disponíveis na base de dados e quase 100 mil traduções de sequências no conjunto de dados. Desta forma, uma grande quantidade de comparações foi necessária para que cada epítopo fosse pareado com todas as feições do conjunto de dados.

Para melhor abordar o problema, foi utilizada uma técnica computacional distribuída. Onde uma instância do problema é criada e dividida em diversas pequenas tarefas independentes entre si. Estas pequenas tarefas são distribuídas em uma série de máquinas físicas ou virtuais destinadas a receber, processar e retornar apenas o resultado daquela tarefa. As tarefas e seus resultados são escalonados por um gerenciador de tarefas, cujo papel é controlar o processo a partir de uma perspectiva global. Uma visão esquemática do processo de distribuição das tarefas de mapeamento pode ser visualizada em forma de diagrama de atividades na Figura 10.

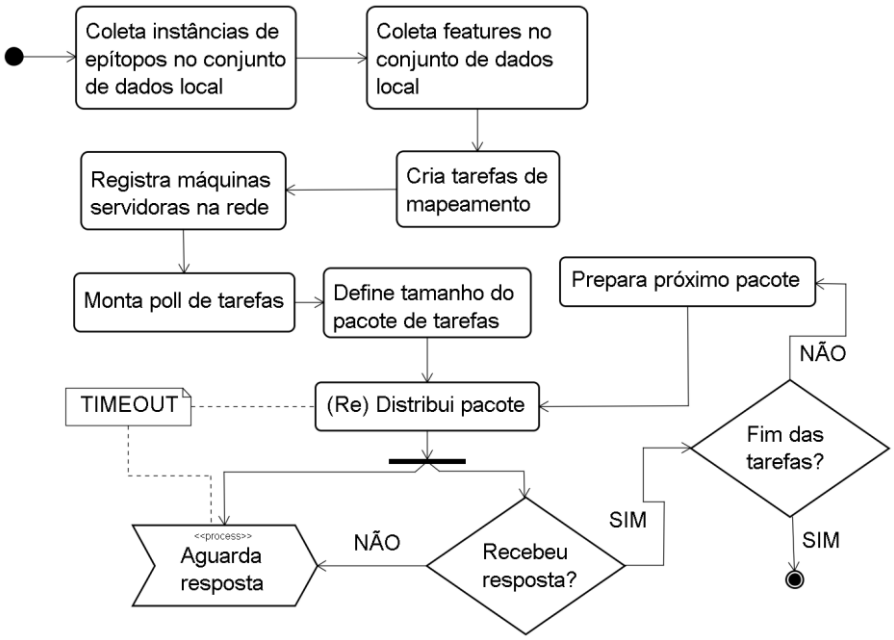


Figura 10 - Diagrama de atividades representado o escalonamento de tarefas do mapeamento genômico

4.6 DESENVOLVIMENTO DO *FRONTEND*

Após o processo de coleta e análise de informações, foi necessário distribuir publicamente os resultados, onde equipes de pesquisadores possam utilizar as informações para novos estudos sobre diagnósticos e vacinas para os agente etiológicos em questão. Além disso, o processo de vigilância epidemiológica dos agentes etiológico pode ser melhorado. Uma vez que o desenvolvimento do *frontend web* acoplado ao *backend* criado fornece uma interface rápida e prática para acompanhamento da evolução epidemiológica dos agentes, contida nos dados.

O *frontend WEB* serviu ainda como portal informativo, contendo os resultados das análises realizadas periodicamente no conjunto de dados. Esses resultados ficam em uma sessão separada para cada agente etiológico e para cada tipo de análise. Esta sessão é importante para o desenvolvimento e constante acompanhamento da expansão de conhecimento propiciada pelo conjunto de dados. Outrossim, isso torna o *frontend* responsável pelo acompanhamento da periodicidade recursiva das diversas análises realizadas pelo *backend*.

Além de uma interface amigável e intuitiva, foi desenvolvido um subsistema de *backend* para suportar as demandas associadas da comunidade. Este subsistema integra uma API com padronização REST e com microsserviços associados ao acesso direto aos dados. Além disso, existem *endpoints* direcionados ao acesso às funções específicas e algoritmos exclusivos.

Assim, caberá ao subsistema do *frontend* em conjunto com a *API RESTful* definir os momentos e a periodicidade em que os dados serão atualizados e analisados. Desta forma, o conjunto de dados passará a ser independente de curagem e praticamente livre de manutenção. Este processo de curagem e manutenção representa um dos principais desafios em manter um conjunto de dados biológico atualmente, uma vez que isso depende de recursos e pessoal especializado. Assim, o conjunto de dados e o sistema em si se tornam mais completos, seguros, autônomos e baratos. Além de desempenharem técnicas de ponta de análise e oferecerem informações relevantes para o desenvolvimento de vacinas e tratamentos eficazes para as etiologias associadas aos organismos em questão.

5 RESULTADOS

5.1 BANCO DE DADOS SAGA

A vigilância epidemiológica e sorotípica dos agentes etiológicos em questão, representa um relevante desafio no contexto atual. Desta forma, o desenvolvimento da plataforma de acompanhamento, visualização e operação de dados, apresenta uma nova perspectiva para o estudo e controle dos agentes.

O banco de dados SAGA é uma plataforma centralizada de informações sobre os organismos em questão. Esta plataforma aborda diversos aspectos como classificação, frequência de regiões subgenômicas, imunogenicidade e georreferenciamento. Todas organizadas por organismo, possibilitando ainda a interação com o usuário. Uma vez que o software recebe entradas do usuário através de movimentos de *mouse* e entradas de funções. Oferecendo informações personalizadas, mesmo mantendo um visual limpo e elegante.

Na página, é possível acompanhar o crescimento do conjunto de dados, suas características principais e os desenvolvimentos da equipe. Para cada organismo, é possível visualizar diversos aspectos do conjunto de dados e das informações contidas no mesmo.

A exemplo do ZKV, onde é possível verificar (Figura 11) a quantidade atual de sequências no conjunto de dados, crescimento do conjunto de dados específico em relação ao conjunto de dados global e mapa de georreferenciamento de dados do conjunto. Passando o *mouse* sobre cada país é possível verificar o valor bruto de submissões deste país, e sobre os pontos nas linhas do gráfico de crescimento para verificar este valor bruto.

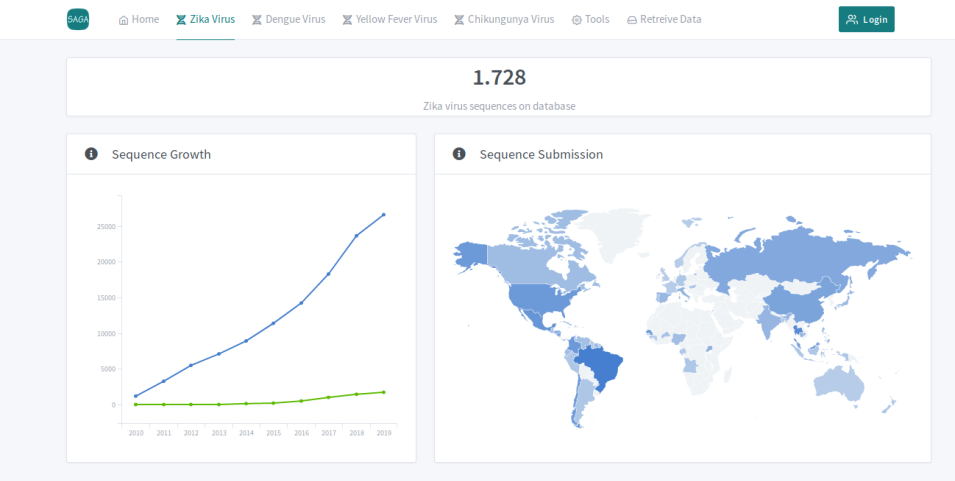


Figura 11 - Recorte de tela do protótipo funcional da plataforma com representação gráfica do crescimento do conjunto de dados e georreferenciamento das sequências por organismo

É possível ainda obter informações atualizadas a respeito da subtipagem, para cada organismo, onde é possível ver a separação das sequências por subtipo. A exemplo do YFV

como pode ser visto na Figura 12. Este gráfico é portável e pode ser extraído da plataforma a qualquer momento pelo usuário.

Acessando a área destinada às regiões genômicas do organismo, a exemplo do DENV (Figura 13) é possível verificar a distribuição das sequências no conjunto de dados em relação ao mapa genômico criado. As informações são separadas pelas sequências de genoma completo e pelas sequências com cobertura parcial. Desta forma é possível acompanhar a qualidade e completude do conjunto de dados à medida que o banco de dados aumenta.

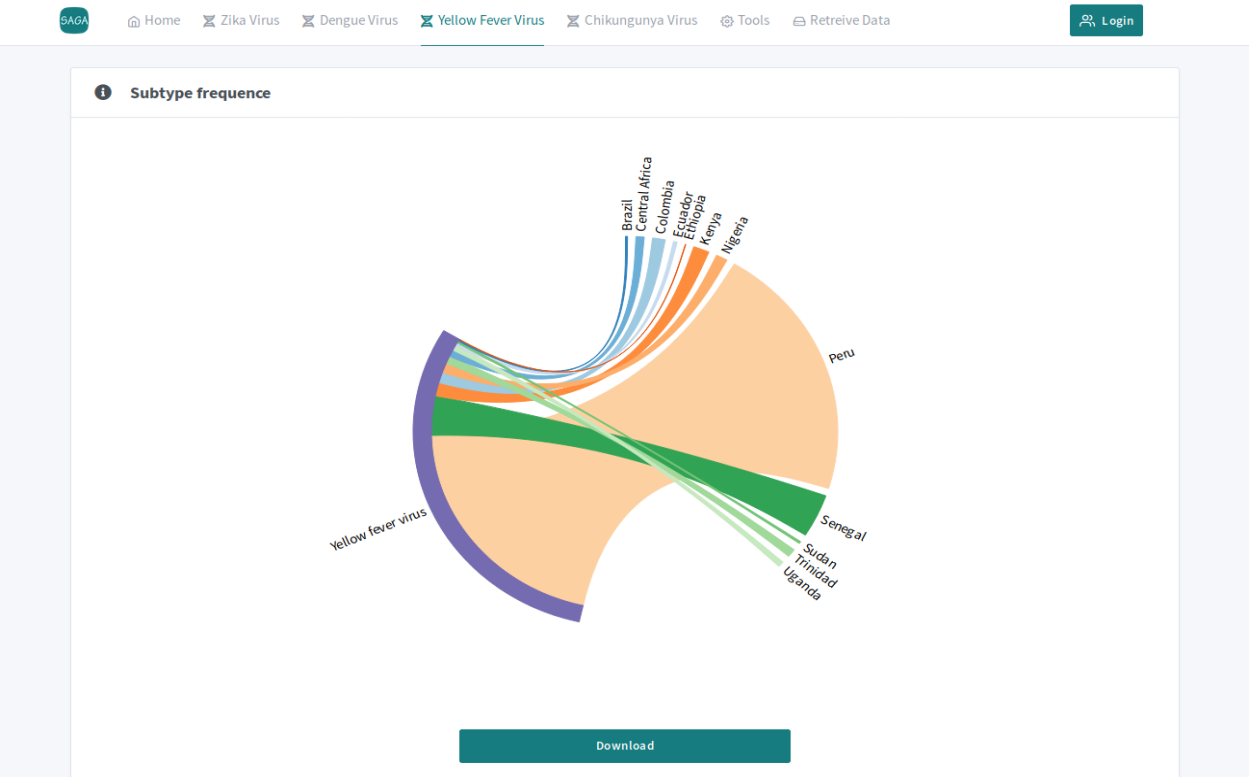


Figura 12 - Recorte de tela do protótipo funcional da plataforma com representação gráfica da distribuição genotípica das cepas do conjunto de dados por organismo.

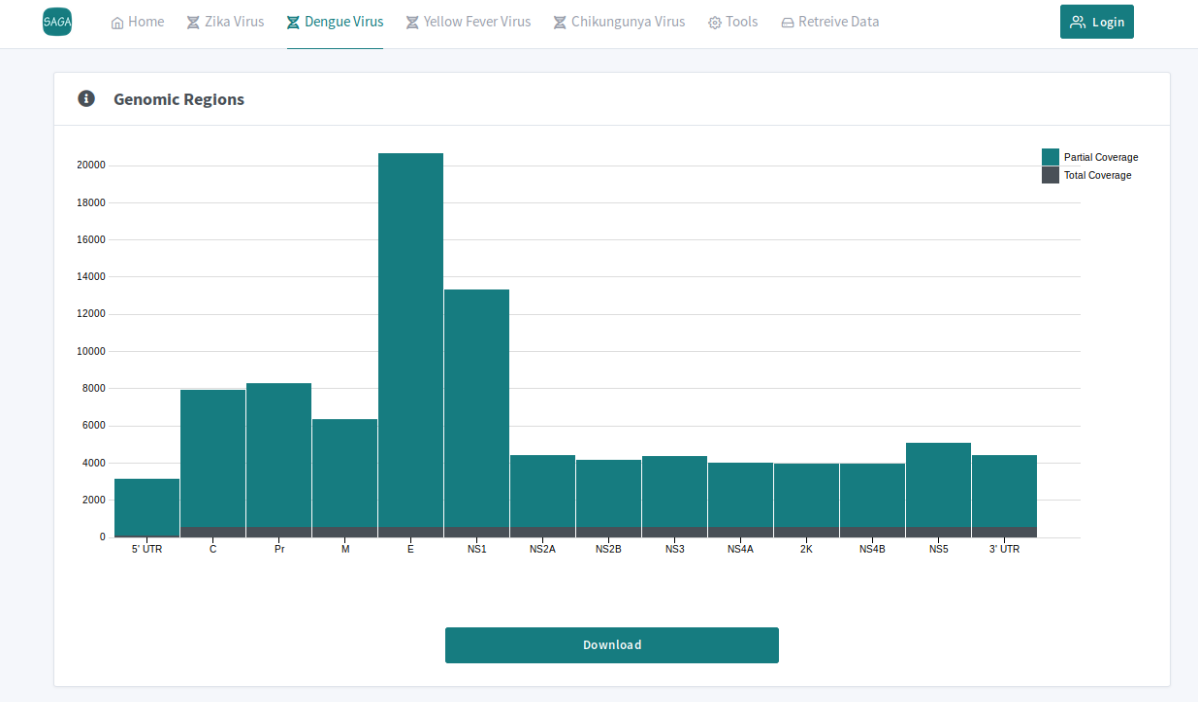
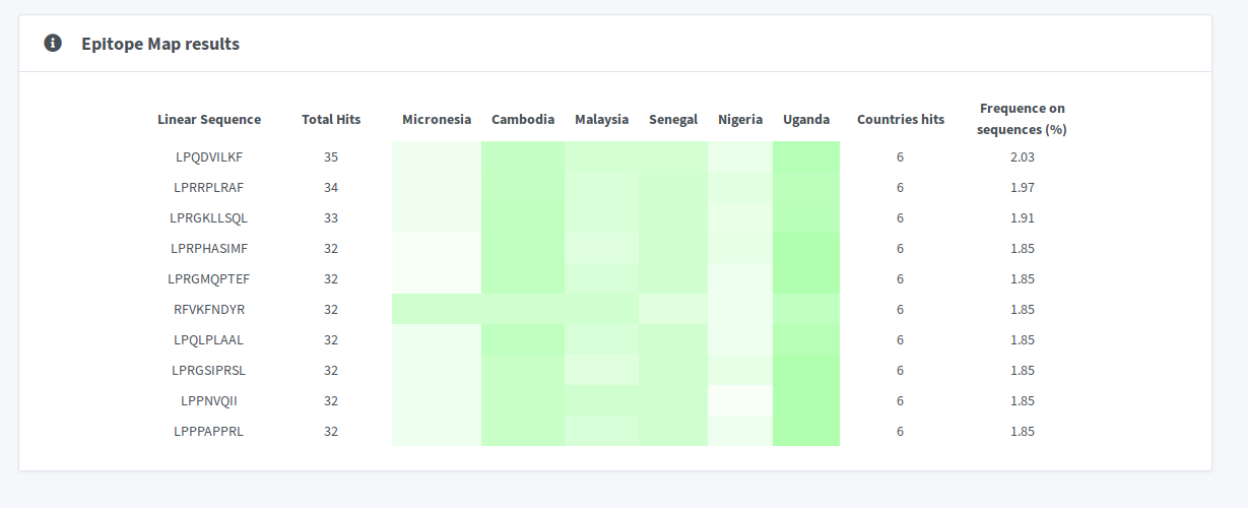


Figura 13 - Recorte de tela do protótipo funcional da plataforma com representação gráfica da distribuição das cepas do conjunto de dados em relação ao genoma completo dividindo entre cobertura total e parcial.

Para cada organismo é disponibilizada ainda uma tela contendo os 10 epítopos mais frequentes, usando escala de cor gradual. No exemplo do ZKV (Figura 14), a frequência deste epítopo nos subtipos do conjunto de dados pode ser observada. Além disso, é possível obter informações quantitativas sobre os países com sequências que possuem associação com um determinado epítopo. Interessante observar que os 10 epítopos que são comuns entre as 6 linhagens endêmicas do ZKV mapearam com sequências de 6 países.

Na seção de ferramentas (*Tools*) é possível acessar a ferramenta de mapeamento rápido de epítopos. Onde é possível entrar com a sequência de aminoácidos do epítopo e uma chamada ao *endpoint* da API é feito e após o processamento, a frequência relativa separada por subtipo para todos os organismos é exibida. Facilitando assim o processo de análise de epítopos novos ou putativos utilizando o conjunto de dados completo. Na Figura 15 é possível ver o mapeamento do epítopo “GKAKGSRAIWYMWLG” que é comum entre os arbovírus, exceto CKV.



Epitope Map results

Linear Sequence	Total Hits	Micronesia	Cambodia	Malaysia	Senegal	Nigeria	Uganda	Countries hits	Frequence on sequences (%)
LPQDVILKF	35							6	2.03
LPRRPLRAF	34							6	1.97
LPRGKLLSQL	33							6	1.91
LPRPHASIMF	32							6	1.85
LPRGMQPTEF	32							6	1.85
RFVKFNDYR	32							6	1.85
LPQLPLAAL	32							6	1.85
LPRGSIPRSL	32							6	1.85
LPPNVQII	32							6	1.85
LPPPAPPRL	32							6	1.85

Figura 14 - Recorte de tela do protótipo funcional da plataforma com representação gráfica dos epítopos mais frequentes no conjunto de dados do organismo, por subtipo e com representação do universo de países os quais as cepas associadas estão relacionadas.

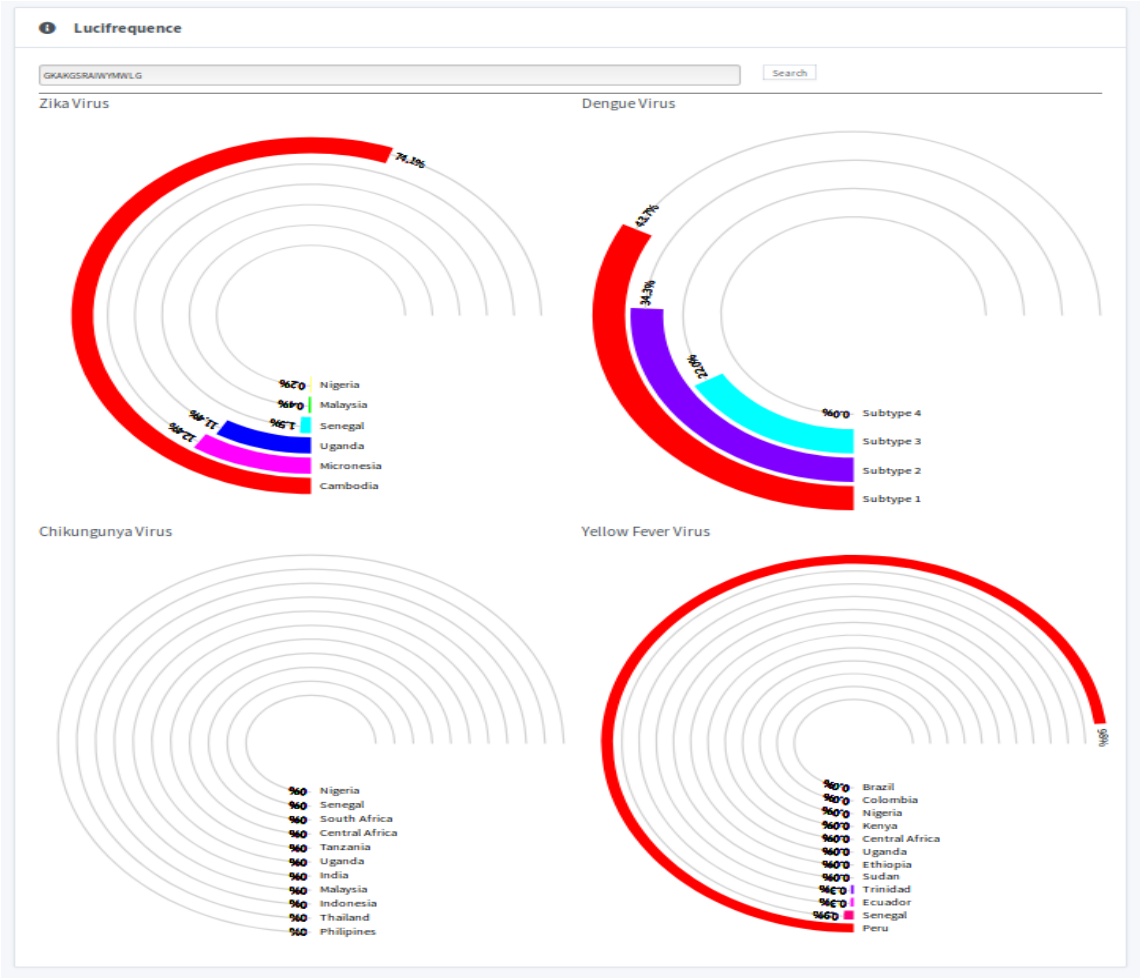


Figura 15 - Recorte de tela do protótipo funcional da plataforma com representação gráfica da execução do mapeamento do epítopo "GKAKGSRAIWYMWLG" no conjunto de dados e seu respectivo resultado.

5.2 *STATUS* DO CONJUNTO DE DADOS

Após coleta intensiva de dados, subtipagem de sequências, mapeamento das mesmas no genoma completo, mapeamento de epítopos e compilação dos dados, o conjunto de dados foi filtrado e sumarizado, buscando identificar o *status* do mesmo em julho de 2019. Considerando o crescimento do conjunto de dados e a recursividade das análises, o conjunto de dados atual pode não refletir o extrato feito no momento supracitado. Isto porque, novas submissões são feitas, análises refeitas, novos dados, resultados agregados periodicamente.

O tamanho do conjunto de dados é relativo à perspectiva que se quer analisar. O conjunto de dados central, ou seja, o conjunto de sequências coletadas é o de menor tamanho em armazenamento, uma vez que cerca de 70% dos dados são gerados e armazenados após análise e processamento destes. A extensão do conjunto de dados assim como o crescimento de cada organismo pode ser visto na Tabela 5.

Tabela 5 - Extensão do banco dados em 2019 e crescimento de cada organismo no período de 2017 a 2019.

Organismo	Total de Sequências	Crescimento %
ZKV	1442	62%
YFV	813	22%
DENV	25821	21%
CKV	4008	10%

A sequência coletada de menor tamanho, possui apenas 41 pb, enquanto a maior sequência coletada possui 15739 pb. Destas sequências em média temos ~22% de sequências que possuem o genoma completo do agente.

No processo de subtipagem e mapeamento de sequências são gerados dados de alinhamento, similaridade e referência das sequências individualmente, escalando consideravelmente o custo computacional da ferramenta. Este conjunto de dados equivale à n^m, sendo “n” o número total de sequências e “m” o total de referências dos grupos genotípicos, uma vez que cada sequência é comparada com o conjunto de referências que por sua vez é comparada a todo o conjunto. Desta forma, observamos que o crescimento do conjunto de dados é exponencial.

Foram coletados ainda 830.102 epítopos até a data de corte. Estes epítopos representam uma fração relativamente pequena do custo computacional da ferramenta, uma vez que possuem pouca informação associada diretamente. No modelo de dados, a maioria

das informações associadas aos epítopos são referências cruzadas de outros conjuntos de dados, facilitando o seu gerenciamento.

Entretanto o mapeamento de epítopos é a análise que mais gera dados, uma vez que cada sequência pode ter diversas áreas de tradução e o número absoluto de epítopos é relativamente grande. Assim, um total de 86.557.225.846 mapeamentos foram realizados para este conjunto de dados, resultando num subconjunto de 28.005.257 *hits*. Estes mapeamentos ocupam a maioria do espaço reservado para o banco de dados e é responsável pelo maior tempo de processamento, análise e custo computacional da ferramenta.

5.3 OBTENÇÃO DE SEQUÊNCIAS NUCLEOTÍDICAS

Durante o processo de obtenção de sequências, diversos desafios foram encontrados. É possível destacar que os principais desafios neste processo são de natureza tecnológica. Isso se dá pelo fato de que limitações de serviços e de recursos são aplicadas pelos detentores das informações. Diversas limitações de acesso aos serviços de download de sequências são aplicadas no intuito de mitigar possíveis ataques de softwares maliciosos, garantindo integridade dos dados e serviços oferecidos pelo provedor.

Desta forma, foi preciso montar uma estratégia de obtenção de sequências que levasse em consideração as limitações impostas, tanto por parte do provedor de serviços, quanto as limitações tecnológicas aplicadas a qualquer processo de download. Dentre essas limitações estão principalmente destacadas as limitações de rede. Onde o limite de banda e taxa de dados reduzem a disponibilidade de serviços.

Considerando os desafios impostos, o processo de obtenção das sequências nucleotídicas foi realizado com sucesso. Sendo criada uma estratégia de adequação aos limites encontrados, o tempo de download do conjunto de sequências de cada organismo não foi proporcional ao seu tamanho real. Assim, devido ao tamanho e complexidade do processo de download, o algoritmo que realiza o processo é de complexidade O(n), sendo n a quantidade de sequências disponíveis. Entretanto, feita uma análise de pior hipótese, a relação de complexidade do tempo levado pelo algoritmo em relação à quantidade de sequências disponíveis para *download* é $O(n^2)$.

Assim, em 4 dias foi realizado o *download* inicial das sequências disponíveis do GenBank. Mais precisamente no banco de dados nucleotide. Foram adquiridas inicialmente 20.325 sequências associadas com o organismo Dengue Virus. Relacionadas ao Chikungunya

Virus, foram obtidas 3.613 sequências, 545 sequências referentes ao Zika Virus e 637 sequências relacionadas ao vírus causador da febre amarela.

Das sequências obtidas, relacionadas ao DENV, 6512 de 25821 são genomas completos. Enquanto 456 de 4008 das sequências de CKV obtidas inicialmente representam genomas completos as de ZKV apresentam 650 de 1442 de genomas completos e 130 de 813 das sequências obtidas de YFV são genomas completos. Isso significa que das 32.804 sequências obtidas, 7.748 (23,61%) são genomas completos como pode ser visto na Figura 16.

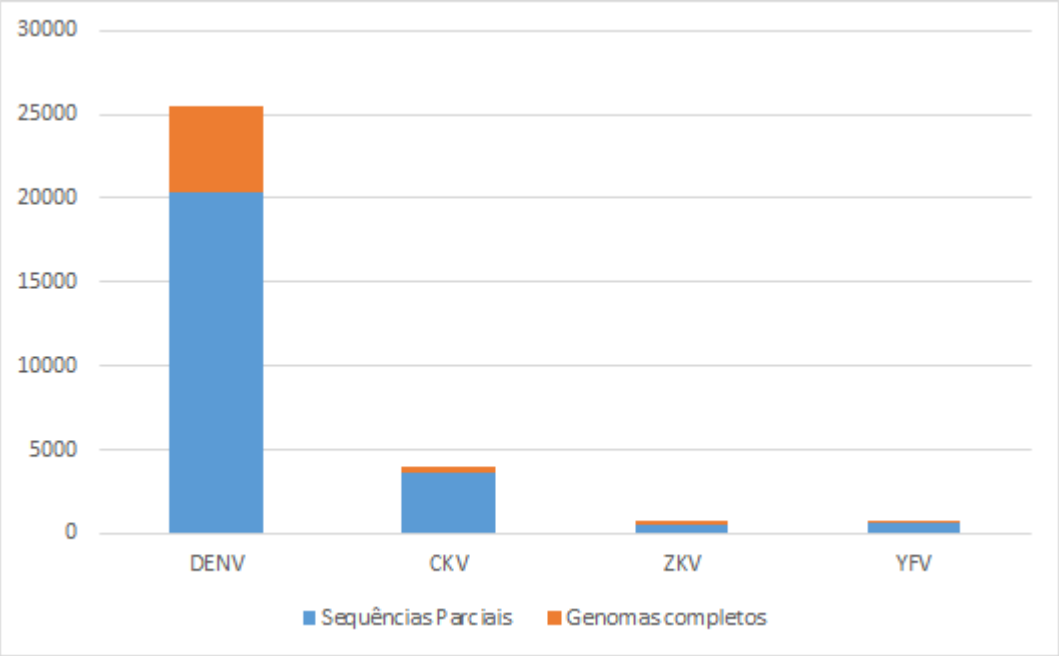


Figura 16 - Gráfico de barras do número de sequencias de cada vírus mapeado: A cor azul representa as sequencias que cobrem parcialmente o genoma enquanto a cor laranja representa as sequencias que cobrem o genoma completo.

As sequências contidas no conjunto de dados representam aquelas obtidas inicialmente. A primeira aquisição foi realizada no mês de julho de 2017. Em novembro de 2017 foi realizada nova obtenção e assim por diante. Mesmo tendo iniciado o processo de coleta em 2017, com a informação de data de submissão das sequências, é possível retro datar as sequências montando um crescimento temporal mesmo maior que o tempo de existência do SAGA. Assim, foi possível observar a taxa de crescimento do conjunto de dados por organismo como pode ser visto na Figura 17.

Observando o crescimento do conjunto de dados de ZKV de 2015 a 2019, é possível identificar um pico (140%) de interesse e por tanto de submissão de sequências do agente etiológico de 2015 a 2016 e uma posterior queda (-100%) até 2019. Observasse ainda um crescimento de sumbissões (~35%) de sequências de YFV entre 2017 e 2018, uma vez que houve surto deste agente etiológico no Brasil neste ano.

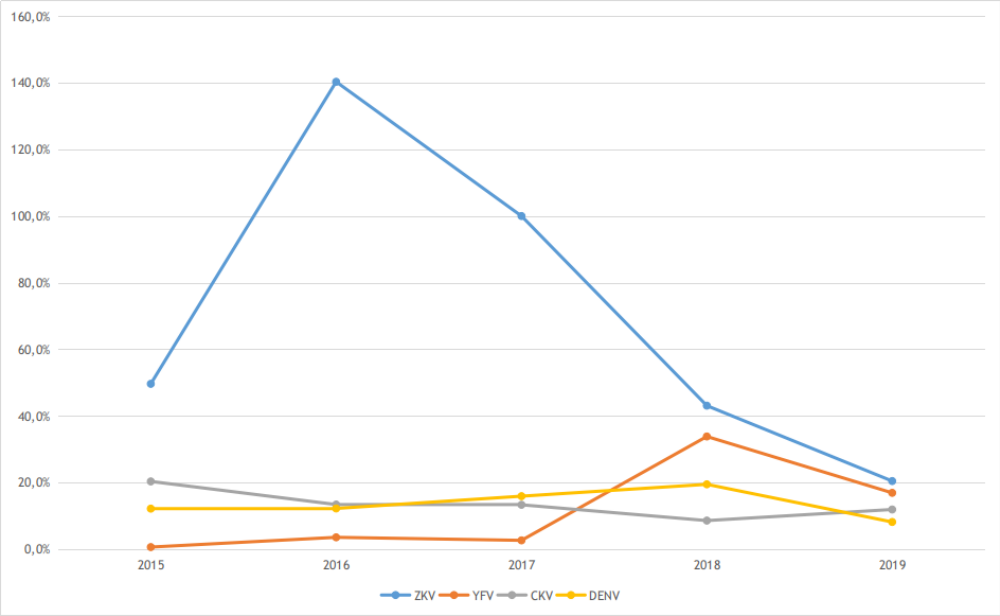


Figura 17 - Gráfico de linhas representando a taxa de crescimento não cumulativa da quantidade de sequencias de cada vírus de 2015 a 2019.

5.4 CLASSIFICAÇÃO GENOTÍPICA

Considerando a estratificação por organismo, foi possível observar que o subtipo originário e endêmico da Tanzânia do CKV é mais presente no conjunto de dados, representando 73,11% da classificação. Considerando a classificação das sequências do DENV, observa-se que os subtipos 1 e 2 são mais frequentes, com 34,54% e 32,85% respectivamente.

Analisando as sequências associadas ao vírus causador da febre amarela, foi identificado que o subtipo peruano é mais frequente, representando 59,18% da classificação. Assim como as sequências do ZKV. Onde, durante a classificação, foi observado que o subtipo originário do Camboja é mais frequente, representando 51,35%.

Na Figura 18, é possível avaliar graficamente a representação das percentagens gerais de cada genótipo viral por vírus. A figura demonstra uma plotagem circular onde são criadas relações entre o genótipo e o vírus. São então calculados os pesos proporcionais de cada

subtipo em relação à classificação total sendo representadas em cores diferentes para cada organismo.

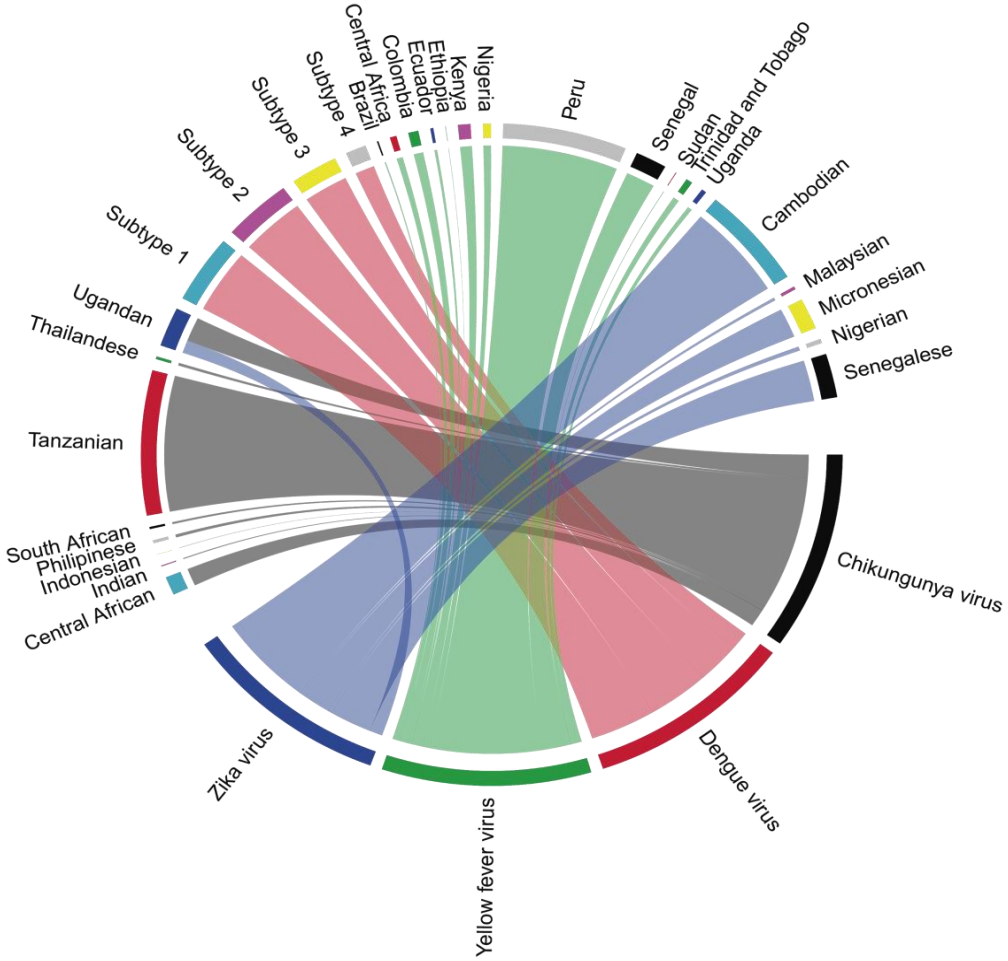


Figura 18 - Gráfico de plotagem circular representando as percentagens gerais de classificação das cepas do conjunto de dados em cada subtipo associado ao organismo. Cada organismo recebe uma cor para relação, e a largura dos fusos representa a percentagem relativa ao organismo associado. n=32.804

Foi realizada ainda a estratificação da percentagem de cada genótipo por país. Esta informação pode ser obtida através dos meta dados do XML principal do Genbank ou através de registros no banco de dados de coleta de amostras (*biosample*), o último sendo mais preciso que o primeiro. Esta separação paramétrica possibilita observar as submissões em regiões endêmicas em relação à acurácia de classificação. Estas informações podem vir a ser relevantes para o processo de vigilância epidemiológica dos organismos em questão. Assim, é

possível observar o local ou país de submissão de determinada cepa, associando a mesma ao genótipo classificado, criando assim um perfil endêmico para cada subtipo.

É possível observar a plotagem circular (Figura 19) que representa a estratificação por país de submissão de cepas classificadas nos genótipos do DENV. As cepas que não possuem a definição de localização no campo *country* estão classificadas como n*ot available* (não disponível).

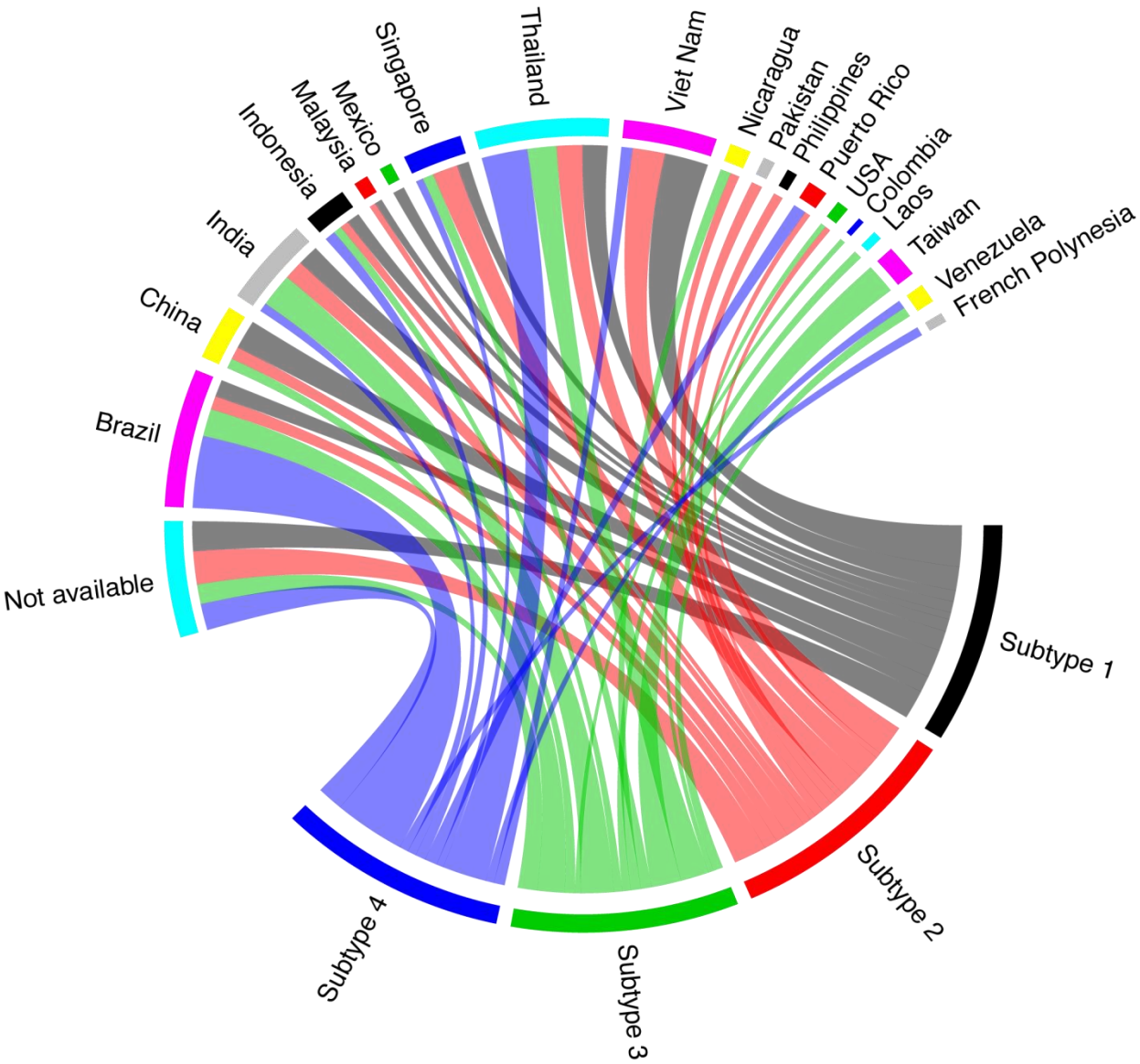


Figura 19 - Gráfico de plotagem circular representando a distribuição das sequências classificadas em cada subtipo do DENV para os países para as que possuem esta informação e N*ot available* para as que não possuem. n=25.821

É possível observar, separado por país de submissão a classificação genotípica das sequências do CKV. Destaca-se a prevalência do subtipo Senegalese no país Semnegal.

Considerando ainda que a maior concentração de sequências do subtipo Central-Africano é na Índia (45,37%). Enquanto a maioria das sequências do subtipo Indiano não possui localização definida (94,12%) como visto na Figura 20.

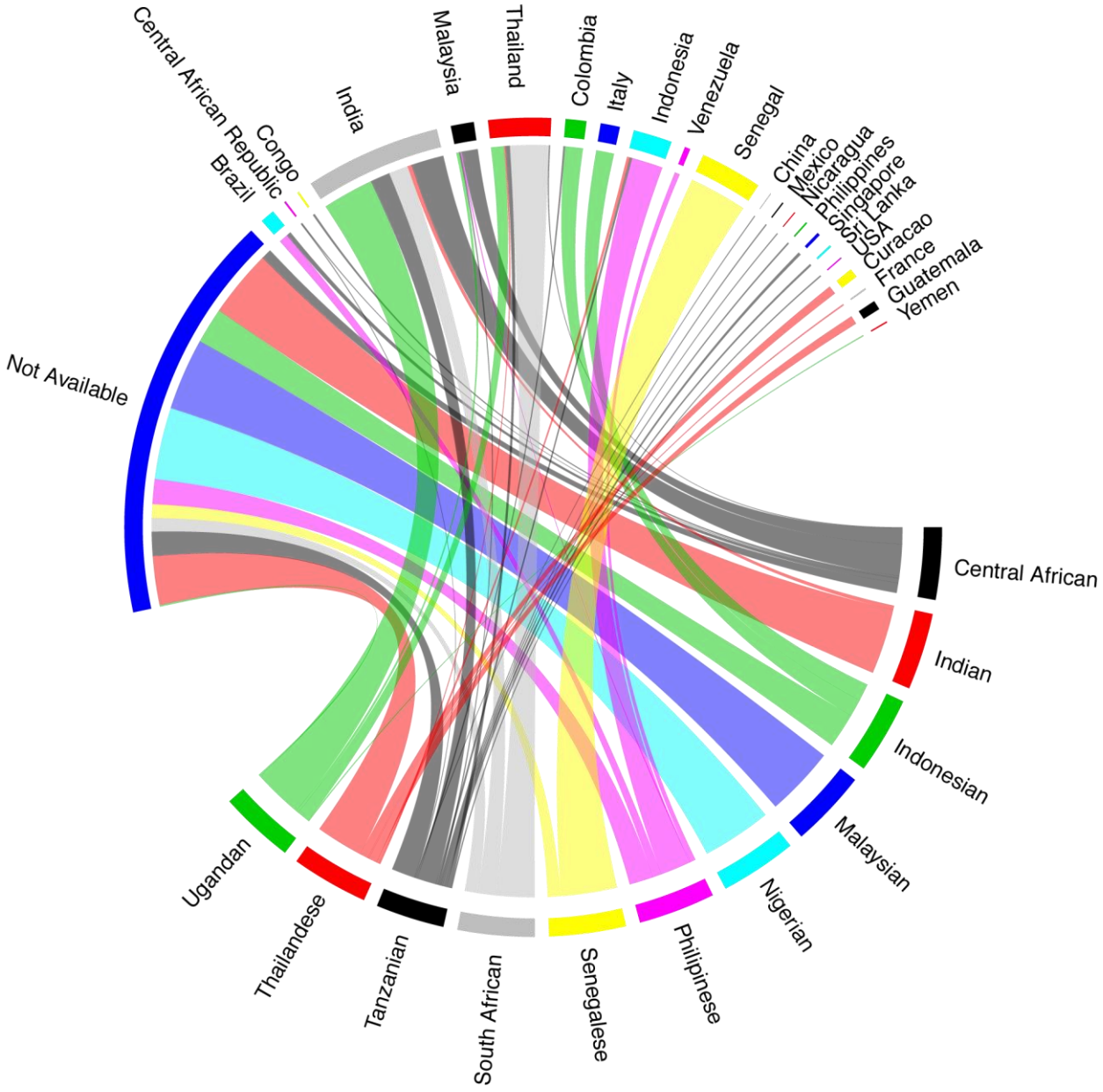


Figura 20 - Gráfico de plotagem circular representando a distribuição das sequências classificadas em cada subtipo do CKV para os países para as que possuem esta informação e *not available* para as que não possuem. n=4.008.

Observando a Figura 21, onde a classificação genotípica do ZKV separada por país é representada, é possível observar sua frequência geográfica. Com isso, é possível observar que o subtipo associado geograficamente ao Camboja, possui sequências bem distribuída entres os países. Além de este genótipo ser o mais frequente grupo classificatório. Entretanto, o genótipo identificado como endêmico na Malásia possui, em sua maioria, sequências identificadas e submetidas neste país.

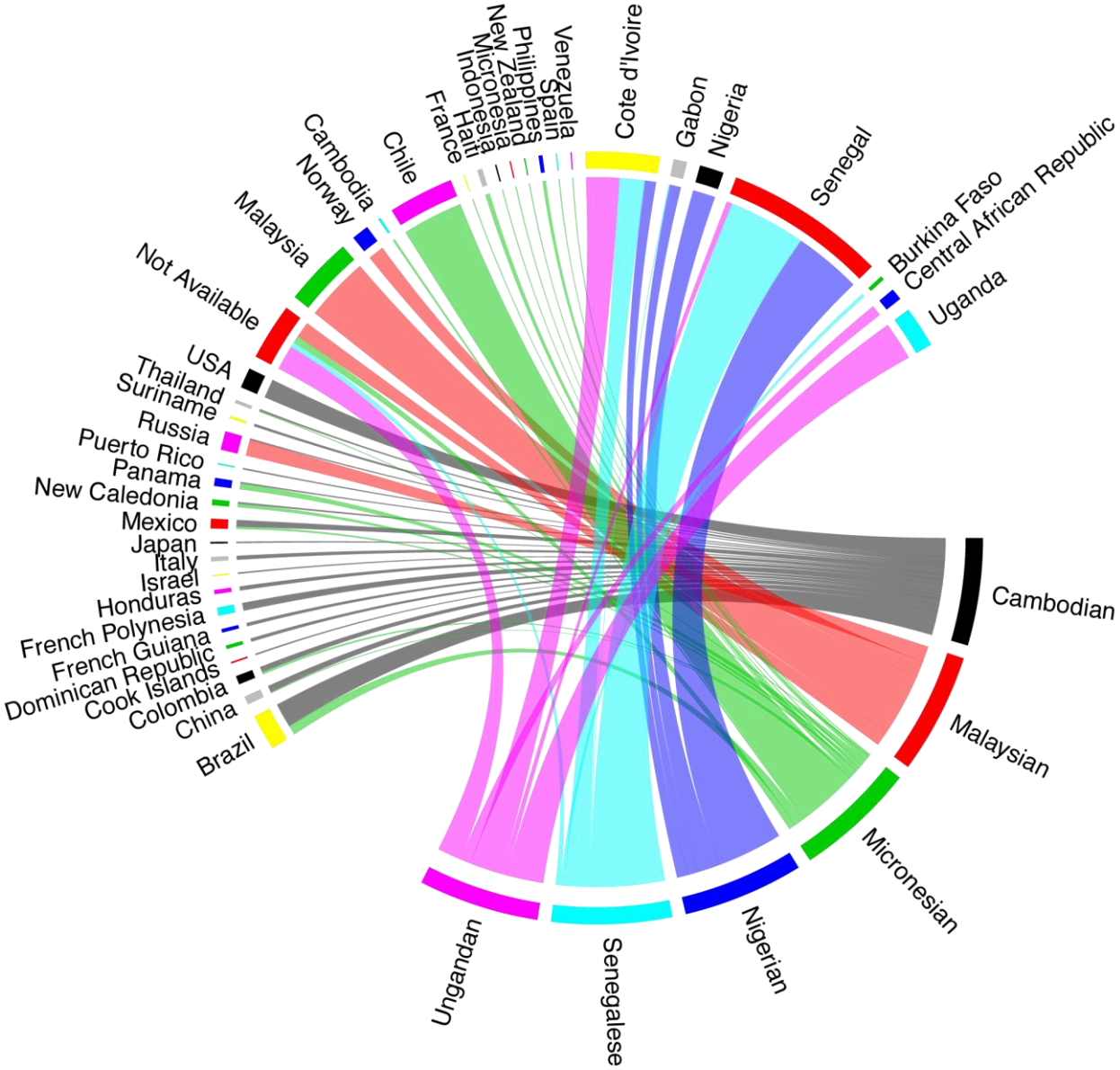


Figura 21 - Gráfico de plotagem circular representando a distribuição das sequências classificadas em cada subtipo do ZKV para os países para as que possuem esta informação e *not available* para as que não possuem. n=1442.

Ainda sobre o organismo causador da febre zika, foram estratificados os perfis de submissão das sequências deste organismo por subtipo no Brasil de 2015 a 2019. Desta forma

criando uma comparação na série temporal, demonstrando a alteração de perfil genotípico das submissões feitas no país neste período, como pode ser visto na Figura 22.

Nesta análise, foi possível observar que em 2019 foram submetidas sequências do subtipo Uganda, que além de não aparecer nos anos anteriores, representa uma diferente linhagem do agente etiológico (Africana) em detrimento da Asiática exibida nos anos anteriores.

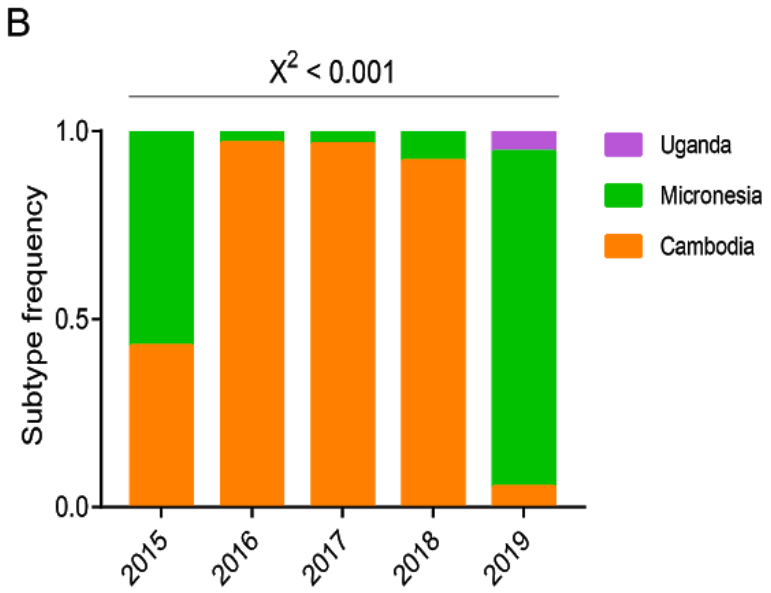


Figura 22 - Gráfico de barras sobrepostas representando a proporção do perfil genotípico de submissão de sequências de ZKV de 2015 a 2019.

Outrossim, na Figura 23, observa-se a plotagem circular estratificada das cepas classificadas por país do YFV, neste, é possível observar que a maioria das sequências não possui informação sobre a localização no campo *country* (país). Entretanto, das remanescentes, observa se que a maior frequência de sequenciamento entre os subtipos ocorreu no Brasil, que apresentou todos subtipos do YFV ocorrendo no país. Com uma maioria de sequências do subtipo Colombiano.

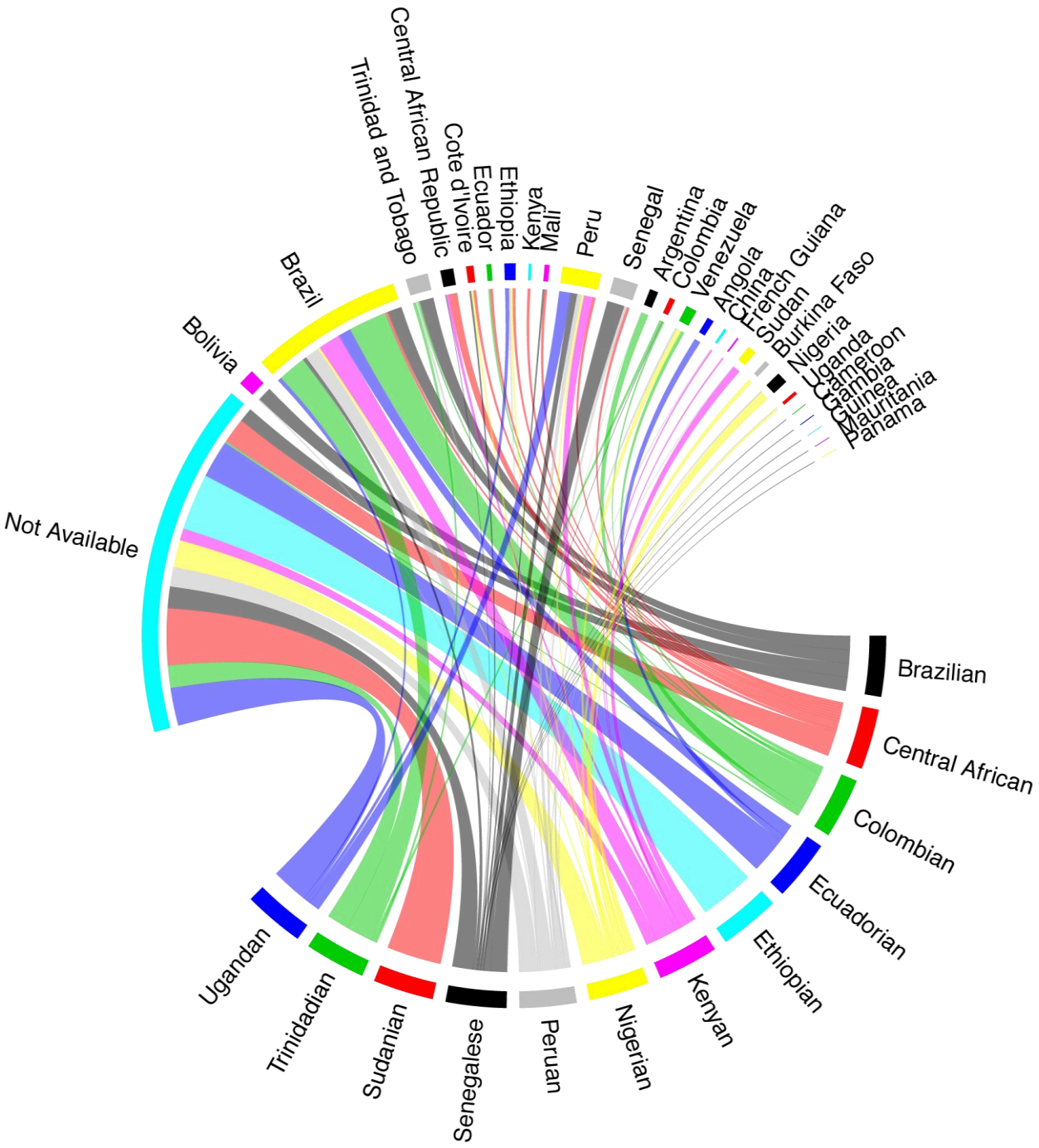


Figura 23 - Gráfico de plotagem circular representando a distribuição das sequências classificadas em cada subtipo do YFV para os países para as que possuem esta informação e *not available* para as que não possuem.n=813

5.5 MAPEAMENTO GENÔMICO

Informações traducionais sobre as sequências são de grande valia para compreender diversos aspectos de organismos virais. Estas características são associadas diretamente às regiões gênicas que podem ser classificadas em genes estruturais e não estruturais,

apresentando funções diversas. As mesmas são anotadas nos genomas de referência, contendo os coordenadas de tradução das mesmas.

Os limites das regiões codificantes permite criar um mapa de características de cada genoma viral. Este mapa foi utilizado para indexar as sequências disponíveis no banco de acordo com seu respectivo genoma. O processo de mapeamento fornece informações posicionais sobre a disposição genômica do organismo no conjunto de dados. Estas informações podem ser utilizadas para definir conjuntos de dados específicos e regiões alvos para estudos e/ou tratamentos e vacinas.

Foram identificadas as frequências para cada região gênica. Entretanto, nem todas as regiões gênicas foram completamente alinhadas durante o mapeamento por que algumas sequências são fragmentos genômicos. Assim, quando apenas parte de uma região é coberta pelo fragmento, é definido que há apenas cobertura parcial. Enquanto se o fragmento em questão corresponde e cobre completamente uma região gênica, este é classificado como total.

Na Figura 24, é possível observar a distribuição de mapeamento das sequências disponíveis no banco no genoma do vírus causador da febre chikungunya. Nesta é possível observar que a região de maior frequência no conjunto de dados é a região gênica que codifica o envelope viral. Outrossim, observa se que a maioria das sequências mapeadas nesta região, se trata de fragmentos, e portanto cobrem apenas parcialmente a região E1 (35,95% contra apenas 11% de coberturas totais). Entretanto, a região E2 possui uma quantidade maior de coberturas totais (18,77% contra 8,46% de parciais).

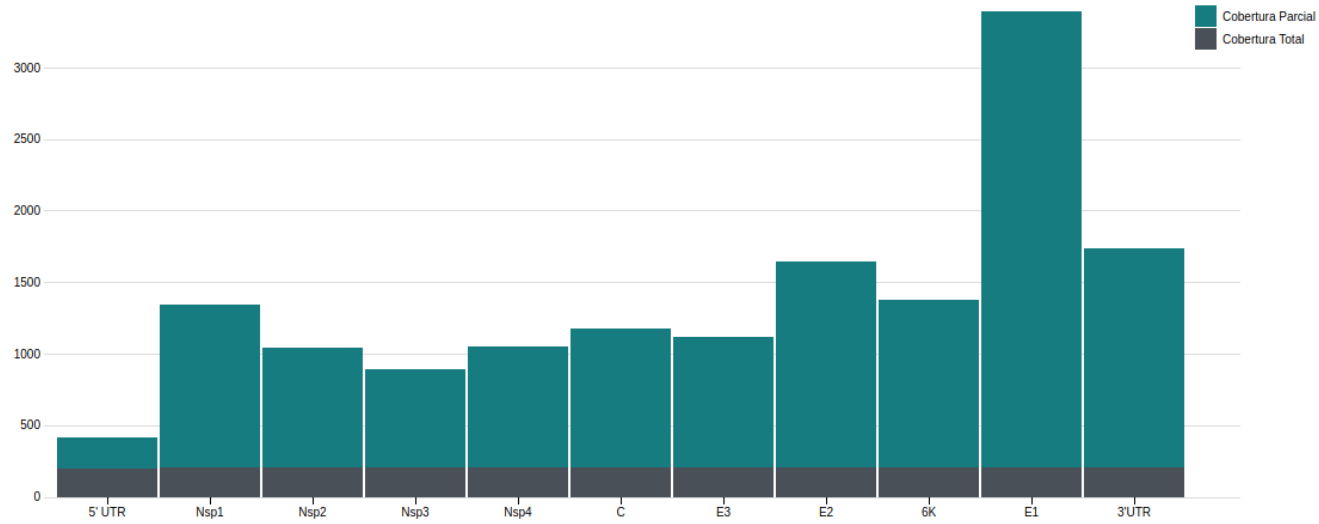


Figura 24 - Gráfico de barras representando a cobertura em % realizada pelo mapeamento genômico das cepas coletadas de CKV. Em cor cinza, as cepas em que o mapeamento compreende completamente a região gênica e em verde as sequências cujo mapeamento cobre parcialmente a região.

A distribuição de mapeamento das sequências disponíveis no banco no genoma do agente etiológico causador da dengue pode ser vista graficamente na Figura 25. É possível observar que a região mais presente é, também, a que codifica o envelope viral. Esta região é coberta por todas as sequências do conjunto parcial ou completamente. Ou seja, todos os mapeamento compreendem esta região gênica. Entretanto, 25,59% das sequências, possuem *indels* neste alinhamento.

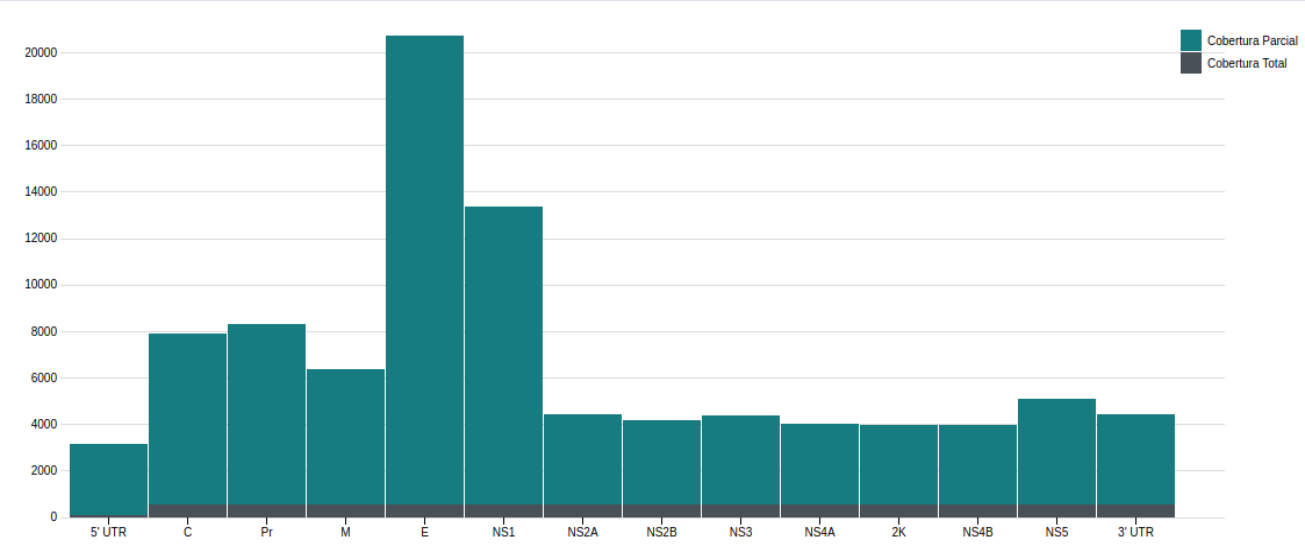


Figura 25 - Gráfico de barras representando a cobertura em % realizada pelo mapeamento genômico das cepas de coletadas de DENV. Em cor cinza, as cepas em que o mapeamento compreende completamente a região gênica e em verde as sequências cujo mapeamento cobre parcialmente a região.

Já na Figura 26, é possível observar o resultado da distribuição de mapeamento das sequências disponíveis no banco no genoma do *Zika virus*. Nesta é possível observar que a região mais frequente também é a que codifica as proteínas do envelope viral, presente em metade do conjunto de dados (50,46%). Entretanto, no caso deste agente etiológico, as outras regiões gênicas também estão bem representadas. Demonstrando assim, que este agente etiológico possui um conjunto de dados relativamente mais coeso. Ainda observamos que para quase todas as regiões gênicas a grande maioria das coberturas é total, mostrando um menor grau de fragmentação.

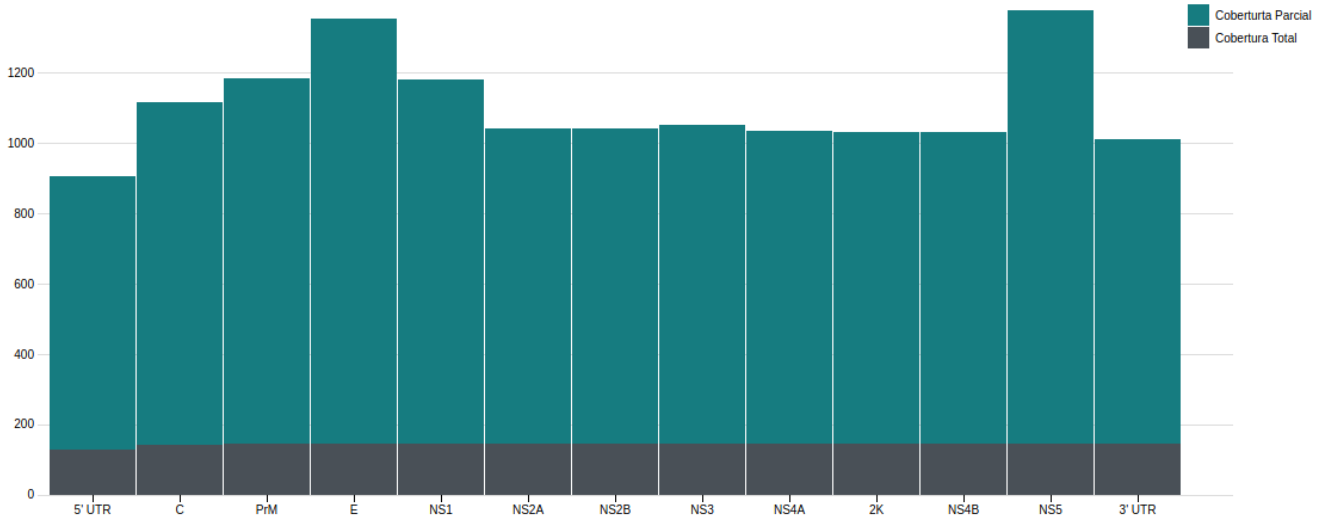


Figura 26 - Gráfico de barras representando a cobertura em % realizada pelo mapeamento genômico das cepas de coletadas de ZKV. Em cor cinza, as cepas em que o mapeamento compreende completamente a região gênica e em verde as sequências cujo mapeamento cobre parcialmente a região.

O mapeamento do vírus da febre amarela demonstrou um padrão semelhante aos outros arbovírus. Na Figura 27, é possível observar que a região do envelope é bem frequente no conjunto de dados. Entretanto, a região mais frequente é aquela que codifica a pré membrana (prM - 33,59%). Esta região tem alta frequência de pareamentos parciais e baixa frequência relativa de pareamentos completos ou totais (19,78%).

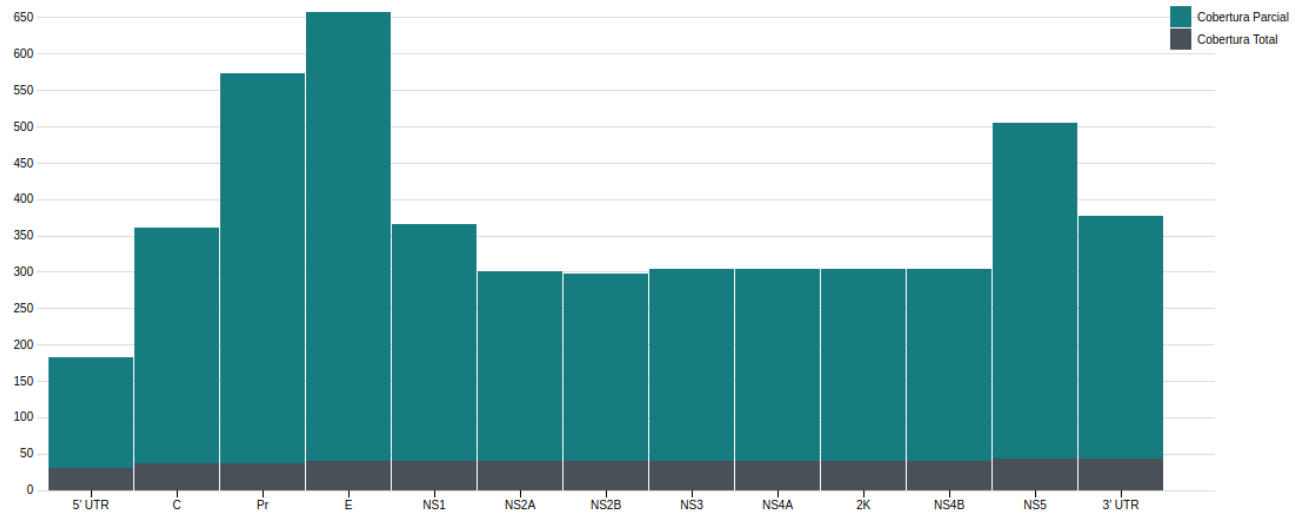


Figura 27 - Gráfico de barras representando a cobertura em % realizada pelo mapeamento genômico das cepas de coletadas de YFV. Em cor cinza, as cepas em que o mapeamento compreende completamente a região gênica e em verde as sequências cujo mapeamento cobre parcialmente a região.

5.6 OBTENÇÃO DE EPÍTOPOS

A identificação de regiões imunogênicas nos genomas dos agentes etiológicos é um importante ponto de influxo de informações para a plataforma. No entanto, o *IEDB* não possui serviço de acesso externo aos dados. Isso se apresentou como um importante desafio na coleta de epítopos para posterior mapeamento no conjunto de dados. Outrossim, a ausência desse serviço dificulta o processo de recursividade proposto pela plataforma SAGA. Dessa forma um processo alternativo de download do conjunto de dados via FTP foi criado.

Durante o processo inicial, 779.362 epítopos foram coletados em apenas duas horas. Entretanto, foram necessárias 27 horas para realizar a conversão dos arquivos coletados para o modelo de dados do SAGA. Durante o período de análise, e considerando os novos filtros impostos, o conjunto de dados apresentou um crescimento de apenas 6.11%. O que corrobora com a latente necessidade da coleta recursiva dos dados.

5.7 MAPEAMENTO DE EPÍTOPOS

A identificação de estruturas imunogênicas nos genomas virais é de real importância. Este processo gera informações que podem ajudar a selecionar alvos para estudos. Outrossim, considerando um banco de dados integrado de multi-organismos da mesma família, foi possível ainda identificar os epítopos comumente frequentes entre os organismos. Assim, aplicadas análises de correlação entre epítopos mapeados em mais de um organismo, observamos que o ZKV, DENV e YFV possuem 27 epítopos frequentes em comum. Enquanto o CKV, apesar de possuir mais epítopos frequentes em seu conjunto de dados, não apresenta sobreposição com epítopos frequentes nos outros organismos, como pode ser visto na Figura 28.

Além disso, é possível ainda identificar que o epítopo “GKAKGSRAIWYMWLG” é o mais frequente entre o DEV, YFV e ZKV, com 24,98%, 22,54% e 28,17% de frequência nas sequências destes organismos, respectivamente. Entretanto, o mapeamento dos epítopos presentes nas sequências do CKV estão mais fragmentadas, havendo uma diferença de apenas 0,52% entre o primeiro e segundo mais frequente. O que pode ser devido ao fato de o CKV pertencer a um complexo imunogênico diferente dos outros organismos.

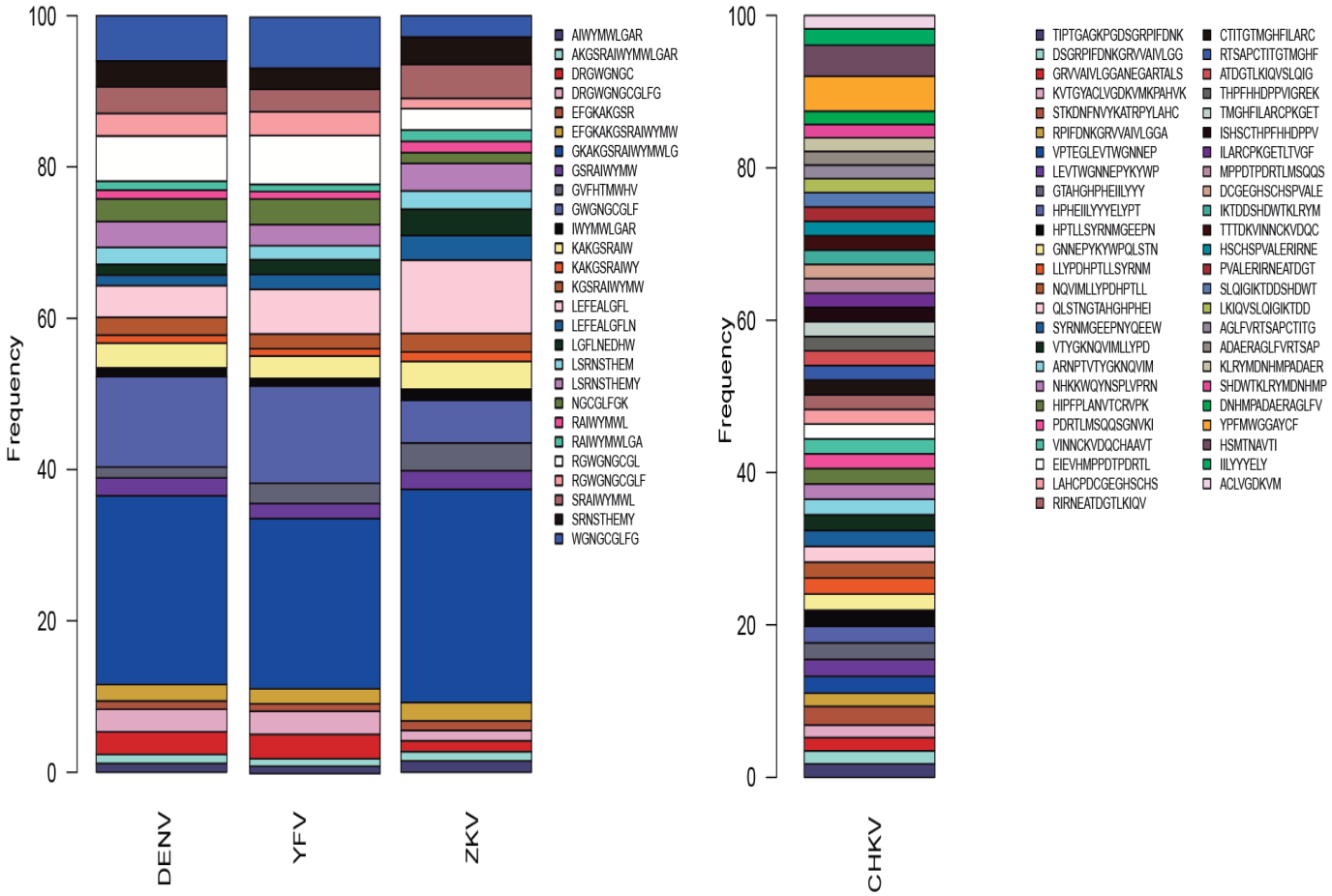


Figura 28 - Gráfico de barras empilhadas representando a frequência em % apresentada pelos epítopos mapeados nos respectivos organismos e em comum. Exceto CHKV que não possui epítopo em comum, então foram compilados os mais frequentes

Além das frequências gerais dos epítopos nos organismos, foram ainda extraídas informações sobre a frequência de cada epítopo por classificação genotípica da cepa. Assim, é possível identificar os epítopos mais frequentes por subtipo de cada organismo. Estes resultados estão representados na Figura 29, onde observamos uma proporção relativamente alta de epítopos com frequência menor que 3%.

Entretanto, podemos observar que existem epítopos como o "AMTDTTPFGQQRVFK" que além de serem frequentes em todos os subtipos do agente etiológico, apresentam valores de frequência relativa acima de 5% para os subtipos. Sendo estes valores 5,04%, 5,02%, 6,10% e 4,89% respectivamente para os subtipos 1, 2, 3 e 4.

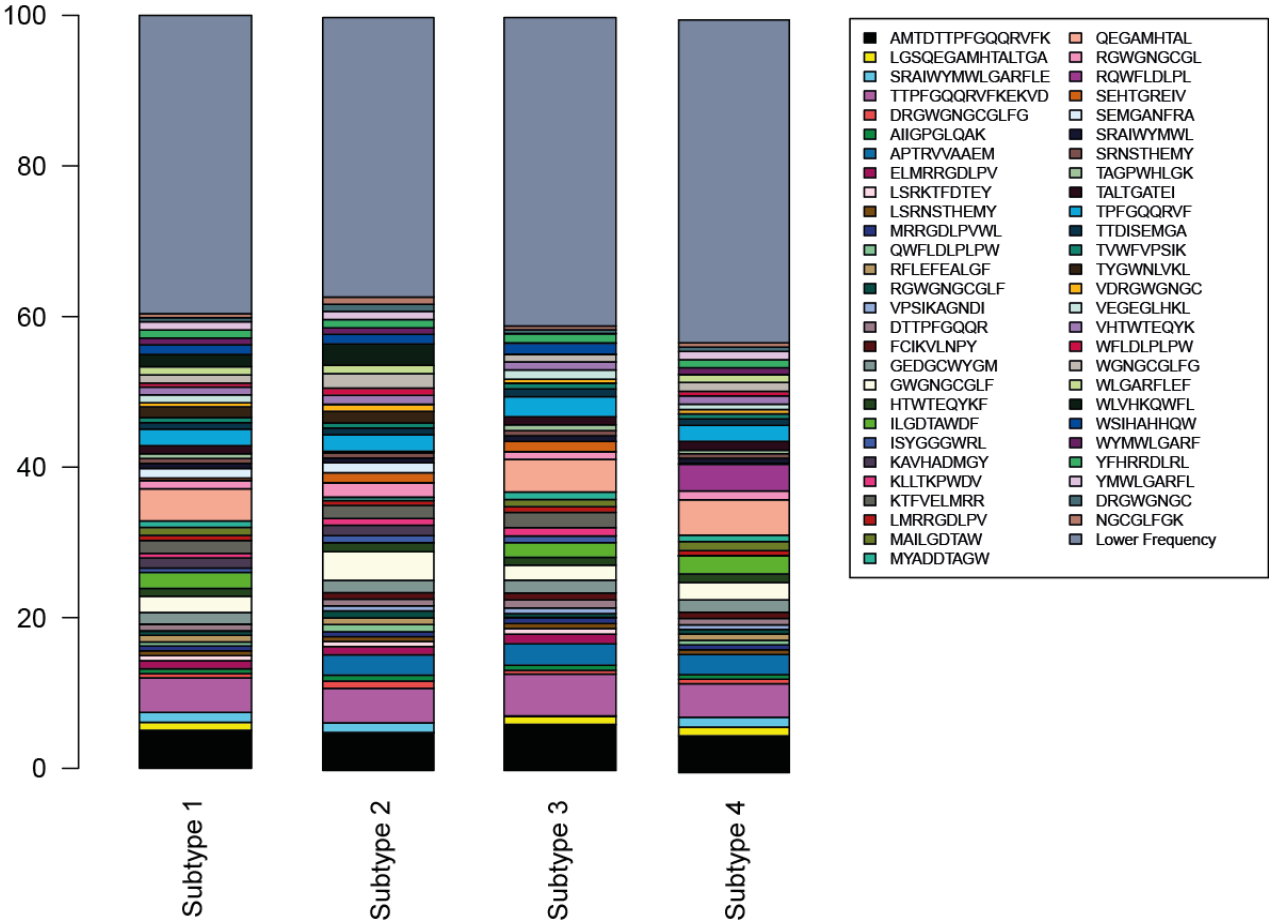


Figura 29 - Gráfico de barras empilhadas representando os epítopos de maior frequência em percentagem no conjunto de dados relativos ao DENV. Agrupando epítopos com menos de 3% de frequência no grupo *"Lower Frequency"*

O ZKV também apresentou epítopos frequentes comuns entre os subtipos. Entretanto, como pode ser visto na Figura 30, os epítopos com frequência absoluta menor que 3% representam a maioria dos pareamentos deste subconjunto. Outrossim, é possível observar que o epítopo "DQRGSGQVVTYALNT" apresentou a maior frequência relativa, tendo chegado a 18,57% de frequência no subtipo *Micronesian* e 10,77% no subtipo *Senegalese*.

O mapeamento de epítopos nas sequências do YFV, também apresentaram ambiguidades quanto à classificação genotípica das sequências. Diversos epítopos apresentaram altas frequências e presença em todos os subtipos, como pode ser visto na Figura 31. No entanto, que a distribuição das frequências dos epítopos entre os genótipos do vírus é mais fragmentada e uniforme.

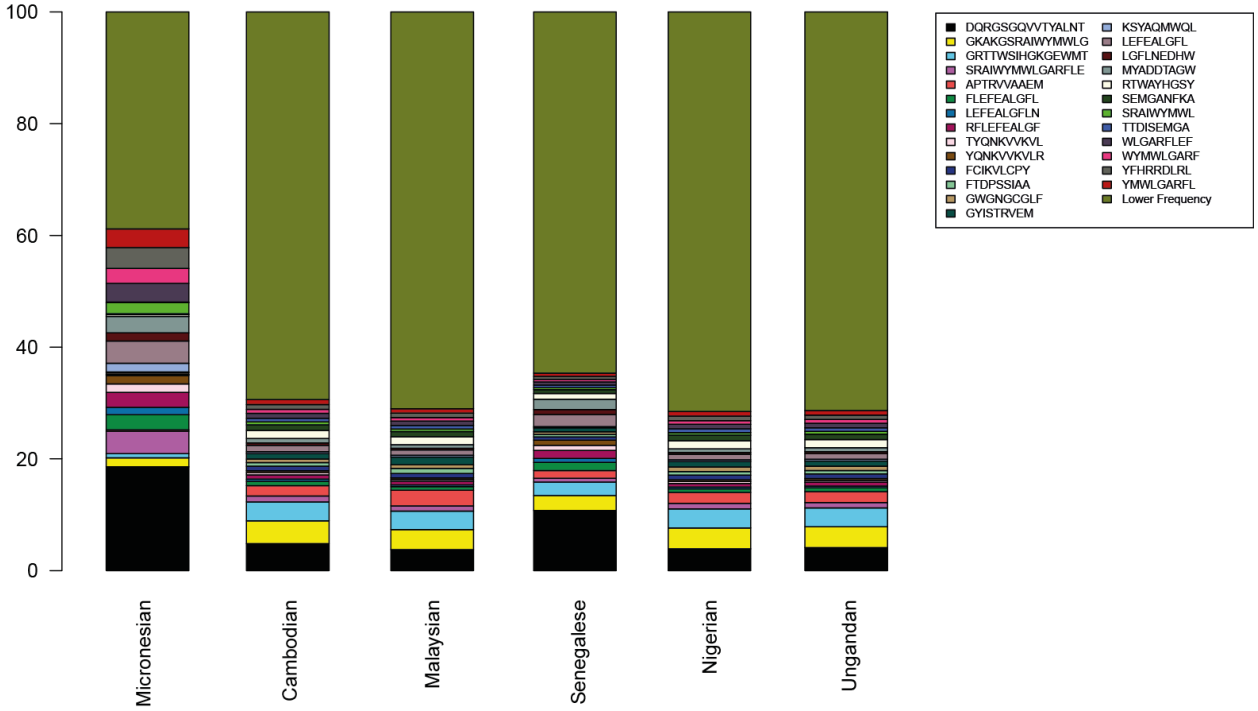


Figura 30 - Gráfico de barras empilhadas representando os epítopos de maior frequência em percentagem no conjunto de dados relativos ao ZKV. Agrupando epítopos com menos de 3% de frequência no grupo *"Lower Frequency"*

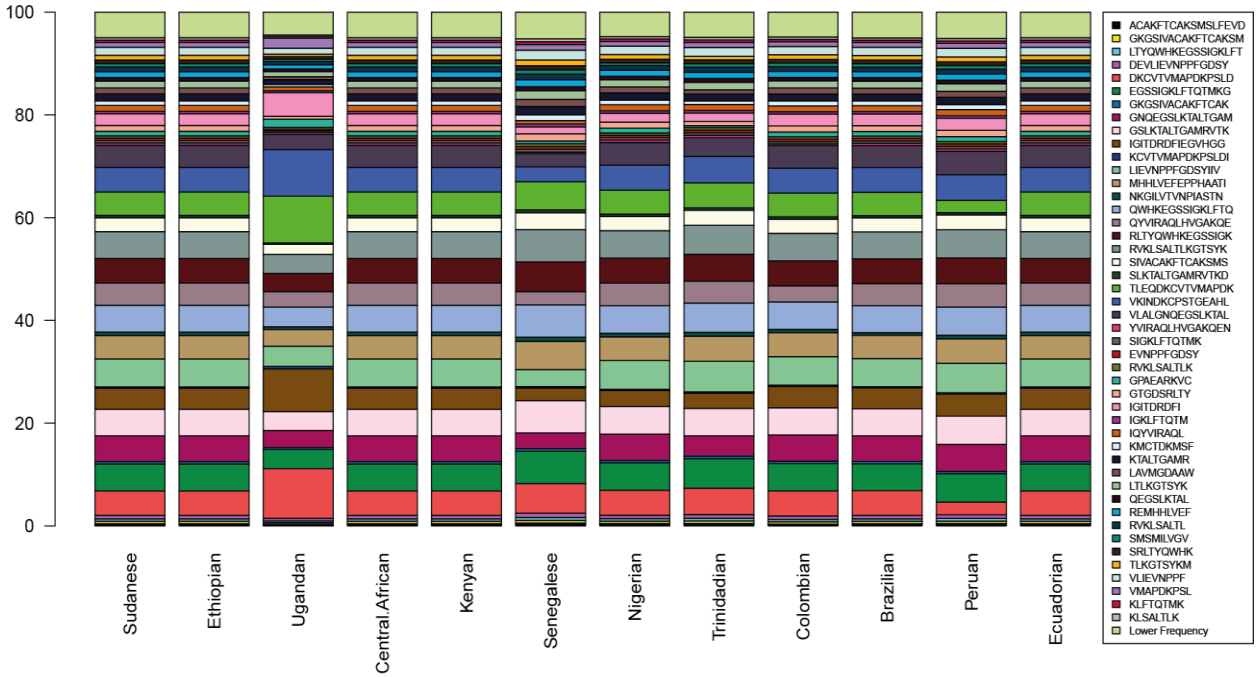


Figura 31 - Gráfico de barras empilhadas representando os epítopos de maior frequência em percentagem no conjunto de dados relativos ao YFV. Agrupando epítopos com menos de 3% de frequência no grupo *"Lower Frequency"*

O mapeamento de epítopos realizado nas sequências de CKV, não geraram ambiguidades entre os subtipos identificados. O que significa que não há no conjunto de dados atual, epítopos com mais de 3% de frequência no conjunto de dados que estejam

mapeados em mais de um subtipo do CKV. Desta forma, observa-se que não houve mapeamentos de mesmos epítopos em sequências classificadas em grupos genotípicos diferentes. Assim, a sobreposição entre a classificação genotípica e o mapeamento de epítopos não possui intersecções exploráveis.

6 DISCUSSÃO

Para melhor compreender e combater os agentes etiológicos é necessário obter informações sob diversos aspectos diferentes (MARCONDES, 2016; DONATELI, 2019). Para tal, é necessário conhecer, por exemplo, o genoma completo do organismo e suas estruturas. Estes genes dispostos no conjunto genômico do organismo podem ser divididos em estruturais e não estruturais. Entretanto, independente de sua classificação, possuem em sua anotação, uma limitação de escopo, ou definição de limites de tradução.

Os limites de regiões gênicas no processo de anotação destes organismos representam uma informação de relevante importância para o processo de mapeamento do conjunto de dados. Uma vez que são utilizadas para montar um mapa de características estruturais e não estruturais deste organismo. Esta informação posicional é utilizada principalmente para definir alvos de estudos e/ou tratamentos. As diversas sequências nucleotídicas inseridas nos conjuntos de dados primários nem sempre são indexadas de acordo com o genoma completo anotado (MORGULIS, 2008). Devidamente mapeadas, as sequências do conjunto de dados podem ser classificadas quanto às estruturas que codificam, e quanto a ou as regiões que representam em relação ao genoma completo dos organismos.

Foi observado após o mapeamento genômico das sequências entre os organismos em questão, que a maior frequência de pareamento ocorreu na região do envelope viral. Isso se dá pela larga ampliação desta região para estudos de filogenia (ENFISSI, 2016), presença de regiões imunodominantes, interação agente/hospedeiro (ENFISSI, 2016) e, no caso do YFV, a maior frequência de ampliação dessa região pode ser explicada por que esta região é o principal alvo para desenvolvimento de vacina (POST, 1992).

Além de informações posicionais e frequência de regiões representadas no conjunto de dados e objetivando melhor compreensão da patogênese e dos processos epidemiológicos dos agentes etiológicos em questão, se faz necessário compreender sua heterogeneidade genotípica (CHRISTOFFERSON, 2016). Levando ainda em consideração sua disposição e frequência no conjunto de dados atual. O procedimento de subtipagem gera informações consideradas de real valia para o estudo das etiologias associadas a estes patógenos (CHAN, 2014). Além disso, o processo de classificação e melhor caracterização da variabilidade genotípica dos vírus em questão, auxilia no processo de desenvolvimento de novos métodos diagnósticos, principalmente em casos de infecção cruzada (PABBARAJU, 2016).

Com o conjunto de dados atual é possível monitorar o crescimento quantitativo dos subtipos virais, baseando-se na periocidade de submissão de novas cepas e posterior

classificação das mesmas. Este processo pode ajudar inclusive na vigilância epidemiológica dos subtipos sendo possível realizar diversos tipos de extração e agrupamentos no conjunto de dados, utilizando características relevantes como região geográfica de submissão, data de submissão, subtipagem, mapeamento, classificação, quantitativos, métodos estatísticos dentre outros.

Por exemplo, estratificando as sequências do ZKV por país (Brasil); e por ano de submissão, agregando por subtipo, temos a realidade do período de 2016 até 2018, onde temos dois subtipos circulantes sendo submetidos "*Cambodia*" com 96,88%, 97,73% e 92% das sequências respectivamente em 2016, 2017 e 2018, assim como "*Micronesia*" com 3,13%, 3,41% e 8% respectivamente nos mesmos anos. Entretanto, em 2019 temos uma mudança nas características das sequências submetidas, com "*Cambodia*" representando apenas 5,41%, "*Micronesia*" com 89,19% e o surgimento de um novo subtipo, de uma outra linhagem no subconjunto: "*Uganda*" com 5,41%. A predominância de subtipos da linhagem asiática do agente etiológico causador da zica é compreensível, uma vez que desde 2015 há evidências de que esta linhagem seja predominante (SHI, 2016). Entretanto, observamos a inserção de uma nova linhagem no contexto de submissão nacional. Isto pode ser explicado pelo fato de que as linhagens de ZKV circulantes no país compartilham de 87 a 90% de características genômicas com a linhagem africana (CUGOLA, 2016; FAYE, 2014), mais especificamente com o subtipo "*Uganda*". Apesar de que, as características clínicas associadas ao acometimento são diferentes, uma vez que a linhagem africana não está associada diretamente às malformações congênitas (CUGOLA, 2016). Outrossim, é importante observar que a diversidade deste agente etiológico e suas cepas circulantes em território nacional tem aumentado (SHI, 2016), o que reforça a necessidade de um conjunto de dados que se renova e se atualiza periodicamente.

Além do estudo classificatório do conjunto de dados, objetivando identificar melhores alvos para estudos de desenvolvimento de fármacos, tratamentos e vacinas, é preciso conhecer informações sobre epítopos no conjunto de dados (PRATHEEK, 2015; SWANSTROM, 2016). Estas informações são geradas a partir do mapeamento entre sequências de epítopos disponíveis em bancos de dados públicos desta natureza e o conjunto de dados local.

Considerando um conjunto de dados multi-organismo e um conjunto de epítopos genéricos, é possível ainda com a plataforma em seu estado atual, identificar epítopos que são frequentes e comuns entre os organismos. Este tipo de informação é importante para o desenvolvimento de tratamentos e vacinas, além de auxiliar no desenvolvimento de métodos

diagnósticos de coinfecção (PABBARAJU, 2016). Foram identificados com a análise dos dados gerados pela plataforma que existem diversos epítopos com frequência relevante entre o ZKV, YFV e DENV. Destes se destaca o epítopo “KAKGSRAIWYMWLG”, identificado inicialmente no *YFV*. Este epítopo possui reatividade imunogênica atestada em mais de 20 trabalhos publicados segundo as referências cruzadas do próprio *IEDB.* Além deste, outro epítopos se mostraram relevantes alvos para estudos sobre os organismos em questão. Além de epítopos comuns, as frequências gerais de epítopos podem evidenciar epítopos exclusivos de um determinado organismo, linhagem ou subtipo, servindo como biomarcador ou alvo para estudos sobre coinfecção (KAM, 2014). Os resultados da análise mostram que o CKV não possui epítopos em comum com os outros vírus avaliados. Ainda é possível observar com as análises que sua distribuição é mais fragmentada, não havendo um epítopo claramente mais frequente. Entretanto, observando aquele que obteve mais hits no mapeamento o “YPFMWGGAYCF”, que foi identificado inicialmente no *Semliki Forest virus* o qual, pertence ao mesmo grupo imunogênico que o CKV (CLETON, 2012) e que foi atestado positivamente imunorreativo (GIBBONS, 2004).

Ademais, é preciso considerar ainda que nem todos os possíveis epítopos foram gerados, identificados ou estudados nos diversos organismos virais existentes. Constantemente estão sendo desenvolvidos, identificados e/ou preditos epítopos. Entretanto enquanto não depositados, processados, curados e disponibilizados, a plataforma não pode testar este contra o conjunto de dados. Isto causa um sério atraso no processo de validação de epítopos comuns e exclusivos, burocratizando e dificultando o processo de combate aos agentes etiológicos e doenças associadas. Buscando mitigar este fato, foi criada uma ferramenta de mapeamento por submissão, que recebe uma sequência, realiza o mapeamento dessa sequência e retorna a sua frequência por vírus. O resultado final é a frequência geral do epítopo no conjunto de dados, no subconjunto de vírus e/ou no subconjunto de seus subtipos. Essa ferramenta poderá facilitar e acelerar o processo de identificação *in silico* de alvos de diagnósticos e vacinas.

Finalmente, apresentamos o SAGA, como uma base de dados e modelagem sólida, sem perder flexibilidade para inserção de novos organismos. Este conjunto de dados é recursivamente atualizado por uma série de fontes e tipos de dados diferentes, gerando uma base secundária com essas informações, base esta inédita para esses vírus. Não somente, o conjunto de dados é acompanhado, mantido, organizado e atualizado por uma plataforma automatizada de bioinformática. Esta, desenvolvida com modernas técnicas de programação que não são afetadas pelo tamanho do conjunto de dados. Desta forma preparando a

plataforma para trabalhar com o constante crescimento quantitativo e qualitativo do conjunto de dados. Além disso, buscando reduzir o tempo de desenvolvimento de trabalhos da mesma natureza e facilitar trabalhos que utilizem dados ou análises similares, foi criada a *API* utilizando o padrão *REST* dando acesso a todas as funcionalidades públicas e dados do SAGA.

Desta forma, observamos que nesta tese foi possível desenvolver uma plataforma unificada para análise genômica e armazenamento de informações sobre *Zika virus*, *Dengue virus*, *Chikungunya virus* e *Yellow Fever Virus*. E que esta pode ser uma importante ferramenta de centralização e geração de informações sobre os agentes etiológicos em questão e as doenças relacionadas. Desta forma facilitando o combate, acelerando a vigilância genotípica e epidemiológica, dando suporte a tomadas de decisão mais rápidas e assertivas em relação à saúde pública.

7 CONCLUSÃO

Doenças causadas por arbovírus são um grave problema de saúde pública. Vírus como ZKV, DENV, YFV e CKV são transmitidos pelo vetor *Aedes* possuem o agravante da prevalência do vetor dado o difícil controle populacional do mesmo. Estes organismos estão associados a doenças debilitantes que representam um problema econômico, uma vez que tiram o cidadão do seu local de trabalho. Outrossim, estão associados à doenças crônicas que oneram o Estado na assistências aos acometidos.

Para abordar este problema, diversos esforços vêm sendo feitos para melhor compreender os agentes etiológicos em questão, seu perfil genotípico, patogênese, imunogenicidade, interação com hospedeiro, entre outros. Entretanto este tipo de análise é dependente de dados consolidados e informações não disponíveis em bancos de dados biológicos primários. Tais como Classificação, mapeamento e frequência de regiões subgenômicas e busca e identificação de epítopos.

Desta forma, este trabalho objetivou o desenvolvimento de uma plataforma unificada de coleta e análise de sequências nucleotídicas destes agentes etiológicos disponíveis nos bancos de dados globais. Além da obtenção e indexação de epítopos do IEDB no conjunto de dados. Agregando e qualificando informações sobre os organismos.

Uma vez concluído, o SAGA representa uma nova perspectiva no perfil dos organismos. Capaz de realizar coleta e análise recursiva sem interação humana, a plataforma se coloca como um marco na independência e viabilidade dos bancos de dados biológicos, contribuindo cada vez mais para estudos sobre os vírus em questão e por tanto contribuindo para o desenvolvimento de diagnósticos e vacinas para as doenças associadas.

Além disso, a plataforma oferece acesso direto às técnicas de classificação, mapeamento, busca de epítopos e geração de dados. Servindo como um centro de dados biológicos e uma importante plataforma de análises de bioinformática, oferecendo sob demanda funcionalidades essenciais ao estudo de diversos agentes etiológicos, evitando retrabalho e acelerando o desenvolvimento de novos estudos sobre estes e outros agentes etiológicos.

REFERÊNCIAS

AHOLA, T. *et*. al. Therapeutics and Vaccines Against Chikungunya Virus. **Vector-Borne and Zoonotic Diseases**, v. 15, n.4, 250–257, 2015. Disponível em: https://doi.org/10.1089/vbz.2014.1681

ALI OU ALLA, S.; COMBE, B. Arthritis after infection with Chikungunya virus. **Best Practice & Research Clinical Rheumatology**, v. 25, n. 3, p. 337–346, 2011. Disponível em: https://doi.org/10.1016/j.berh.2011.03.005

ÁLVAREZ-ARGÜELLES, E. *et. al*. Diagnosis and Molecular Characterization of Chikungunya Virus Infections. **Current Topics in Neglected Tropical Diseases**, 2019. IntechOpen. Disponível em: https://doi.org/10.5772/intechopen.86957

ARCURI, A. RESTful API Automated Test Case Generation with EvoMaster. *ACM* **Transactions on Software Engineering and Methodology**, v. *28, n.* 1, p. 1–37, 2019. Disponível em: https://doi.org/10.1145/3293455

BAO, Y. *et. al*. The Influenza Virus Resource at the National Center for Biotechnology Information. **Journal of Virology**, v. 82, n. 2, 596–601, 2008. Disponível em: https://doi.org/10.1128/JVI.02005-07

BARRETT, A. D. T.; HIGGS, S. Yellow Fever: A Disease that Has Yet to be Conquered. **Annual Review of Entomology**, v. *52, n.* 1, p. 209–229, 2007. Disponível em: https://doi.org/10.1146/annurev.ento.52.110405.091454

BESNARD, M. *et. al*. (2014). Evidence of perinatal transmission of Zika virus, French Polynesia, December 2013 and February 2014. **Eurosurveillance**, v. 19, n. 13, 20751, 2014. Disponível em: https://doi.org/10.2807/1560-7917.ES2014.19.13.20751

BHAMARAPRAVATI, N.; SUTEE, Y. Live attenuated tetravalent dengue vaccine. **Vaccine**, v. 18, p. 44–47, 2000. Disponível em: https://doi.org/10.1016/S0264-410X(00)00040-2

BRASIL. MINISTÉRIO DA SAÚDE. **Centro de Operações em Emergências em Saúde Pública. Monitoramento dos casos e óbitos de febre amarela no Brasil** - Informe n. 08/2017. Disponível em: http://portalsaude.saude.gov.br/ images/pdf/2017/fevereiro/06/COES-FEBRE-AMARELA- INFORME-8-Atualização.pdf.

BRASIL. MINISTÉRIO DA SAÚDE. SECRETARIA DE VIGILÂNCIA EM SAÚDE. Zika Vírus: perfil epidemiológico em mulheres. **Boletim Epidemiológico,** v. 47, n. 37, p. 7, 2016. Disponível em: http://portalarquivos2.saude.gov.br/images/pdf/2016/novembro/15/2016_031-Mulheres_publicacao.pdf

CAMPOS, G. S.; BANDEIRA, A. C.; SARDI, S. I. Zika Virus Outbreak, Bahia, Brazil. **Emerging Infectious Disease**, v. 21, n. 10, 2015. Disponível em: https://doi.org/10.3201/eid2110.150847

CARDOSO, C. W. *et. al*. Unrecognized Emergence of Chikungunya Virus during a Zika Virus Outbreak in Salvador, Brazil. **PLOS Neglected Tropical Diseases**, v. 11, n. 1, e0005334, 2017. Disponível em: https://doi.org/10.1371/journal.pntd.0005334

CARTER H.R. Yellow Fever: An Epidemiological and Historical Study of Its Place of Origin. Baltimore, MD:Williams &Wilkins. **Nature**, v. 130, p. 646–647, 1931. Disponível em: https://doi.org/10.1038/130646a0

CAVALCANTI, L. P. de G. *et. al.* Surveillance of deaths caused by arboviruses in Brazil: from dengue to chikungunya. **Memórias do Instituto Oswaldo Cruz**, v. 112, n. 8, p. 583–585, 2017. https://doi.org/10.1590/0074-02760160537

CHA, G. W. *et. al*. Travel-Associated Chikungunya Cases in South Korea during 2009–2010. **Osong Public Health and Research Perspectives**, v. 4, n. 3, p. 170–175, 2013. Disponível em: https://doi.org/10.1016/j.phrp.2013.04.008

CHAKRABORTY, A.; BANDYOPADHYAY, S. FOGSAA: Fast Optimal Global Sequence Alignment Algorithm. **Scientific Reports**, v. 3, n.1, 1746, 2013. Disponível em: https://doi.org/10.1038/srep01746

CHAN, P. A. *et. al*. Phylogenetic and geospatial evaluation of HIV-1 subtype diversity at the largest HIV center in Rhode Island. **Infection, Genetics and Evolution**, v. 28, p. 358–366, 2014. Disponível em: https://doi.org/10.1016/j.meegid.2014.03.027

CHAN, P. A. *et. al*. Phylogenetic and geospatial evaluation of HIV-1 subtype diversity at the largest HIV center in Rhode Island**. Infection, Genetics and Evolution**, v. 28, p. 358–366, 2014. Disponível em: https://doi.org/10.1016/j.meegid.2014.03.027

CHANG, G. J. *et. al*. Nucleotide sequence variation of the envelope protein gene identifies two distinct genotypes of yellow fever virus. **Journal of Virology**, v. 69, n. 9, p. 5773–5780, 1995. Disponível em: https://doi.org/10.1128/jvi.69.9.5773-5780.1995

CHRISTOFFERSON, R. C. Zika Virus Emergence and Expansion: Lessons Learned from Dengue and Chikungunya May Not Provide All the Answers. **The American Journal of Tropical Medicine and Hygiene**, v. 95, n. 1, p. 15–18, 2016. Disponível em: https://doi.org/10.4269/ajtmh.15-0866

CLETON, N. *et. al.* Come fly with me: Review of clinically important arboviruses for global travelers. **Journal of Clinical Virology**, v. 55, n. 3, p. 191–203, 2012. Disponível em: https://doi.org/10.1016/j.jcv.2012.07.004

COFFEY, L.; FAILLOUX, A.B.; WEAVER, S. Chikungunya Virus–Vector Interactions. **Viruses**, v. 6, n. 11, p. 4628–4663, 2014. Disponível em: https://doi.org/10.3390/v6114628

COMBE, M.; SANJUÁN, R. (2014). Variation in RNA Virus Mutation Rates across Host Cells. ***PLoS Pathogens***, *10*(1), e1003855. Disponível em: https://doi.org/10.1371/journal.ppat.1003855

CUGOLA, F. R *et. al* (2016). The Brazilian Zika virus strain causes birth defects in experimental models. ***Nature***, *534*(7606), 267–271. Disponível em: https://doi.org/10.1038/nature18296

CUNHA, R. V. da; TRINTA, K. S. (2017). Chikungunya virus: clinical aspects and treatment - A Review. ***Memórias Do Instituto Oswaldo Cruz***, *112*(8), 523–531. Disponível em: https://doi.org/10.1590/0074-02760170044

WORLD HEALTH ORGANIZATION (2009) Dengue: guidelines for diagnosis, treatment, prevention and control: New Edition. Geneva: World Health Organization, 2009.

DESHMUKH, K. B.; KHARAT, M. U. (2015). *Review on Retrieving Biological Sequence Alignment using Smith-Waterman Algorithm*. ***International Journal of Innovative Research in Computer Science & Technology*** *1*, 24–26. Disponível em: https://www.ijircst.org/DOC/6_review_on_retrieving_biological_sequence_alignment_using_smith-waterman_algorithm.pdf

DIAMOND, M. S.; PIERSON, T. C. (2015). Molecular Insight into Dengue Virus Pathogenesis and Its Implications for Disease Control. ***Cell***, *162*(3), 488–492. Disponível em: https://doi.org/10.1016/j.cell.2015.07.005

DOLMA, K. (2017). *Zika Virus (ZIKV) Infection*: A Review**, *Radiology of Infectious Diseases* 4**(2), 2–7. Disponível em: https://doi.org/10.1016/j.jrid.2017.01.002

DONATELI, C. P. *et al.* Endemic disease control agents' perception on the fight against aedes aegypti and the prevention of arbovirus infections in Brazil. **PLoS Neglected Tropical Diseases**, v. 13, n. 10, p. 1–15, 2019. Disponível em: https://doi.org/10.1371/journal.pntd.0007741.

DUFFY, M. R. *et. al* (2009). Zika Virus Outbreak on Yap Island, Federated States of Micronesia. ***New England Journal of Medicine***, *360*(24), 2536–2543. Disponível em:https://doi.org/10.1056/NEJMoa0805715

EICKMANN, S. H. *et. al* (2016). Síndrome da infecção congênita pelo vírus Zika. ***Cadernos de Saúde Pública***, *32*(7). Disponível em: https://doi.org/10.1590/0102-311X00047716

ENFISSI, A *et. al*. (2016). Zika virus genome from the Americas. ***The Lancet***, *387*(10015), 227–228. Disponível em: https://doi.org/10.1016/S0140-6736(16)00003-9

FARIA, N. R. *et. al* (2016). Epidemiology of Chikungunya Virus in Bahia, Brazil, 2014-2015. ***PLoS Currents***. Disponível em: https://doi.org/10.1371/currents.outbreaks.c97507e3e48efb946401755d468c28b2

FARIA, N. R., Lourenço, J., Marques de Cerqueira, E., Maia de Lima, M., & Carlos Junior Alcantara, L. (2016). Epidemiology of Chikungunya Virus in Bahia, Brazil, 2014-2015. ***PLoS Currents***. Disponível em: https://doi.org/10.1371/currents.outbreaks.c97507e3e48efb946401755d468c28b2

FAYE, O *et. al*. (2014). Molecular Evolution of Zika Virus during Its Emergence in the 20th Century. ***PLoS Neglected Tropical Diseases***, *8*(1), e2636. Disponível em: https://doi.org/10.1371/journal.pntd.0002636

FAYE, O. *et. al*. (2014). Molecular Evolution of Zika Virus during Its Emergence in the 20th Century. ***PLoS Neglected Tropical Diseases***, *8*(1), e2636. Disponível em: https://doi.org/10.1371/journal.pntd.0002636

FERNÁNDEZ-SUÁREZ, X. M.; RIGDEN, D. J.; GALPERIN, M. Y. (2014). The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. ***Nucleic Acids Research***, *42*(D1), D1–D6. Disponível em: https://doi.org/10.1093/nar/gkt1282

FILIPE, A. R.; MARTINS, C. M. V.; ROCHA, H. (1973). Laboratory infection with Zika virus after vaccination against yellow fever. ***Archiv Für Die Gesamte Virusforschung***, *43*(4), 315–319. Disponível em: https://doi.org/10.1007/BF01556147

FOY, B. D. *et. al*. (2011). Probable Non–Vector-borne Transmission of Zika Virus, Colorado, USA. ***Emerging Infectious Diseases***, *17*(5), 880–882. Disponível em: https://doi.org/10.3201/eid1705.101939

FUKUTANI, K. F. *et. al* (2018). Meta-Analysis of Aedes aegypti Expression Datasets: Comparing Virus Infection and Blood-Fed Transcriptomes to Identify Markers of Virus Presence. ***Frontiers in Bioengineering and Biotechnology***, *5*. Disponível em: https://doi.org/10.3389/fbioe.2017.00084

GAFFIGAN T.V.W.R. *et. al* (2016) Smithsonian Institution; Silver Spring (MD): **Walter Reed Army Institute of Research**; 2016. Disponível em: http://www.mosquitocatalog.org/taxon_table.aspx. Acessado em: 03 de março de 2016.

GALÁN-HUERTA, K. A. *et. al* (2015). Chikungunya virus: A general overview. **Medicina Universitaria**, *17*(68), 175–183. Disponível em: https://doi.org/10.1016/j.rmu.2015.06.001

GALÁN-HUERTA, K. A. *et. al*. (2016). The Zika virus disease: An overview. **Medicina Universitaria**, *18*(71), 115–124. Disponível em: https://doi.org/10.1016/j.rmu.2016.05.003

GATHERER, D.; KOHL, A. (2016). Zika virus: a previously slow pandemic spreads rapidly through the Americas. ***Journal of General Virology***, *97*(2), 269–273. Disponível em: https://doi.org/10.1099/jgv.0.000381

GATHERER, D.; KOHL, A. (2016). Zika virus: a previously slow pandemic spreads rapidly through the Americas. ***Journal of General Virology***, *97*(2), 269–273. Disponível em: https://doi.org/10.1099/jgv.0.000381

GERSHONI, J. M. *et. al*. (2007). Epitope Mapping. ***BioDrugs***, *21*(3), 145–156. Disponível em: https://doi.org/10.2165/00063030-200721030-00002

GIBBONS, D. L *et. al*. (2004). Multistep Regulation of Membrane Insertion of the Fusion Peptide of Semliki Forest Virus. ***Journal of Virology***, *78*(7), 3312–3318. Disponível em: https://doi.org/10.1128/JVI.78.7.3312-3318.200

GIMENEZ, G.; BENEDÉ, S.; LIN, J. (2016). IgE Epitope Mapping Using Peptide Microarray Immunoassay ***Peptide Microarrays*** (pp. 251–261). Disponível em: https://doi.org/10.1007/978-1-4939-3037-1_19

GOUPIL, B. A.; MORES, C. N. (2016). A Review of Chikungunya Virus-induced Arthralgia: Clinical Manifestations, Therapeutics, and Pathogenesis. ***The Open Rheumatology Journal***, *10*(1), 129–140. Disponível em: https://doi.org/10.2174/1874312901610010129

GREEN, S.; ROTHMAN, A. (2006). Immunopathological mechanisms in dengue and dengue hemorrhagic fever. ***Current Opinion in Infectious Diseases***, *19*(5), 429–436. Disponível em: https://doi.org/10.1097/01.qco.0000244047.31135.fa

GUABIRABA, R.; RYFFEL, B. (2014). Dengue virus infection: current concepts in immune mechanisms and lessons from murine models. ***Immunology,*** *141*(2), 143–156. 144p. Disponível em: https://doi.org/10.1111/imm.12188

GUANCHE GARCELL, H. *et. al* (2020). Clinical relevance of Zika symptoms in the context of a Zika Dengue epidemic. ***Journal of Infection and Public Health***, *13*(2), 173–176. Disponível em: https://doi.org/10.1016/j.jiph.2019.07.006

GUZMAN, M. G. *et. al* (2010). Dengue: a continuing global threat. ***Nature Reviews Microbiology***, *8*(S12), S7–S16. Disponível em: https://doi.org/10.1038/nrmicro2460

HADDOW, A. D. *et. al* (2012). Genetic Characterization of Zika Virus Strains: Geographic Expansion of the Asian Lineage. ***PLoS Neglected Tropical Diseases***, *6*(2), e1477. Disponível em: https://doi.org/10.1371/journal.pntd.0001477

HALSTEAD, S. B. (1988). Pathogenesis of dengue: Challenges to molecular biology. ***Science***, *239*(4839), 476–481. https://doi.org/10.1126/science.3277268

HAMEL, R. *et. al*. (2015). Biology of Zika Virus Infection in Human Skin Cells. ***Journal of Virology***, *89*(17), 8880–8896. Disponível em: https://doi.org/10.1128/JVI.00354-15

HASAN, S. *et. al* (2016). Dengue virus: A global human threat: Review of literature. ***Journal of International Society of Preventive and Community Dentistry***, *6*(1), 1. Disponível em: https://doi.org/10.4103/2231-0762.175416

HOLMES, E.; TWIDDY, S. (2003). The origin, emergence and evolutionary genetics of dengue virus. ***Infection, Genetics and Evolution***, *3*(1), 19–28. Disponível em: https://doi.org/10.1016/S1567-1348(03)00004-2

HONÓRIO, N. A. *et al.* Chikungunya: uma arbovirose em estabele cimento e expansão no Brasil. **Cadernos de Saúde Pública,** v. 31, n. 5, p. 906–908, 2015. Disponível em: https://doi.org/10.1590/0102-311XPE020515

IOOS, S. *et. al*. (2014). Current Zika virus epidemiology and recent epidemics. ***Médecine et Maladies Infectieuses***, *44*(7), 302–307. Disponível em: https://doi.org/10.1016/j.medmal.2014.04.008

KAM, *et. al* (2014). Unique Epitopes Recognized by Antibodies Induced in Chikungunya Virus-Infected Non-Human Primates: Implications for the Study of Immunopathology and Vaccine Development. ***PLoS ONE***, *9*(4), e95647. Disponível em: https://doi.org/10.1371/journal.pone.0095647

KAM, Y.W. *et. al* (2009). Immuno-biology of Chikungunya and implications for disease intervention. ***Microbes and Infection***, *11*(14–15), 1186–1196. Disponível em: https://doi.org/10.1016/j.micinf.2009.09.003

KAMARULZAMAN, A. *et. al*. (2019, November 11). Reducing Integrated Operation Project Deployment Complexity through Application Programming Interface API Enabled Software. ***Abu Dhabi International Petroleum Exhibition & Conference***. Disponível em: https://doi.org/10.2118/197165-MS

KASPRZYKOWSKI, J. I. (2016). ***Sistema para an análise de sequências nucleotídicas do HIV disponíveis no GenBank***. Dissertação (Dissertação em Computação Aplicada) - Universidade Estadual de Feira de Santana.

KASPRZYKOWSKI, J. I. *et. al*. (2017). HIV-1 Nucleotide Sequence Comprehensive Analysis: A Computational Approach. ***Current Bioinformatics***, *12*(4). Disponível em: https://doi.org/10.2174/1574893611666161027142611

KAST, W. M. (2002). Ins and outs of clinical trials with peptide-based vaccines. ***Frontiers in Bioscience***, *7*(5), e204-213. Disponível em: https://doi.org/10.2741/a916

KAUFMANN, B.; ROSSMANN, M. G. (2011). Molecular mechanisms involved in the early steps of flavivirus cell entry. ***Microbes and Infection***, *13*(1), 1–9. Disponível em: https://doi.org/10.1016/j.micinf.2010.09.005

KRAEMER, M. U. G. *et. al* (2015). *The global compendium of Aedes aegypti and Ae. Albopictus occurrence*. ***Scientific Data***, *2*(1), 150035. Disponível em: https://doi.org/10.1038/sdata.2015.35

KUIKEN, C. *et. al* (2005). The Los Alamos hepatitis C sequence database. ***Bioinformatics***, *21*(3), 379–384. Disponível em: https://doi.org/10.1093/bioinformatics/bth485

KUIKEN, C. *et. al*. (2005). The Los Alamos hepatitis C sequence database. ***Bioinformatics***, *21*(3), 379–384. Disponível em: https://doi.org/10.1093/bioinformatics/bth485

KUIKEN, C. *et. al*. (2012). The LANL hemorrhagic fever virus database, a new platform for analyzing biothreat viruses. ***Nucleic Acids Research***, *40*(D1), D587–D592. Disponível em: https://doi.org/10.1093/nar/gkr898

KUMARIA, R. (2010). Correlation of disease spectrum among four Dengue serotypes: a five years hospital based study from India. ***The Brazilian Journal of Infectious Diseases***, *14*(2), 141–146. Disponível em: https://doi.org/10.1016/S1413-8670(10)70027-1

KUNO, G.; CHANG, G.J. J. (2007). Full-length sequencing and genomic characterization of Bagaza, Kedougou, and Zika viruses. ***Archives of Virology***, *152*(4), 687–696. Disponível em: https://doi.org/10.1007/s00705-006-0903-z

KURANE, I. (2007). Dengue hemorrhagic fever with special emphasis on immunopathogenesis. ***Comparative Immunology, Microbiology and Infectious Diseases***, *30*(5–6), 329–340. 338p. Disponível em: https://doi.org/10.1016/j.cimid.2007.05.010

KURANE, I. *et. al* (1994). Immunopathologic mechanisms of dengue hemorrhagic fever and dengue shock syndrome. In *Positive-Strand RNA Viruses* (pp. 59–64). Vienna: **Springer Vienna**. Disponível em: https://doi.org/10.1007/978-3-7091-9326-6_7

KYLE, J. L.; HARRIS, E. (2008). Global Spread and Persistence of Dengue. ***Annual Review of Microbiology***, *62*(1), 71–92. Disponível em: https://doi.org/10.1146/annurev.micro.62.081307.163005

LANCIOTTI, R. S. *et. al*. (2008). Genetic and Serologic Properties of Zika Virus Associated with an Epidemic, Yap State, Micronesia, 2007. ***Emerging Infectious Diseases,*** *14*(8), 1232–1239. Disponível em: https://doi.org/10.3201/eid1408.080287

LEBRUN, G. *et. al* (2009). Guillain-Barré Syndrome after Chikungunya Infection. ***Emerging Infectious Diseases***, *15*(3), 495–496. Disponível em: https://doi.org/10.3201/eid1503.071482

LEDERMANN, J. P. *et. al* (2014). Aedes hensilli as a Potential Vector of Chikungunya and Zika Viruses. ***PLoS Neglected Tropical Diseases***, *8*(10), e3188. Disponível em: https://doi.org/10.1371/journal.pntd.0003188

LEITNER, T. *et. al*. (2017). HIV Sequence Compendium 2017 ***HIV Sequence Database*** Editors. *La-Ur-17-25240*, 1–454. Disponível em: https://www.hiv.lanl.gov./content/sequence/HIV/COMPENDIUM/2017/sequence2017.pdf

LEUNG, G. H. Y. *et. al*. (2015). Zika virus infection in Australia following a monkey bite in Indonesia. ***Southeast Asian Journal of Tropical Medicine and Public Health***, *46*(3), 460–464.Disponível em: https://www.tm.mahidol.ac.th/seameo/2015-46-3/09-642913p460.pdf

LI PIRA, G. *et. al*. (2010). High Throughput T Epitope Mapping and Vaccine Development. ***Journal of Biomedicine and Biotechnology***, *2010*, 1–12. Disponível em: https://doi.org/10.1155/2010/325720

LINARES, E. M. *et. al* (2013). Immunospot assay based on fluorescent nanoparticles for Dengue fever detection. ***Biosensors and Bioelectronics***, *41*, 180–185. Disponível em: https://doi.org/10.1016/j.bios.2012.08.005

LINDENBACH B.D. *et. al*. (2013) Flaviviridae: The viruses and their replication**. Fields Virology**, eds Knipe DM, Howley PM (Wolters Kluwer/Lippincott Williams & Wilkins Health, Philadelphia), 6th Ed. Disponível em: http://www.viala.org/img/Lindenbach2007.pdf

LO PRESTI, A. *et. al*. (2014). Chikungunya virus, epidemiology, clinics and phylogenesis: A review. ***Asian Pacific Journal of Tropical Medicine***, *7*(12), 925–932. Disponível em: https://doi.org/10.1016/S1995-7645(14)60164-4

MARCONDES, C. B.; XIMENES, M. DE F. F. DE M. Zika virus in Brazil and the danger of infestation by aedes (Stegomyia) mosquitoes. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 49, n. 1, p. 4–10, 2016. Disponível em: https://doi.org/10.1590/0037-8682-0220-2015.

MINISTÉRIO DA SAÚDE. (2015). Febre pelo vírus Zika: uma revisão narrativa sobre a doença. ***Boletim Epidemiológico***, *46*(26), 1–7 Disponível em: https://www.saude.gov.br/images/pdf/2015/agosto/26/2015-020-publica----o.pdf

MONATH, T. P. *et. al*. (2013). Yellow fever vaccine. ***Vaccines*** (pp. 870–968). Elsevier. Disponível em: https://doi.org/10.1016/B978-1-4557-0090-5.00043-4

MONATH, T. P.; VASCONCELOS, P. F. C. (2015). Yellow fever. ***Journal of Clinical Virology***, *64*, 160–173. Disponível em: https://doi.org/10.1016/j.jcv.2014.08.030

MORGULIS, A. *et. al*. (2008). Database indexing for production MegaBLAST searches. ***Bioinformatics***, *24*(16), 1757–1764. Disponível em: https://doi.org/10.1093/bioinformatics/btn322

MORRISON, C. R.; PLANTE, K. S.; HEISE, M. T. (2016). Chikungunya Virus: Current Perspectives on a Reemerging Virus. ***Emerging Infections*** *10* (pp. 143–161). Washington, DC, USA: ASM Press. Disponível em: https://doi.org/10.1128/9781555819453.ch8

MUSSO, D *et. al*. (2014). Potential for Zika virus transmission through blood transfusion demonstrated during an outbreak in French Polynesia, November 2013 to February 2014. ***Eurosurveillance***, *19*(14), 20761. https://doi.org/10.2807/1560-7917.ES2014.19.14.20761

MUSSO, D. *et. al*. (2015). Potential Sexual Transmission of Zika Virus. ***Emerging Infectious Diseases***, *21*(2), 359–361. Disponível em: https://doi.org/10.3201/eid2102.141363

MUSTAFA, M. S. *et. al* (2015). Discovery of fifth serotype of dengue virus (DENV-5): A new public health dilemma in dengue control. ***Medical Journal Armed Forces India***, *71*(1), 67–70. Disponível em: https://doi.org/10.1016/j.mjafi.2014.09.011

MUTEBI, J.-P. *et. al*. (2001). Phylogenetic and Evolutionary Relationships among Yellow Fever Virus Isolates in Africa. ***Journal of Virology***, *75*(15), 6999–7008. Disponível em: https://doi.org/10.1128/JVI.75.15.6999-7008.2001

NAVECA, F. G. *et. al*. (2019). Genomic, epidemiological and digital surveillance of Chikungunya virus in the Brazilian Amazon. ***PLOS Neglected Tropical Diseases***, *13*(3), e0007065. Disponível em: https://doi.org/10.1371/journal.pntd.0007065

NEEDLEMAN, S. B.; CHRISTIAN, L. N. D.; WUNCH, D. (1970). *A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins*. ***Journal of Molecular Biology***Disponível em: https://doi.org/10.1016/0022-2836(70)90057-4

NUNES, M. R. T. *et. al* (2015). Emergence and potential for spread of Chikungunya virus in Brazil. ***BMC Medicine***, *13*(1), 102. Disponível em: https://doi.org/10.1186/s12916-015-0348-x

OLIVEIRA MELO, A. S. *et. al*. (2016). Zika virus intrauterine infection causes fetal brain abnormality and microcephaly: tip of the iceberg? ***Ultrasound in Obstetrics & Gynecology***, *47*(1), 6–7. Disponível em: https://doi.org/10.1002/uog.15831

OLSON, J. G. *et. al* (1981). Zika virus, a cause of fever in Central Java, Indonesia. ***Transactions of the Royal Society of Tropical Medicine and Hygiene***, *75*(3), 389–393. Disponível em: https://doi.org/10.1016/0035-9203(81)90100-0

OPS/OMS. (2013). *Fiebre por Chikungunya. Alerta Epidemiológica 9 de diciembre 2013.* 1–5 ***World Health Organization*** Disponível em: https://www.paho.org/hq/dmdocuments/2013/9-diciembre-2013-Chikungunya-Alerta-Epi.pdf

PABBARAJU, K. *et. al*. (2016). Simultaneous detection of Zika, Chikungunya and Dengue viruses by a multiplex real-time RT-PCR assay. ***Journal of Clinical Virology***, *83*, 66–71. Disponível em: https://doi.org/10.1016/j.jcv.2016.09.001

PANNING, M. *et. al* (2008). Chikungunya Fever in Travelers Returning to Europe from the Indian Ocean Region, 2006. ***Emerging Infectious Diseases***, *14*(3), 416–422. Disponível em: https://doi.org/10.3201/eid1403.070906

PATKAR, C. G. *et. al*. (2009). Identification of Inhibitors of Yellow Fever Virus Replication Using a Replicon-Based High-Throughput Assay. ***Antimicrobial Agents and Chemotherapy,*** *53*(10), 4103–4114. Disponível em: https://doi.org/10.1128/AAC.00074-09

PETERSEN, L. R.; POWERS, A. M. (2016). Chikungunya: epidemiology. ***F1000Research***, *5*, 82. Disponível em: https://doi.org/10.12688/f1000research.7171.1

PINEDA-PEÑA, A. *et. al*. (2013). 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***, *19*, 337–348. Disponível em: https://doi.org/10.1016/j.meegid.2013.04.032

PLOURDE, A. R.; BLOCH, E. M. (2016). A Literature Review of Zika Virus. ***Emerging Infectious Diseases,*** *22*(7), 1185–1192. Disponível em: https://doi.org/10.3201/eid2207.151990

POLANSKI, A.; Kimmel, M. *Bioinformatics*. (2007). **Springer** Berlin Heidelberg. Disponível em: https://doi.org/10.1007/978-3-540-69022-1

POST, P. R. *et. al.* (1992). Heterogeneity in envelope protein sequence and N-Linked glycosylation among yellow fever virus vaccine strains. ***Virology***, *188*(1), 160–167. Disponível em: https://doi.org/10.1016/0042-6822(92)90745-B

POWELL, J. R.; TABACHNICK, W. J. (2013). History of domestication and spread of Aedes aegypti - A Review. ***Memórias Do Instituto Oswaldo Cruz***, *108*(suppl 1), 11–17. Disponível em: https://doi.org/10.1590/0074-0276130395

POWERS, A. M.; LOGUE, C. H. (2007). Changing patterns of chikungunya virus: re-emergence of a zoonotic arbovirus. ***Journal of General Virology***, *88*(9), 2363–2377. Disponível em: https://doi.org/10.1099/vir.0.82858-0

PRATHEEK, B. M. *et. al* (2015). In silico analysis of MHC-I restricted epitopes of Chikungunya virus proteins: Implication in understanding anti-CHIKV CD8+ T cell response and advancement of epitope based immunotherapy for CHIKV infection. ***Infection, Genetics and Evolution***, *31*, 118–126. Disponível em: https://doi.org/10.1016/j.meegid.2015.01.017

PRUESSE, E.; PEPLIES, J.; GLÖCKNER, F. O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. ***Bioinformatics***, *28*(14), 1823–1829. Disponível em: https://doi.org/10.1093/bioinformatics/bts252

REZZA, G.; WEAVER, S. C. (2019). Chikungunya as a paradigm for emerging viral diseases: Evaluating disease impact and hurdles to vaccine development. ***PLOS Neglected Tropical Diseases***, *13*(1), e0006919. Disponível em: https://doi.org/10.1371/journal.pntd.0006919

RICE, C. *et. al*. (1985). Nucleotide sequence of yellow fever virus: implications for flavivirus gene expression and evolution. ***Science***, *229*(4715), 726–733. Disponível em: https://doi.org/10.1126/science.4023707

ROBINSON, M. C. (1957). An epidemic of virus disease in Southern Province, Tanganyika Territory, in 1952–1953. ***Transactions of the Royal Society of Tropical Medicine and Hygiene***, *51*(3), 238–240. Disponível em: https://doi.org/1016/0035-9203(55)90080-8

ROTH, A. *et. al* (2014). Concurrent outbreaks of dengue, chikungunya and Zika virus infections – an unprecedented epidemic wave of mosquito-borne viruses in the Pacific 2012–2014. ***Eurosurveillance***, *19*(41). Disponível em: https://doi.org/10.2807/1560-7917.ES2014.19.41.20929

RÜCKERT, C., *et. al* (2017). Impact of simultaneous exposure to arboviruses on infection and transmission by Aedes aegypti mosquitoes. ***Nature Communications***, *8*(1), 15412. Disponível em: https://doi.org/10.1038/ncomms15412

SAAVEDRA, M. (2017). Early-life disease exposure and occupational status: The impact of yellow fever during the 19th century. ***Explorations in Economic History***, *64*, 62–81. Disponível em: https://doi.org/10.1016/j.eeh.2017.01.003

SAM, I.C. *et. al* (2012). Genotypic and Phenotypic Characterization of Chikungunya Virus of Different Genotypes from Malaysia. ***PLoS ONE***, *7*(11), e50476. Disponível em: https://doi.org/10.1371/journal.pone.0050476

SAM, I.C. *et. al* (2015). Updates on Chikungunya Epidemiology, Clinical Disease, and Diagnostics. ***Vector-Borne and Zoonotic Diseases***, *15*(4), 223–230. Disponível em: https://doi.org/10.1089/vbz.2014.1680

SAMUEL, G. H. *et. al*. (2016). Yellow fever virus capsid protein is a potent suppressor of RNA silencing that binds double-stranded RNA. ***Proceedings of the National Academy of Sciences***, *113*(48), 13863–13868. Disponível em: https://doi.org/10.1073/pnas.1600544113

SANGKAWIBHA, N. *et. al* (1984). Risk Factors In Dengue Shock Syndrome: A Prospective Epidemiologic Study In Rayong, Thailand. ***American Journal Of Epidemiology***, 120(5), 653–669. Disponível em: https://doi.Org/10.1093/Oxfordjournals.Aje.A113932

SANTOS, J. J. da S. *et. al* (2015). Full-length infectious clone of a low passage dengue virus serotype 2 from Brazil. ***Memórias Do Instituto Oswaldo Cruz***, *110*(5), 677–683. Disponível em: https://doi.org/10.1590/0074-02760150053

SCHWARTZ, O.; ALBERT, M. L. (2010). Biology and pathogenesis of chikungunya virus. ***Nature Reviews Microbiology***, *8*(7), 491–500. Disponível em: https://doi.org/10.1038/nrmicro2368

SETTE, A.; PETERS, B. (2007). Immune epitope mapping in the post-genomic era: lessons for vaccine development. ***Current Opinion in Immunology***, *19*(1), 106–110. Disponível em: https://doi.org/10.1016/j.coi.2006.11.002

SHI, W *et. al*. (2016). Increasing genetic diversity of Zika virus in the Latin American outbreak. ***Emerging Microbes & Infections***, *5*(1), 1–3. Disponível em: https://doi.org/10.1038/emi.2016.68

SIMON, F. *et. al* (2015). French guidelines for the management of chikungunya (acute and persistent presentations). November 2014. ***Medecine et Maladies Infectieuses***, *45*(7), 243–263. Disponível em: https://doi.org/10.1016/j.medmal.2015.05.007

SIMPSON, D. I. H. (1964). Zika virus infection in man. ***Transactions of the Royal Society of Tropical Medicine and Hygiene***, *58*(4), 339–348. Disponível em: https://doi.org/10.1016/0035-9203(64)90201-9

SISSOKO, D. *et. al* (2008). Outbreak of Chikungunya fever in Mayotte, Comoros archipelago, 2005–2006. ***Transactions of the Royal Society of Tropical Medicine and Hygiene,*** *102*(8), 780–786. Disponível em: https://doi.org/10.1016/j.trstmh.2008.02.018

SMITH, T. F.; WATERMAN, M. S. (1981). *Identification of Common Molecular Subsequences*. ***Journal of Molecular Biology*** 195–197. Disponível em: https://doi.org/10.1016/0022-2836(81)90087-5

SORIA SEGARRA, C. *et. al* (2018). Aplicación y aceptabilidad de la Guía Clínica de Dengue OMS-2009: la percepción de Ecuador - Aplicability and Aceptability and of the Dengue Clinical Guidelines WHO 2009: Ecuador's perception. ***Rev. Medica Electron***, *40*(4), 989–1001. Disponível em: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1684-18242018000400007

STATES, E. E. A. M., *et. al*. (2016). RAPID RISK ASSESSMENT Zika virus disease epidemic: potential association with microcephaly and Guillain-Barré syndrome (1 st update) Main conclusions. ***European Centre for Disease Prevention and Control***, *January*, 1–20.Disponível em: https://www.ecdc.europa.eu/en/publications-data/rapid-risk-assessment-zika-virus-disease-epidemic-potential-association-4 Acessado em: 20 de março de 2016.

SUAYA, J. A. *et. al* (2009). Cost of dengue cases in eight countries in the Americas and asia: A prospective study. ***American Journal of Tropical Medicine and Hygiene***, *80*(5), 846–855. 849p. Disponível em: https://doi.org/10.4269/ajtmh.2009.80.846

SWANSTROM, J. A *et. al* (2016). Dengue Virus Envelope Dimer Epitope Monoclonal Antibodies Isolated from Dengue Patients Are Protective against Zika Virus. ***MBio***, *7*(4). Disponível em: https://doi.org/10.1128/mBio.01123-16

TAMI, A.; GRILLET, M. E.; GROBUSCH, M. P. Applying geographical information systems (GIS) to arboviral disease surveillance and control: A powerful tool. **Travel Medicine and Infectious Disease,** v. 14, n. 1, p. 9–10, 2016. Disponível em: https://doi.org/10.1016/j.tmaid.2016.01.002.

THIBERVILLE, S.D. *et. al*. (2013). Chikungunya fever: Epidemiology, clinical syndrome, pathogenesis and therapy. ***Antiviral Research***, *99*(3), 345–370. Disponível em: https://doi.org/10.1016/j.antiviral.2013.06.009

THOMAS, E. A.; JOHN, M; BHATIA, A. (2007). Cutaneous manifestations of dengue viral infection in Punjab (north India). ***International Journal of Dermatology***, *46*(7), 715–719. Disponível em: https://doi.org/10.1111/j.1365-4632.2007.03298.x

TSETSARKIN, K. A. *et. al* (2007). A Single Mutation in Chikungunya Virus Affects Vector Specificity and Epidemic Potential. ***PLoS Pathogens***, *3*(12), e201. Disponível em: https://doi.org/10.1371/journal.ppat.0030201

TUBOI, S. H. *et. al*. (2007). Clinical and epidemiological characteristics of yellow fever in Brazil: analysis of reported cases 1998–2002. ***Transactions of the Royal Society of Tropical Medicine and Hygiene***, *101*(2), 169–175. Disponível em: https://doi.org/10.1016/j.trstmh.2006.04.001

US FOOD AND DRUG ADMINISTRATION. Recommendations for donor screening, deferral, and product management to reduce the risk of transfusion transmission of Zika virus, recommendations for industry. ***Food and Drug Administration*** 2016 Disponível em: http://www.fda.gov/ downloads/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/Guidances/Blood/UCM486360.pdf. Acessado em: 20 de março de 2016.

VALDERRAMA, A.; DÍAZ, Y.; LÓPEZ-VERGÈS, S. (2017). Interaction of Flavivirus with their mosquito vectors and their impact on the human health in the Americas. ***Biochemical and Biophysical Research Communications***, *492*(4), 541–547. Disponível em: https://doi.org/10.1016/j.bbrc.2017.05.050

VALESKA ROSSETTO, E.; PAULO, S. What to expect from the 2017 yellow fever outbreak in Brazil? 2017. **Revista do Instituto de Medicina Tropical de São Paulo** Disponível em: https://doi.org/10.1590/S1678-9946201759017

VAUGHN, D. W. *et. al* (1996). Testing of a dengue 2 live-attenuated vaccine (strain 16681 PDK 53) in ten American volunteers. ***Vaccine***, *14*(4), 329–336. Disponível em: https://doi.org/10.1016/0264-410X(95)00167-Y

VENTURA, C. V. *et. al* (2016). Ophthalmological findings in infants with microcephaly and presumable intra-uterus Zika virus infection. *Arquivos Brasileiros de Oftalmologia*, *79*(1). https://doi.org/10.5935/0004-2749.20160002

VITA, R. *et. al* (2015). The immune epitope database (IEDB) 3.0. ***Nucleic Acids Research***, *43*(D1), D405–D412. Disponível em: https://doi.org/10.1093/nar/gku938

VITAL, C. *et. al*. (2002). Postvaccinal inflammatory neuropathy: peripheral nerve biopsy in 3 cases. ***Journal of the Peripheral Nervous System***, *7*(3), 163–167. Disponível em: https://doi.org/10.1046/j.1529-8027.2002.02010.x

VOIGT, U. *et. al*. (2001). Optikusneuritis nach Impfung gegen Hepatitis A, B und Gelbfieber mit irreversiblem Visusverlust1. ***Klinische Monatsblätter Für Augenheilkunde***, *218*(10), 688–690. Disponível em: https://doi.org/10.1055/s-2001-18392

VOLK, S. M. *et. al* (2010). Genome-Scale Phylogenetic Analyses of Chikungunya Virus Reveal Independent Emergences of Recent Epidemics and Various Evolutionary Rates. ***Journal of Virology***, *84*(13), 6497–6504. Disponível em: https://doi.org/10.1128/JVI.01603-09

WEAVER, S. C. *et. al* (2012). Chikungunya virus and prospects for a vaccine. ***Expert Review of Vaccines***, *11*(9), 1087–1101. Disponível em: https://doi.org/10.1586/erv.12.84

WHITEHORN, J.; Farrar, J. (2010). Dengue. ***British Medical Bulletin***, *95*(1), 161–173. Disponível em: https://doi.org/10.1093/bmb/ldq019

WHITEHORN, J.; SIMMONS, C. P. (2011). The pathogenesis of dengue**.** ***Vaccine***, *29*(42), 7221–7228. Disponível em: https://doi.org/10.1016/j.vaccine.2011.07.022

WILSON, M. E.; SCHLAGENHAUF, P. Aedes and the triple threat of DENV, CHIKV, ZIKV – Arboviral risks and prevention at the 2016 Rio Olympic games. **Travel Medicine and Infectious Disease**, v. 14, n. 1, p. 1–4, 2016. Disponível em: https://doi.org/10.1016/j.tmaid.2016.01.010

WORLD HEALTH ORGANIZATION (WHO). Global strategy for dengue prevention and control 2012-2020. New ed. Geneva, Switzerland: **World Health Organization**; 2012

ZOU, D. *et. al*. (2015). Biological Databases for Human Research. ***Genomics, Proteomics & Bioinformatics***, *13*(1), 55–63. Disponível em: https://doi.org/10.1016/j.gpb.2015.01.006