FUNDAÇÃO OSWALDO CRUZ

INSTITUTO GONÇALO MONIZ

FIOCRUZ

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

DISSERTAÇÃO DE MESTRADO

FERRAMENTA DE ANÁLISE DE HEPATITE, BANCO DE DADOS DE SEQUÊNCIAS - HATsDB - *HEPATITIS ANALYSIS TOOL, SEQUENCE DATABASE*

HELTON FÁBIO SANTOS DE ARAÚJO JÚNIOR

Salvador – Bahia

2020

FUNDAÇÃO OSWALDO CRUZ

INSTITUTO GONÇALO MONIZ

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

FERRAMENTA DE ANÁLISE DE HEPATITE, BANCO DE DADOS DE SEQUÊNCIAS - HATsDB - *HEPATITIS ANALYSIS TOOL, SEQUENCE DATABASE*

HELTON FÁBIO SANTOS DE ARAÚJO JÚNIOR

Orientador: Prof. Dr. Artur Trancoso Lopo de Queiroz

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

Salvador – Bahia

2020

"HATsDB - Hepatitis Analysis Tool, sequence DataBase".

HELTON FÁBIO SANTOS DE ARAÚJO JÚNIOR

FOLHA DE APROVAÇÃO

Salvador, 03 de novembro de 2020.

COMISSÃO EXAMINADORA

Dra. Isabel Maria Vicente Guedes de Carvalho Mello
Pesquisadora
INSTITUTO BUTANTAN

Dr. Phellippe Arthur Santos Marbach
Professor
UFRB

Dr. Pablo Ivan Pereira Ramos
Pesquisador
IGM/FIOCRUZ

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

Ao **Prof. Dr. Artur Trancoso Lopo de Queiroz**, pela orientação ao longo destes anos entre iniciação científica e mestrado, por toda a competência, companheirismo, disponibilidade e resiliência que tornou toda esta jornada uma experiência incrível, de muito aprendizado e crescimento. Muito obrigado por acreditar no meu potencial, sem o seu apoio nada disso seria possível, agradeço profundamente pelos ensinamentos, pelas brincadeiras para descontrair os diversos momentos em que me via ilhado em limitações técnicas e por demonstrar que eu tenho com quem contar.

Ao **Dr. José Irahe Kasprzykowski Gonçalves**, pela coorientação, pelo universo de novos conhecimentos computacionais, por ser o primeiro a confiar nas minhas habilidades e potenciais e me convidar a ingressar na iniciação científica, pelos direcionamentos e todo o companheirismo. Certamente você é responsável também por possibilitar tudo isso, muito obrigado, eu “te vejo na tia do *pot* em *Darashia*”.

Aos membros da banca examinadora, **Dra. Isabel Maria Vicente Guedes de Carvalho Mello**, **Dr. Phellippe Arthur Santos Marbach** e **Dr. Pablo Ivan Pereira Ramos**, que prontamente se disponibilizaram a colaborar com esta dissertação com comentários e apontamentos enriquecedores.

Aos demais membros que integram a nossa equipe de pesquisa e em especial ao **Dr. Kiyoshi Ferreira Fukutani** por sempre me trazer crescimento pessoal, a **Eduardo Fukutani Rocha** pelo companheirismo e trocas de conhecimentos desde o IC e gostaria de citar mais uma vez o **Dr. Pablo Ivan Pereira Ramos**, sinto que deve ser pontuado que foi muito enriquecedor ter tido a oportunidade de ser aluno durante o mestrado e sou muito grato pelos ensinamentos e pelos apontamentos ao longo da minha formação.

Aos demais membros do CPqGM, especialmente os que integram o PGBSMI, como o vice-coordenador, **Dr. Prof. Luciano Kalabrick**, a amiga **Caroline Sousa** e a **Simone Farias Silva** por toda a paciência e disponibilidade. Também deixo aqui o meu enorme agradecimento a **Noélia Santos** e à Biblioteca do IGM pela disponibilização do espaço e pela revisão deste trabalho, em particular à **Ana Maria Fiscina**.

À **Fiocruz** e à **CAPES** pela realização do curso e por tornar possível a execução deste trabalho, disponibilizando infraestrutura, financiamento, acervo acadêmico e

todo ecossistema necessário.

À minha família, em especial aos meus pais, **Helton Araújo** e **Rita Canário Araújo**, e irmã **Janaína Andrade**, por me apoiarem e terem sido compreensivos em relação à minha ausência. Obrigado por tudo… Por se preocuparem, pelo carinho, pelo afeto, o zelo de vocês é imprescindível e eu não me arrisco a dizer qual o rumo as coisas teriam sem vocês ao meu lado.

À minha namorada **Luana Karam** que me apoiou, não me deixou desanimar um minuto, e me ensinou muito no decorrer do curso. Além de ser uma parceira incrível, ajudou na confecção deste trabalho. Obrigado por estar sempre ao meu lado, mesmo em alguns momentos sem poder dar a atenção que você merecia e sem ter tido os momentos de lazer. Obrigado pela felicidade de partilhar a dádiva da vida, e pelos “23:08, 1, 2, 3, e play”.

Finalizo agradecendo a todos aqueles que direta ou indiretamente contribuíram para a execução deste trabalho e/ou para a escrita dos manuscritos, e agradeço enormemente a quem contribuiu com o meu crescimento acadêmico, pessoal e técnico.

ARAÚJO JÚNIOR, Helton Fábio Santos de. Ferramenta de análise de hepatite, banco de dados de sequências – HATsDB. 2020. 65 f. Dissertação (Mestrado em Biotecnologia em Saúde e Medicina Investigativa) – Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, 2020

RESUMO

INTRODUÇÃO: A Hepatite C, doença silenciosa causada pelo Vírus da Hepatite C (VHC), representa um problema na saúde pública global. Segundo a OMS (Organização Mundial da Saúde) cerca de 399.000 pessoas morrem anualmente por Hepatite C ou pelas doenças associadas, como cirrose ou carcinoma hepatocelular. Para lidar com a hepatite e evitar as suas possíveis progressões, é necessária a criação de vacinas. No entanto, os dados especializados necessários para o desenvolvimento de novos fármacos e tratamentos não estão disponíveis de forma centralizada em uma base de dados. Embora o VHC apresente risco severo à saúde pública, não há um banco de dados atualizado com análises biológicas sobre o vírus. Existem apenas bancos de dados primários e bancos de epítopos, e a única base de dados específica para o VHC foi descontinuada. **OBJETIVO:** Para obter uma vacina, é de grande valia que haja a informação sobre a distribuição dos subtipos e o mapeamento do genoma. Assim é possibilitado também que os epítopos imunogênicos, representados pela menor parcela do antígeno capaz de gerar uma resposta imune no hospedeiro, sejam mapeados, consequentemente se encontre a ordem de relevância de todos os epítopos. Desta forma, objetivou-se o desenvolvimento de um sistema autônomo que foca em obter dados primários e realizar mapeamento de sequências, classificação destas sequências e mapeamento de epítopos. **MATERIAL e MÉTODOS:** Neste trabalho, foi realizado um estudo in-sílico no propósito da obtenção dos dados primários, mapeamento e subtipagem de sequências genômicas utilizando modelos matemáticos para realizar alinhamentos conhecidos como matrizes de pontuação por posição, fazendo assim a classificação destas sequências em genótipos e subtipos. Desse modo, o próximo passo corresponde ao processo de mapeamento dos epítopos imunogênicos através de um algoritmo de janela deslizante, buscando encontrar regiões genômicas nas sequências de VHC correspondentes a um epítopo dentre todos os epítopos presentes no conjunto de dados que foram obtidos de um banco especializado. **RESULTADOS:** Se observa que o epítopo de sequência linear “YLLPRRGPRL” tem uma das maiores quantidades de correspondências (937.738 ocorrências totais e 17.384 ocorrências em registros únicos de sequências) em todas as sequências estando presente em 16 subtipos, e apresenta uma frequência de correspondência em sequências de 7,15% na quantidade total de sequências (não estando presente apenas nos subtipos 3b e 3k). Todavia, o epítopo de sequência linear “GSWHINRT” tem cerca da metade do total de correspondências (48.626 ocorrências totais e também para registros únicos de sequências), porém, está globalmente mais presente (em um total de 20% de todas as sequências). **CONCLUSÃO:** O HATsDB, os dados estatísticos dos mapeamentos, os backups de banco e os dados de

epítopos estão disponíveis em: http://pah.bahia.fiocruz.br/hat/

Palavras-chave: Sistema; Análise Genômica; Banco De Dados; Hepatite C; VHC; Epítopos Imunogênicos; Sequências Genômicas;

ARAÚJO JÚNIOR, Helton Fábio Santos de. HATsDB - Hepatitis Analysis tool,

sequence database.2020. 65 f. Dissertação Mestrado em Biotecnologia em Saúde e Medicina Investigativa) – Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, 2020.

ABSTRACT

INTRODUCTION: Hepatitis C, a silent disease caused by the Hepatitis C Virus (HCV), represents a global public health issue. According to WHO (World Health Organization), about 399,000 people die annually from Hepatitis C or associated diseases such as cirrhosis or hepatocellular carcinoma. In order to deal with hepatitis and prevent its possible progressions, it is necessary to create vaccines. However, the specialized data necessary for the development of new drugs and treatments are not available centrally in a database. Although HCV presents a severe public health risk, there is no updated database with biological analyzes of the virus. There are only primary databases and epitope banks, and the only HCV-specific database has been discontinued. **OBJECTIVE**: To obtain a vaccine, it is nice-to-have information on the distribution of the subtypes and the mapping of the genome. Thus, it is also possible that the immunogenic epitopes, represented by the smallest portion of the antigen capable of generating an immune response in the host, are mapped, consequently finding all epitopes relevance order. Thus, the aim was to develop an autonomous system that focuses on obtaining primary data and performing sequence mapping, classification of these sequences, and mapping of epitopes. **MATERIAL and METHODS**: In this work, an in-silicon study was carried out in order to obtain primary data, mapping and subtyping genomic sequences using mathematical models to perform alignments known as position scoring matrices, thus making the classification of these sequences into genotypes and subtypes. Thus, the next step corresponds to the process of mapping immunogenic epitopes through a sliding window algorithm, seeking to find genomic regions in the HCV sequences corresponding to an epitope among all the epitopes present in the data set that was obtained from a specialized bank. **RESULTS**: It is observed that the linear sequence epitope “YLLPRRGPRL” has one of the largest number of matches (937,738 total occurrences and 17,384 occurrences in single sequence records) in all sequences being present in 16 subtypes, and has a correspondence frequency in 7.15% of sequences in the total number of sequences (not only present in subtypes 3b and

3k). However, the linear sequence epitope “GSWHINRT” has about half of the total matches (48,626 total occurrences and also for single sequence records), however, it is globally more present (in a total of 20% of all sequences). **CONCLUSION**: HATsDB, mapping statistical data, database backups, and epitope data are available at http://pah.bahia.fiocruz.br/hat/

Keywords: System; Genomic Analysis; Databases; Hepatitis C; HCV; Immunogenic Epitopes; Genomic Sequences;

LISTA DE ABREVIAÇÕES E SIGLAS

aa	Aminoácidos
DDBJ	DNA Database Bank of Japan
DDoS	Distributed Denial-of-Service
DoS	Denial-of-Service
EMBL	European Molecular Biology Laboratory
FPR	False Positive Rate
FTP	File Transfer Protocol
GI	GenInfo Identifier
GPU	Graphics Processor Unit
HATsDB	Hepatitis C Analysis Tool, sequence DataBase
HCV	Hepatitis C Virus
HCV Databases	Hepatitis C Virus Databases
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IEDB	Immune Epitope DataBase
IP	Internet Protocol

LANL	Los Alamos National Laboratory
NCBI	National Center for Biotechnology Information
OMS	Organização Mundial da Saúde
RNA	Ribonucleic Acid
SOAP	Simple Object Access Protocol
TPR	True Positive Rate
VHC	Vírus da Hepatite C
XML	Extensive Markup Language
ZIKV	Zika Vírus

SUMÁRIO

1 INTRODUÇÃO E JUSTIFICATIVA 12

2 OBJETIVOS 14

2.1 OBJETIVO GERAL 14

2.2 OBJETIVOS ESPECÍFICOS 14

3 REVISÃO DE LITERATURA 15

3.1 SOBRE O VÍRUS 15

3.2 SOBRE A HEPATITE C 16

3.3 BANCO DE DADOS BIOLÓGICO 17

3.4 ALINHAMENTO DE SEQUÊNCIAS NUCLEOTÍDICAS 18

3.5 MAPEAMENTO DE SEQUÊNCIAS NO GENOMA COMPLETO 19

3.6 SUBTIPAGEM DAS SEQUÊNCIAS NUCLEOTÍDICAS 20

3.7 MAPEAMENTO DE EPÍTOPOS 21

4 MATERIAL E MÉTODOS 22

4.1 DOWNLOAD E ARMAZENAMENTO DE SEQUÊNCIAS 22

4.2 MAPEAMENTO E SUBTIPAGEM DE SEQUÊNCIAS 23

4.2.1 MAPEAMENTO DE SEQUÊNCIAS 23

4.2.2 SUBTIPAGEM DE SEQUÊNCIAS 24

4.3 TRANSFERÊNCIA E ARMAZENAMENTO DE EPÍTOPOS 25

4.4 MAPEAMENTO DE EPÍTOPOS 27

4.5 DESENVOLVIMENTO DO FRONT-END 29

4.6 VALIDAÇÃO DOS DADOS 29

5 RESULTADOS 30

5.1 DADOS 30

5.1.1 ALINHAMENTO E SUBTIPAGEM DE SEQUÊNCIAS 36

5.1.2 MAPEAMENTO DE EPÍTOPOS 36

5.2 PLATAFORMA DIGITAL 37

5.3 BENCHMARK E VALIDAÇÃO DO PROCESSO 38

6 DISCUSSÃO 40

6.1 SOBRE O TRABALHO 40

6.1.1 A discrepância entre número de sequências por subtipo 40

6.1.2 Epítopos 40
6.2 LIMITAÇÕES 41
6.2.1 O 7º genótipo 41
6.2.2 O congelamento do banco de dados 41
6.2.3 A “qualidade das sequências” 41
7 CONCLUSÃO 42
REFERÊNCIAS 43
Material suplementar 47

1 INTRODUÇÃO E JUSTIFICATIVA

As hepatites virais são consideradas importantes problemas de saúde pública (LYONS et al., 2016). Estas doenças são causadas por vírus do gênero *Hepacivirus,* da família *Flaviviridae* (HANAFIAH et al., 2013) e afetam cerca de 71 milhões de pessoas no mundo (World Health Organization, 2020). Destes indivíduos, cerca de 130 a 150 milhões desenvolvem a forma crônica da doença (LAWITZ et al., 2014). O VHC, possui 6 grupos distintos classificados filogeneticamente (FERREIRA, 2004). Apesar de não existirem diferenças dos dados clínicos, os infectados com tipos 1a e 1b apresentam uma menor resposta ao tratamento, resultando em maior tendência a desenvolverem uma progressão mais rápida da hepatite C crônica e da doença hepática, quando comparados aos que apresentam outros genótipos (SIMMONDS et al., 1994).

O VHC é um vírus com genoma em fita simples de sentido positivo (*RNA+ vírus*). Dentre seus genótipos, ainda podemos ter variações do VHC, que são chamadas *quasispecies* (FARCI, 2000), uma estrutura populacional de vírus com grande número de genomas variantes que geram constantes mutações (HANAFIAH et al., 2013). A dificuldade no seu estudo se dá ao fato do vírus ser um patógeno humano, não havendo animais para experimentação que se adaptem à pesquisa, com exceção do chimpanzé (*Pan troglodytes*), que além de ter custos elevados apresenta algumas características que inviabilizam o estudo (STRAUSS, 2001), somente 50% desenvolviam infecções persistentes do VHC enquanto os humanos desenvolvem em 80% dos casos (GOWER et al., 2014).

Para lidar com a hepatite C e evitar as suas progressões, é necessária a criação de vacinas e de mecanismos para tratamentos. Um fator que poderia auxiliar e acelerar esses avanços é a unificação de informações sobre os patógenos que poderiam ser acessadas mais rapidamente e praticamente em uma base de dados centralizada. Apesar de representar este risco severo à saúde pública, não há repositórios atualizados de sequências especializadas para esses vírus e suas variações. O LANL com o banco "*Hepatitis C Virus Database Project*" um dos principais bancos de dados de sequências desse vírus foi descontinuado em 2010. Entretanto, existe uma vasta gama de sequências nucleotídicas disponíveis nos bancos de dados primários como DDBJ, EMBL e GenBank.

Apesar da grande quantidade de sequências contidas nestes bancos de dados

primários, informações como classificação e mapeamento das sequências não estão disponíveis. Além disso, o monitoramento epidemiológico da prevalência de epítopos imunogênicos desse vírus não está abordado nem nos bancos de dados especializados, nem nos bancos de dados primários. Essas informações genômicas auxiliam na identificação de mutações associadas às características clínicas e na classificação de novas variantes sorotípicas (SIMMONDS et al., 1994). Neste projeto foi desenvolvida uma plataforma de aquisição e análise automática das sequências do VHC disponíveis em banco de dados primários. Além das análises genômicas, a plataforma irá realizar o mapeamento de epítopos provenientes de banco de dados especializados como o IEDB. Os resultados gerados estão disponíveis à comunidade científica em um banco de dados biológico especializado, para consulta e monitoramento dos respectivos dados nas seguintes formas:

1. *Front-end* para consulta dos resultados processados
2. Arquivos de *backup* do banco de dados com cortes temporais a cada 5 anos e modelo para transferência e replicação local

O principal objetivo das análises deste mapeamento é encontrar epítopos que sejam globalmente prevalentes, ou seja, epítopos imunogênicos que estejam presentes no RNA dos diversos subtipos do vírus de forma a tornar possível a criação de uma vacina que tenha área de cobertura do maior número de subtipos possível.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Desenvolver uma plataforma online unificada para análise genômica e armazenamento de informações sobre o Vírus da Hepatite C de forma que os seus dados e mapeamentos de epítopos sejam organizados de maneira centralizada.

2.2 OBJETIVOS ESPECÍFICOS

• Desenvolvimento do módulo de importação automática e periódica de sequências provenientes de bancos de dados primários

• Desenvolvimento do módulo de inserção manual por colaboradores do grupo de pesquisa (para uso somente quando houver uma necessidade extraordinária, como indisponibilidade dos repositórios primários, falta de comunicação de rede/internet etc);

• Mapeamento das sequências genômicas de VHC em relação aos genomas completos de referência;

• Desenvolvimento de um algoritmo de subtipagem das sequências nucleotídicas obtidas segundo os genótipos conhecidos e consolidados;

• Implementação de um algoritmo de mapeamento de epítopos T nas sequências do conjunto de dados;

• Desenvolvimento de *front-end web* para disponibilização dos resultados das análises em formatos tabulares e/ou brutos.

3 REVISÃO DE LITERATURA

3.1 SOBRE O VÍRUS

O *VHC* é um hepacivirus pertencente à família *Flaviviridae* (imagem 1) que infecta exclusivamente primatas, nos quais os humanos têm a maior taxa de infecção. Divide-se filogeneticamente em 7 genótipos, em que o genótipo 1 é o mais frequente no mundo inteiro (com uma frequência de cerca de 44%) seguido pelos genótipos 3, 2 e 4 respectivamente (GOWER et al., 2014). O genótipo 1 do VHC é relatado como clinicamente mais agressivo na maioria dos países, com exceção do sul da Ásia, norte da África e regiões próximas ao deserto do Saara (GOWER et al., 2014).

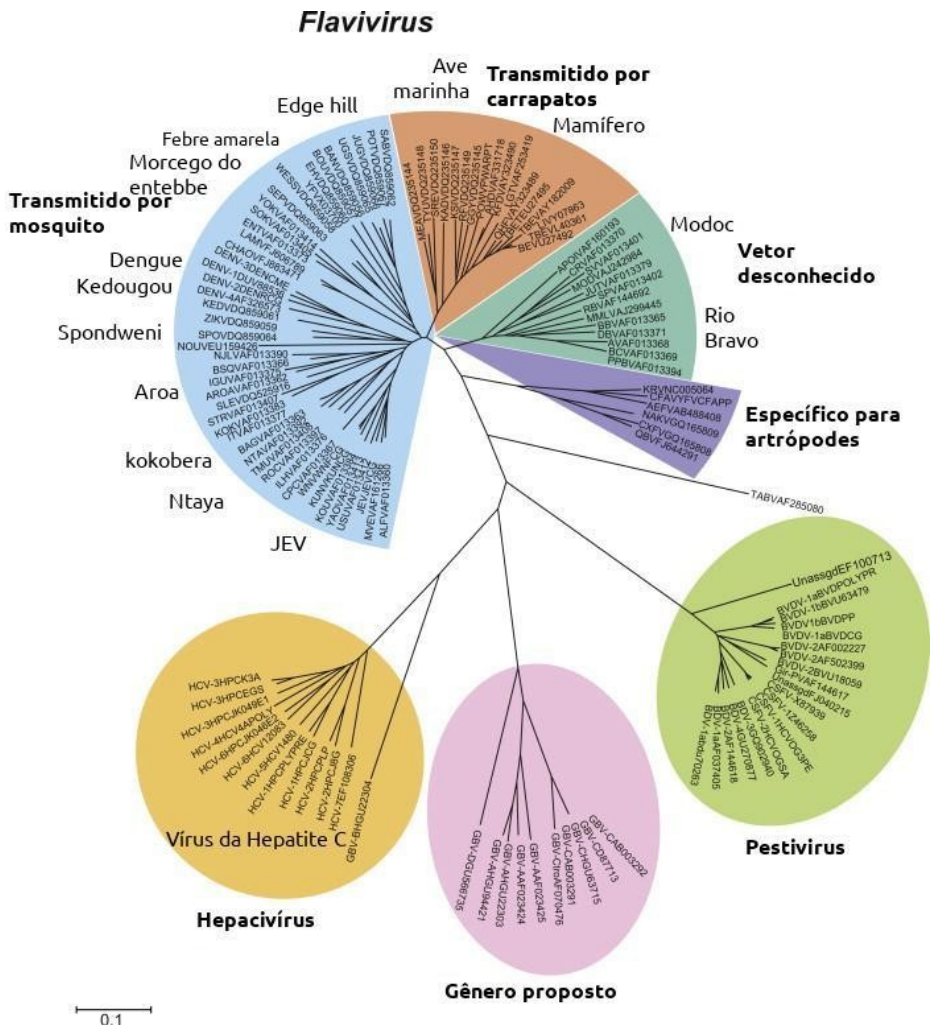


Figura 1 - Árvore filogenética dos Flavivírus. As cores representam os gêneros: em azul os vírus transmitidos por vetores, em laranja os transmitidos por carrapatos (ex. Ixodoidea), verde-escuro os com

vetores desconhecidos, em roxo específicos para artrópodes, em verde os pestivirus, rosa os gêneros propostos e em laranja os hepacivirus. Fonte: Adaptado de: (KING et al., 2012)

O vírus tem uma estrutura simples (Figura 2) e apresenta uma fita de RNA. A presença da enzima RNA polimerase sem atividade corretiva torna o vírus extremamente suscetível às mutações genéticas durante os processos de tradução e replicação, um envelope e algumas proteínas que são divididas em estruturais e não estruturais. Apesar de simples, o genoma do vírus codifica mais de 3000 aminoácidos (aa), a sequência de referência utilizada, NC_004102 que é um isolado H77, apresenta 9646 pares de bases, a estrutura do envelope é baseada em 2 proteínas funcionais (E1 e E2) que são essencialmente proteínas reconhecidas como próprias pelo sistema imune humano garantindo adsorção do vírus, enquanto a proteína funcional core (C) é responsável pelo nucleocapsídeo. Em conjunto, estas proteínas funcionais formam a parte estrutural do vírus. Há também a proteína p7 (uma pequena proteína em que não há informação suficiente ainda sobre sua função) e sete proteínas não estruturais (NS, NS2, NS3, NS4A, NS4B, NS5A e NS5B).

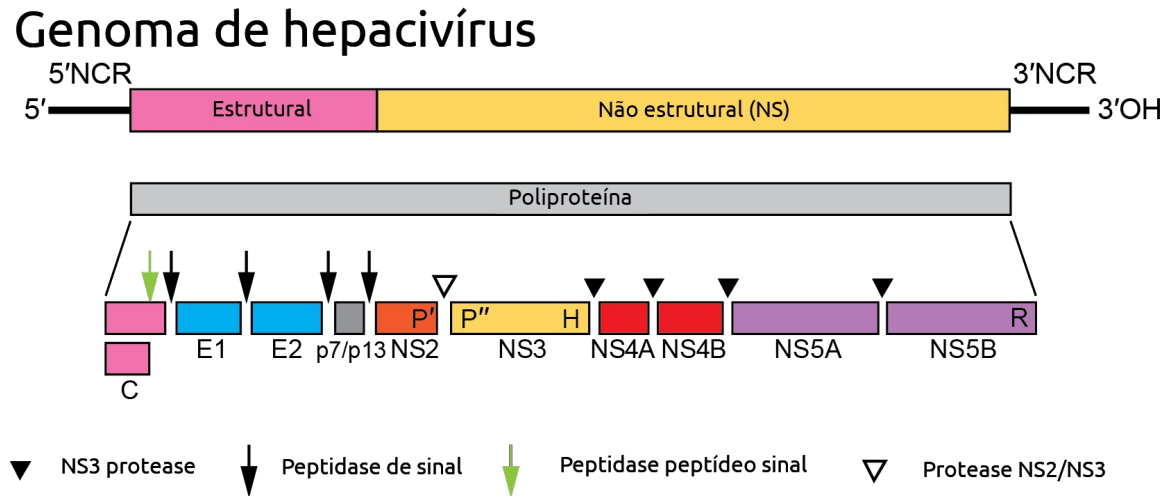


Figura 2 - Estrutura proteica do VHC. Fonte: Adaptado de: (KING et al., 2012)

3.2 SOBRE A HEPATITE C

A Hepatite C, a doença causada pelo *Vírus da Hepatite C* (*VHC*), é uma preocupação da saúde pública global, pois afeta cerca de 180 milhões de pessoas ao redor do mundo. A OMS aponta que a Hepatite C é responsável por 399 mil mortes por

ano (World Health Organization, 2017) e a doença ainda pode evoluir para fibrose hepática, carcinoma hepatocelular e cirrose.

Trata-se de uma doença predominantemente assintomática (PRATI et al., 1996) e a sua progressão para as demais enfermidades depende de cada organismo e do tempo de infecção. Estima-se que a evolução para uma cirrose leve de 20 a 30 anos, podendo variar de acordo com hábitos do paciente, enquanto a evolução para um câncer dure em média de 6 a 10 anos após o quadro de cirrose.

Não existe uma vacina que seja capaz de prevenir a Hepatite C, o que é preocupante num mundo cada vez mais globalizado. Análises epidemiológicas frequentes para esta etiologia são de extrema importância para que haja o direcionamento correto de qual vacina, quando houver, deve ser ministrada em certas regiões de modo a proteger a população local. No entanto, com a volatilidade em que as pessoas transitam de estado, país e de continentes, esta metodologia ainda assim poderá não ser o suficiente para proteger a população de forma eficaz.

Não obstante, os estudos sobre tratamentos são praticados e atualmente existem registros no *clinical trials* de 2402 estudos sobre infecção do VHC, das quais 164 já foram terminados e 1597 já estão completos. Dos estudos completados, 1290 foram de intervenção e tiveram um ensaio clínico. Destes 1290 estudos, 531 estudos são encontrados ao filtrar pelos que têm resultados. Ainda no *clinical trials* buscando por dados de vacinas, 55 estudos foram encontrados onde 38 foram completados e destes 6 houveram resultados.

3.3 BANCO DE DADOS BIOLÓGICO

É possível observar o avanço tecnológico no que diz respeito à quantidade de dados que são armazenados, analisados e tratados cotidianamente com a crescente do *Big Data*, o conjunto de dados biológicos vem acompanhando este aumento tanto em tamanho quanto em complexidade das informações. Desta forma torna-se imprescindível para a bioinformática a criação de uma plataforma para análise e gerenciamento dos dados biológicos (ZOU et al., 2015).

Existem bancos de dados chamados primários, como o GenBank, por exemplo, que contém dados brutos incluindo dados não curados. Bancos de dados chamados secundários por sua vez tiveram seus dados passados por um certo nível de curagem e

os especializados consecutivamente passaram por um processo de avaliação e classificação mais assíduos (ZOU et al., 2015).

Um problema no gerenciamento de bancos de dados biológicos (em particular nos especializados) é a manutenção deste, dado que o processo manual de curagem demanda uma equipe com um quadro equivalente ao tamanho da base de dados, com a qualificação adequada para manter os dados contidos e a sua qualidade. Ao passo que a base de dados aumenta, é comum se tornar inviável a sua manutenção por ser relativamente custosa, podendo ser ou não descontinuada, como, por exemplo, o banco *HCV Databases* do LANL que foi descontinuado em relação à manutenção apesar de continuar *online* e com *updates* de dados.

Com este problema em mente, nossa abordagem visava utilizar uma curagem automatizada, assim o banco de dados gerado sempre está em conformidade com a especialidade e especificidade dos dados. O fato de existir uma base de dados especializada e descontinuada desta etiologia corrobora com a necessidade de um sistema automatizado de curagem.

3.4 ALINHAMENTO DE SEQUÊNCIAS NUCLEOTÍDICAS

Uma das principais análises das sequências nucleotídicas é o alinhamento de sequências. Esse processo consiste na comparação entre duas sequências considerando a similaridade entre regiões dessas sequências. Existem regiões que podem ter o grau de similaridade modificado por ocorrência de inserções ou deleções de nucleotídeos na sequência, entretanto, não é possível determinar sem análises filogenéticas se houve uma inserção ou deleção em um determinado sítio.

Partindo da análise de duas sequências nucleotídicas é possível representá-las bidimensionalmente para a comparação entre elas de forma a identificar os pareamentos. O alinhamento resulta em um coeficiente de similaridade entre as duas sequências que pode ser utilizado, por exemplo, para classificar as sequências, além de uma sequência que representa o pareamento entre as regiões que é utilizada para mapear regiões genômicas (DESHMUKH, 2015).

O processo de alinhamento pode ser classificado em duas frentes principais: alinhamento global e alinhamento local. O alinhamento global assume que as cepas são homólogas, ou seja, têm um ancestral em comum, assim o alinhamento presume que as

sequências têm o mesmo tamanho, enquanto o alinhamento local busca o sítio com o maior nível de similaridade entre as sequências (DESHMUKH, 2015). O alinhamento global por considerar que as sequências são homólogas, busca sítios em uma sequência B onde é encontrado na sequência A (Ex: O sítio 1 da sequência A seria o mesmo da sequência B, assim como o sítio n para ambas) como pode ser observado na Figura 3.

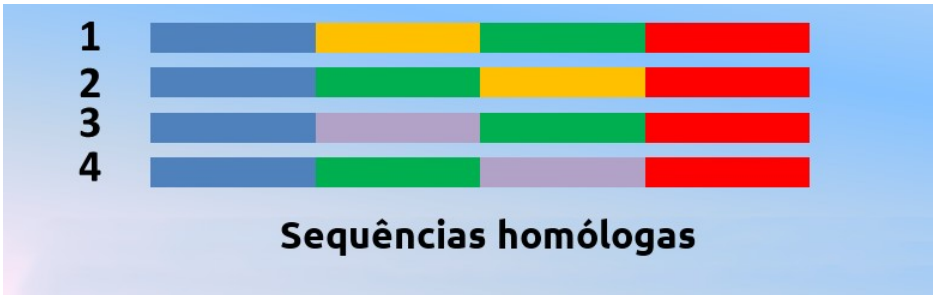


Figura 3 - Representação de homologia entre sequências, Fonte: Adaptado de: Wikimedia. Disponível em: https://commons.wikimedia.org/w/index.php?curid=30542439

Acessado em 20/01/2020

Existem duas principais metodologias para realizar o processo de alinhamento de sequências. Uma parte do princípio exaustivo, ou seja, que utiliza todas as possibilidades de maneira exaustiva até obter o melhor alinhamento possível considerando e ponderando as deleções e inserções na sequência (CHAKRABORTY, 2013) e a outra parte do princípio heurístico, que perde um pouco da acurácia ao não testar exaustivamente, mas é a abordagem que melhora a alocação e otimização dos recursos uma vez que as bases de dados crescem cada vez mais, e o resultado é apenas um dos melhores alinhamentos (CHAKRABORTY, 2013).

Alinhamentos de metodologia exaustiva, como os propostos por Needleman e Wunsch (1970), e Smith e Waterman (1981), apresentam uma ordem de grandeza na complexidade de algoritmos computacional proporcional ao tamanho das sequências, tornando assim a complexidade quadrática em relação a tempo e recurso computacional. Isso acontece porque ambos os algoritmos consistem em matrizes computacionais dinâmicas onde cada possibilidade de alinhamento é testada exaustivamente e anotada com a intenção de garantir sempre o melhor alinhamento entre as sequências.

3.5 MAPEAMENTO DE SEQUÊNCIAS NO GENOMA COMPLETO

De modo a conhecer melhor o genoma completo do organismo e suas estruturas para podermos compreender os diversos aspectos do vírus em estudo, é necessária a obtenção de informações traducionais e pós-traducionais.

Os genes que compõem o genoma viral podem ser divididos em dois grupos principais: estruturais e não estruturais, que se caracterizam por apresentarem aspectos críticos do ciclo de vida viral. Independente da sua classificação, essas estruturas apresentam em sua composição regiões de início e fim para a tradução.

Este processo consiste na identificação do escopo (início e fim) de uma dada sequência em um mapa de características anotadas no genoma completo de um organismo. Este procedimento é imprescindível para análises de dados obtidos de sequenciamento de alto desempenho, pois com este processo podemos identificar quais genes estão contidos na sequência e também, em caso de mutações, é possível identificar o gene em que a mutação ocorreu (COMBE e SANJUÁN, 2014).

Para o mapeamento ser realizado é necessário um mapa de características disponíveis no genoma completo do organismo, que é feito a partir da sequência de referência que representa o genoma completo deste organismo, onde há informações funcionais e posicionais, e estão disponíveis para a consulta. Assim é possível identificar se e onde há um determinado gene ou fragmento genético e determinar características que são comuns.

O mapeamento resulta em regiões de cobertura do genoma, de onde é possível a identificação das características que estão presentes no genoma completo e estão sendo referidas na sequência em questão. Desta forma pode-se classificar as sequências quanto a cobertura, estratificando quanto a um determinado gene ou região, aumentando a qualidade dos dados disponíveis no conjunto. Outrossim, é possível selecionar sequências que codificam regiões muito específicas, como regiões que codificam epítopos imunogênicos (a menor parcela do antígeno capaz de gerar uma resposta no sistema imune do hospedeiro), por exemplo, o que norteia uma abordagem para o desenvolvimento de vacinas e tratamentos (HE e ZHU, 2015).

3.6 SUBTIPAGEM DAS SEQUÊNCIAS NUCLEOTÍDICAS

De modo a melhorar a compreensão sobre um vírus, suas tendências mutacionais e como é dada sua interação com o hospedeiro é essencial classificá-los em grupos onde haja mais similaridades entre os indivíduos presentes em cada grupo. É possível identificar uma classificação assim a partir de um perfil já traçado de um grupo e do mapeamento das sequências em conjunto com as referências destes grupos (CHAN et al., 2014).

Desta forma, subtipar uma sequência consiste em aferir o grau de semelhança desta com os grupos já conhecidos e estabelecidos nas bases de dados e alocá-la no grupo mais adequado para as suas características. Para tal é realizado um mapeamento de sequência local com o intuito de verificar o coeficiente de similaridade para cada alinhamento entre esta sequência e todas as sequências de referência para todos os subtipos. O subtipo que resultar uma maior pontuação de semelhança é então aceito como sendo o grupo que melhor representa a sequência.

3.7 MAPEAMENTO DE EPÍTOPOS

Existem três classificações para epítopos: Lineares, em que o seu reconhecimento independe da estrutura tridimensional da proteína; Conformacionais, que depende da estrutura tridimensional para ser reconhecido; Hipotéticos, epítopos que estão em uma região proteína hipotética.

O foco para a análise e mapeamento está em epítopos lineares e conformacionais dado o foco em possibilidades mais assertivas para dados que levem a resultados com maiores oportunidades de estudos farmacogenômicos.

A identificação e mapeamento de epítopos imunogênicos é crucial para o desenvolvimento de vacinas (JONES, 2015). O processo de mapeamento é dado pela identificação de uma sequência de aminoácidos (*aa*) em peptídeos capazes de fazer com que o sistema imune seja ativado. Embora este processo seja extremamente importante para o entendimento da patogênese e criação de novas vacinas (JONES, 2015) e tratamentos, as informações sobre frequências de epítopos e seus respectivos mapeamentos não estão disponíveis em bancos de dados primários. Isso é explicado pela complexidade do processo, uma vez que tanto os conjuntos de dados de sequências quanto de epítopos são numerosos, o que torna o processo muito custoso.

Um dos repositórios mundiais de epítopos imunogênicos é o IEDB (VITA et al., 2014) que contém atualmente disponíveis para visualização e para *download* da base cerca de 1,275 milhões de epítopos, que podem ser classificados pela cadeia de aminoácidos, tipo de resposta, alelos, agentes etiológicos. Com a utilização de dados desta base é possível assim identificar no conjunto de sequências quais epítopos são cobertos e assim determinar quais regiões e referenciar cada ocorrência.

4 MATERIAL E MÉTODOS

4.1 *DOWNLOAD* E ARMAZENAMENTO DE SEQUÊNCIAS

As sequências utilizadas foram obtidas do NCBI no banco de dados *Nucleotide*. Apesar de existirem diversos métodos para a obtenção de dados, o método escolhido e utilizado foi o *ESearch* (Entrez Programming Utilities Help, 2016). Esta ferramenta foi criada para permitir o acesso à plataforma de base de dados (MCENTYRE, 1998) e selecionar qual base de dados a ser acessada dentro do NCBI. A ferramenta *EFetch* seleciona a lista de identificadores de sequências nucleotídicas. Entretanto, estas ferramentas têm uma limitação de *query* de apenas 200 sequências por requisição (SAYERS, 2009).

Para obter mais sequências nós desenvolvemos a ferramenta chamada *Nseek*, esta usa o *ESearch* para identificar os identificadores únicos numa base de dados. Assim é possível criar uma lista de diferenças entre os números de identificação de sequências já baixados e os identificadores únicos de novas sequências a serem obtidas em subconjuntos de 200 unidades cada. Essas novas sub-listas são separadas em *threads*, permitindo um *download* assíncrono em paralelo. A fim de ter mais segurança na conexão de obtenção, um espaço de tempo de 30 minutos foi implementado para evitar um bloqueio de IP do NCBI (implementado a fim de evitar ataques de negação de serviço dos tipos DDoS e DoS). Consequentemente, quando o *download* de um *SOAP* falha, a ferramenta reinicia o *download* deste enviando uma requisição via *POST* (método de requisição comumente utilizado pelo protocolo de rede *HTTP*). Todos os *downloads* que falham são novamente atribuídos pelo gerenciador de serviços, o que otimiza o tempo de operação e a carga de rede. Depois que todos os dados são baixados, a lista de diferenças contém todos os identificadores da base de dados. O *NSeek* armazena todos os dados no HATsDB e uma rotina de atualização à base local é feita de forma periódica

(tempo arbitrariamente escolhido de 1 semana).

As sequências são armazenadas assim que são baixadas, num processo que roda simultâneo ao *download*. Entretanto, em caso de falha no *download* ou no armazenamento, o arquivo *GenBank* pode ser baixado manualmente. Por causa do tamanho de alguns arquivos multi-gbk (arquivos *GenBank*, que podem ser maiores que 30 GB por instância) nós desenvolvemos um algoritmo de leitura gradual para reduzir a alocação de memória durante o processo de importação. Desta forma, enquanto os arquivos são processados o *NSeek* executa a demodulação e armazenamento no HATsDB usando o método manual ou automático. O processo de *download* está representado de maneira simplificada no Fluxograma 2.

Foi decidido, de forma arbitrária, que 7 dias é um espaço de tempo suficiente para que seja feita a atualização desses dados, assim como o de epítopos também.

As informações que são salvas sobre a sequências são:

- Sequência linear
- Versão de adesão (*accession version*)
- Locus
- Definição (uma breve descrição da sequência)
- Tamanho da sequência
- *GI*
- Tipo de molécula
- Topologia da sequência
- Taxonomia
- País de publicação
- Data de submissão
- Identificador de publicação no pubmed
- Todas as *sequence feature* disponíveis sobre a sequência

4.2 MAPEAMENTO E SUBTIPAGEM DE SEQUÊNCIAS

4.2.1 **Mapeamento de Sequências**

Para realizar o mapeamento, é necessário criar um “mapa” base para o vírus, criado a partir de uma sequência de referência. A sequência de referência (*refseq*) utilizada foi a de identificação NC_004102.1, definida como “*Hepatitis C virus genotype 1*” publicada pelo NCBI (2017) e contém diversas informações como sítios de codificações. Esta sequência foi escolhida por ser identificada como o isolado H77, e ser amplamente aceite como uma referência para o genoma completo. Com o mapa definido, é feito o alinhamento global entre as sequências do HATsDB (o próprio HATsDB) e a sequência do mapa base, desta forma podemos identificar o escopo de cada alinhamento, podendo assim identificar a localização de um determinado fragmento no genoma completo. O processo de mapeamento está exemplificado no Fluxograma 3.

4.2.2 **Subtipagem de Sequências**

Para a subtipagem deve ser criado um mapa classificatório baseado nas variantes existentes e suas respectivas cepas de referência, que foram coletadas na literatura base do VHC, o que resultou na identificação de 6 genótipos (SIMMONDS et al., 1993). Há no HATsDB o registro de 18 cepas de referência que podem ser vistas na tabela 1, que são relativas a quantidades de subtipos, o que gera um aumento da complexidade de classificação tornando-o mais custoso e demorado, além da quantidade relativamente grande de sequências.

Tabela 1 - Lista de genótipos, subtipos e *accession version*.

Fonte: Autor.

Genótipo	Subtipo	Acession version
1	1a	M62321.1, M67463.1
1	1b	D90208.1, M58335.1
1	1c	D14853.1, AY051292.1
2	2a	D00944.1, AB047639.1

2	2b	D10988.1, AB030907.1
2	2c	D50409.1
2	2k	AB031663.1
3	3a	D17763.1, D28917.1
3	3b	D49374.1
3	3k	D63821.1
4	4a	Y11604.1
5	5a	Y13184.1, AF064490.1
6	6a	Y12083.1, AY858526.2
6	6b	D84262.2
6	6d	D84263.2
6	6g	D63822.1
6	6h	D84265.2
6	6k	D84264.2

Após a identificação destas cepas de referência (SIMMONDS et al., 1994) e da metodologia a ser utilizada para o alinhamento é preciso pontuar que o VHC por ter uma conhecida alta taxa de mutação genética termina por tornar imprecisa a análise heurística e assim necessária um mapeamento local da metodologia exaustiva de modo a testar todas as possíveis combinações de inserção e deleção (Fluxograma 4). Assim, se utilizando de técnicas já validadas pela nossa equipe foi possível classificar todas as sequências em uma quantidade de tempo cabível, de algumas horas.

4.3 TRANSFERÊNCIA E ARMAZENAMENTO DE EPÍTOPOS

Ainda que as informações sobre epítopos sejam muito valiosas para análises e criação de vacinas (JONES, 2015), não estão disponíveis em bancos de dados biológicos primários, o que é o esperado, pois é preciso ser feito um processo de mapeamento dos epítopos em proteínas codificadas, sendo necessário primeiramente obter as sequências de aminoácidos de cada epítopo, um processo demorado e de alto custo computacional.

Todavia as informações de epítopos estão contidas em bancos de dados

especializados para este fim como, por exemplo, o IEDB. Os dados obtidos desta base contém uma série de epítopos identificados para diversos organismos e inclusive epítopos hipotéticos (que foram dispensados para a base de dados local), e suas informações como publicação, alelo e sequência linear.

O *IEDB* disponibiliza através de um *FTP* (protocolo de transferência de arquivos do protocolo *HTTP*) um arquivo com um nível de compactação muito alto contendo arquivos do tipo *XML* (arquivo de marcação de informações que serve para definir um padrão de texto). Assim o processo de download e persistência desses dados, que está representado no Fluxograma 5, é feito através de um algoritmo desenvolvido pela nossa equipe que consiste nos seguintes passos:

1 O *back-end* solicita ao servidor via protocolo *FTP* a transferência do arquivo

1.1 Caso haja o arquivo e a requisição seja positiva (código 200) o arquivo é enviado

1.2 Caso não haja o arquivo e/ou a requisição seja recusada é enviada uma mensagem de erro (código 404)

2 Após o *download* do arquivo

2.1 verifica-se se é a primeira vez que foi baixado

2.1.1 caso seja o algoritmo vai para o passo 3

2.1.2 caso não seja a primeira vez, o arquivo é comparado com o já existente

2.1.2.1 caso seja igual é apagado e o log de que não há novos arquivos é persistido

2.1.2.2 caso seja diferente o algoritmo vai para o passo 3

3 O arquivo é aberto em *Streaming*

4 Cada arquivo *XML* contido é aberto

4.1 É feita a decodificação do arquivo e enviado para a inserção

4.1.1 Antes da inserção é verificado se aquela sequência linear existe

5 Após gerar os objetos computacionais relativos a cada epítopo é feita uma lista de distinção única por identificador das sequências lineares

6 Os objetos computacionais com sequências lineares novas são inseridos na base de dados

Obs1.: A decisão de abrir o arquivo por s*treaming* foi tomada a fim de evitar alocar e desalocar memória e aumentar o desempenho da aplicação.

Obs2.: A verificação realizada neste passo é feita na lista de objetos que ainda está sendo criada, não nos dados já existentes no banco de dados.

Obs3.: Este passo gerou um dos problemas de custo de tempo para o algoritmo, uma vez que a quantidade de dados inseridos é grande, apesar de os registros em si serem estruturalmente pequenos. Então, houve a necessidade da implementação da inserção de múltiplos epítopos na base de dados (criação de blocos de informação) de uma única vez para tornar o algoritmo mais performático.

Sendo assim, ao fim do algoritmo e com o seu uso recorrente, sempre haverá no HATsDB a lista de sequências lineares de epítopos mais atualizadas de acordo com a base de dados especializada do IEDB.

4.4 MAPEAMENTO DE EPÍTOPOS

Dada a importância dos epítopos no processo de criação de vacinas (JONES, 2015) e de novos tratamentos, é necessário fazer o processo de mapeamento para identificar quais são os epítopos codificados no genoma do VHC. Este processo é feito a partir do mapa de características já criado, simultaneamente com o conjunto de dados de epítopos no HATsDB obtido do *IEDB* (Fluxograma 6).

Para a realização do mapeamento foi então feito o mapeamento local das cadeias de aminoácidos dos epítopos nas informações de tradução das sequências (que são obtidas com as sequências direto do *NCBI*) utilizando um algoritmo de janela deslizante para que assim seja possível definir frequência dos epítopos em cada grupo genotípico do VHC e também de forma global para instigar investigações para bons candidatos para vacinas (que podem inclusive ser globalmente eficazes a depender do quão frequente o epítopo for em todos os genótipos do organismo) e fármacos.

O processo de mapeamento é computacionalmente complexo e extremamente custoso, afinal é um processo que se dá pela ordem de 2^n(o aumento é exponencial) e há no banco atualmente cerca de 243 mil sequências de VHC e cerca de 1,275 milhões de epítopos. O algoritmo de mapeamento utilizado se baseia em computação paralela e

assíncrona para realizar a tarefa no menor tempo possível. O algoritmo se dá nos seguintes passos:

1 Coleta dos epítopos no HATsDB

2 Coleta das *features* (características anotadas) no HATsDB

3 Criado a *pool* de tarefas de mapeamento

4 Registro de máquinas servidoras e processadores disponíveis para a tarefa

5 *Pool* de tarefas é iniciado

6 Definido o tamanho do pacote de tarefas

7 (Re)Distribuição os pacotes

8 Após o processamento verifica se há mais tarefas

8.1 Caso haja volta à tarefa 7

8.2 Caso não haja vai à tarefa 9

9 Armazena resultado do mapeamento

Obs1.: O processo do mapeamento de epítopos apesar de computacionalmente custoso apresenta uma peculiaridade: a tarefa unitária de um mapeamento não custa tempo em excesso, mas a quantidade de vezes que é feita a faz gastar mais tempo, entretanto, o maior gasto de tempo está no processo de armazenamento dos dados processados. Neste caso, temos um processo de 243.000 * 1.370.000, o que resulta em cerca de 95 milhões de ocorrências unitárias e inseri-los, no banco de dados causava um atraso enorme por consequência de uma trava de banco de dados (chamada de *Table-Lock* para garantir que somente uma transação ocorra e assim evitar colisão de dados) o que gerou um atraso de i/o (entrada e saída). O primeiro mapeamento completo tinha menos sequências e muito menos epítopos (204 mil e 780 mil respectivamente) levou cerca de 16 dias. As soluções implementadas pela nossa equipe chamadas de *MemoryMap* e *LargeInsert* fizeram todo o processamento do mapeamento de epítopos ser diminuído para algum tempo por volta de 2 horas de processamento.

O resultado da correlação destes dados gera informações de origens cruzadas, como, por exemplo, a lista de países que foram submetidas as sequências para as quais

um epítopo foi mapeado.

4.5 DESENVOLVIMENTO DO *FRONT-END*

Feitos os processos de coleta, análise, processamento e armazenagem das informações que dizem respeito ao escopo desta ferramenta deve haver uma maneira de disponibilizar os resultados publicamente para que desta forma outras equipes de pesquisadores tenham acesso a estes dados para identificar assim tratamentos e vacinas para o VHC. Para este fim, foi acordada a criação de um *front-end* desenvolvido para a w*eb* que se acople com o *back-end* e sirva os respectivos dados de maneira simples, conforme está exemplificado no Fluxograma 7. Assim, fica sob responsabilidade do *front-end* de apresentar o resultado das análises e dados principais, como também disponibilizar de maneira pública todo o material para *download* (como o arquivo .sql para a replicação do banco de dados).

Este *front-end* deve servir ainda como um portal informativo contendo os resultados a cada vez que as análises são refeitas.

4.6 VALIDAÇÃO DOS DADOS

Verificar a precisão do algoritmo é extremamente importante, e para possibilitar foram obtidos dados de 79314 sequências do banco de dados do Los Alamos, e após filtrar referências duplicadas, remover fragmentos de sequência e subtipos que não estão no HATsDB obtivemos dados de subtipagem de 29950 sequências únicas.

Para a checagem dos dados foi desenvolvido um algoritmo em Python que gera um relatório em csv utilizando a *accession version* da sequência como identificador e preenche o subtipo que foi definido pelo LANL e o subtipo que foi definido para o HATsDB.

Em poucos casos o LANL não identificou o subtipo específico, apenas o genótipo, e para garantir a validação de genotipagem o algoritmo avaliou se o HATsDB classifica da mesma forma.

Com o arquivo contendo os dados de verificação é possível então verificar a taxa de acerto, e então os dados de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Assim então pode-se obter a acurácia, sensibilidade e especificidade, além do valor preditivo positivo, valor preditivo negativo, a razão de probabilidades.

5 RESULTADOS

5.1 DADOS

O banco de dados biológicos resultante do nosso software contém dados relevantes de VHC que tangem os dados primários e especializados. A base de dados contém 243.169 sequências nucleotídicas com seus respectivos mapeamentos e subtipagem e 1.370.597 epítopos, dos quais 13.133 são epítopos específicos descritos, pelo IEDB, em VHC.

A distribuição da quantidade de subtipos mapeados por epítopos únicos e classificação dos subtipos é mostrada nas Tabelas Suplementares 1 e 2. O número total esperado de epítopos (com base no número de epítopos do IEDB) mapeados para o VHC era de 13.133, no entanto, 12.955 epítopos distintos foram observados após a análise de mapeamento.

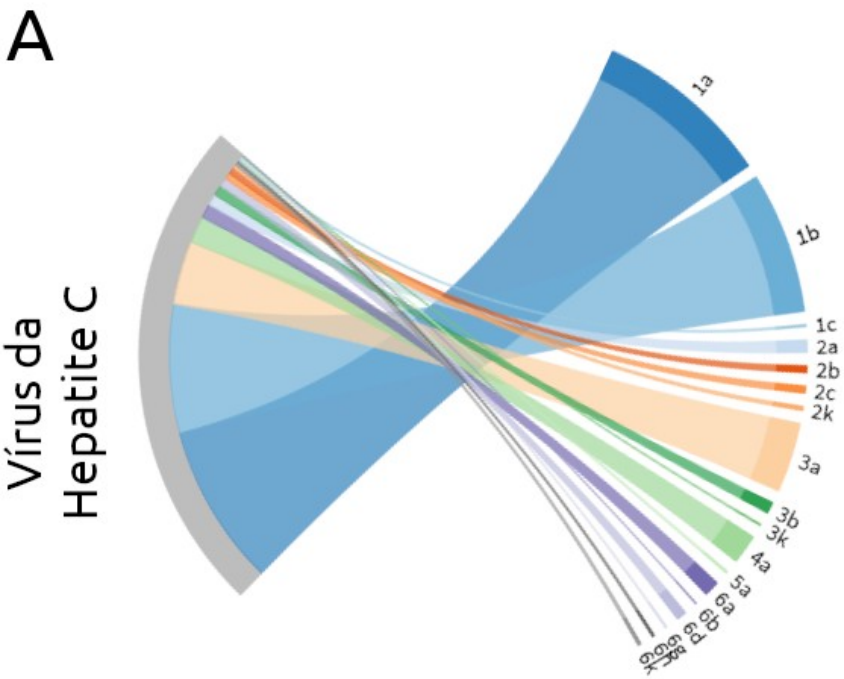
A análise das sequências e epítopos de VHC disponíveis até janeiro de 2020 gerou um total de 75.205.624 alinhamentos e processos de subtipagem de sequência (ou seja, um processo para cada sequência). A descrição dos dados gerados a partir desta análise é compilada na Tabela Suplementar 3 e contém um breve resumo do banco de dados, dos dados primários (quantidade de sequências e epítopos) dos processos de mapeamento.

O número de epítopos na Tabela Suplementar 3 reflete o número total de epítopos contidos no banco de dados IEDB, contendo epítopos hipotéticos e todos os dados adicionais pertencentes a cada epítopo. Os epítopos de VHC marcados pelo *IEDB* representaram cerca de 0,958% (13.133) de todos os epítopos no banco de dados. Os epítopos mapeados representam cerca de 0,945% (12.955) da quantidade total.

Os dados na Tabela 3 foram extraídos do número total de sequências e representam a alta taxa de mutação do VHC. A figura 4A mostra as frequências dos subtipos indicando a distribuição geral dos subtipos do VHC.

Inicialmente, assumindo que esses dados seguem tendências globais, o genótipo 1 do VHC deve ser o mais frequentemente encontrado (73,6%), seguido pelos genótipos 3 (12,0%), 2 (5,5%), 6 (4,2%) e 4 (3,6%). A validação do processo de subtipagem mostrou um alto nível de precisão (*AUC* 0,997), com sensibilidade e especificidade robustas

(sensibilidade de 0,963 e especificidade de 0,999). A taxa positiva verdadeira (*TPR*) foi de 0,997 e a taxa de falso positivo foi de 0,0001. Esses resultados de precisão estão resumidos na curva *ROC* mostrada na Figura 4B.



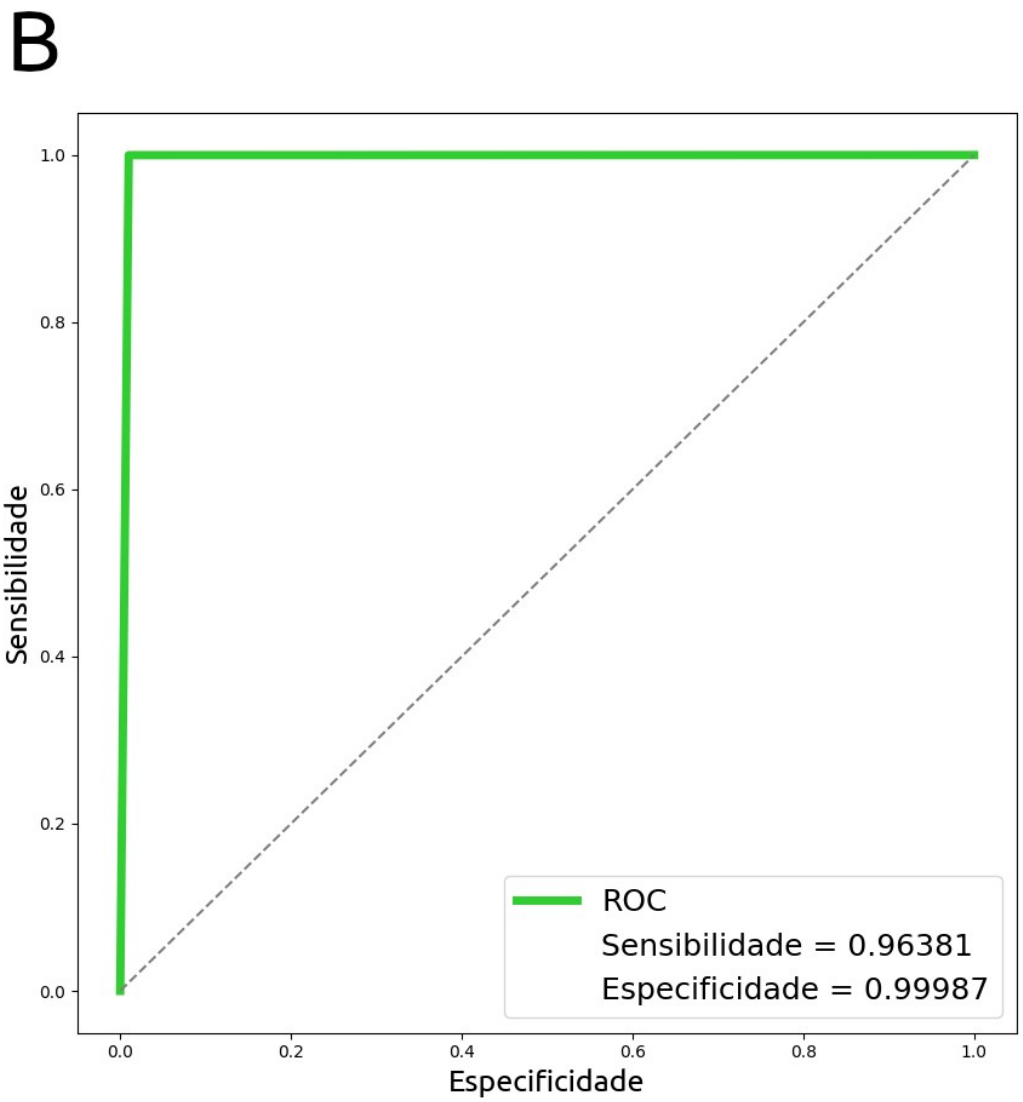


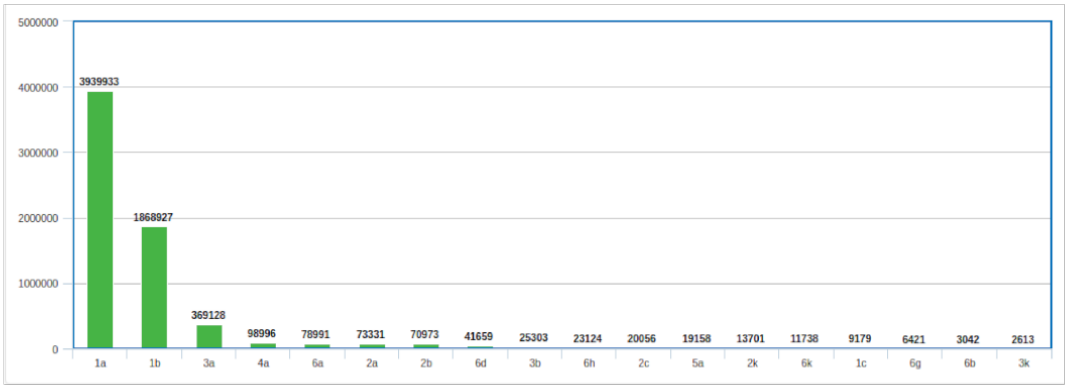
Figura 4 - A: Distribuição dos subtipos de VHC na classificação feita pelo HATsDB. B: Curva ROC demonstrando o desempenho na classificação de sequências em subtipos.

Utilizando os dados de sequência subtipadas, os epítopos mapeados foram agrupados de acordo com o genótipo do VHC, o que pode (ou não) refletir um viés introduzido pela maior submissão de sequências de alguns subtipos. A Figura 5A ilustra os dados de epítopo mapeados de acordo com o subtipo de VHC.

Depois disso, um mapa de calor interativo foi criado para representar dinamicamente o número de epítopos mapeados para cada subtipo de VHC, bem como as regiões e frequências correspondentes. A Figura 5B mostra uma captura de tela do *front-end* das sequências lineares do epítopo e informações relacionadas (ocorrências

totais, ocorrências de subtipos, um mapa de calor representando cada subtipo individual, número de países que relatam uma sequência específica de epítopos e abundância global de sequências) com um filtro mostrando os epítopos com maior número de ocorrências e que tenham tido ocorrências nos 18 subtipos.

A



B

Linear Sequence	Total Hits	Subtype Hits	1a	1b	1c	2a	2b	2c	2k	3a	3b	3k	4a	5a	6a	6b	6d	6g	6h	6k	Countries	Frequency in sequences (%)
QLINTNGSWHIN	349744	18																			96	28.41
YLVAYQATV	300274	18																			111	11.94
KTSERSQPR	228051	18																			149	20.30
HSKKKCDEL	210185	18																			101	10.70
LPGCSFSIF	141342	18																			113	14.87
QLINTNGS	123162	18																			104	33.55
IPFYGKAI	101520	18																			90	12.69
TYSTYGKFL	92065	18																			103	14.67

Figura 5 - A: Quantidade de ocorrências de epítopos por subtipo de VHC no HATsDB, o eixo y representa a quantidade de ocorrências e o eixo x representa os diferentes subtipos. B: *HeatMap* do *front-end* do HATsDB exemplificando os dados de mapeamento de epítopos.

No total, 235 dos epítopos foram mapeados para todos os 18 subtipos, enquanto outros 129 epítopos relevantes foram mapeados para pelo menos 17 dos 18 subtipos. Todos esses epítopos não são hipotéticos, ou seja, possuem uma sequência linear proteica que pode ser relevante para uso em estudos farmacológicos projetados para desenvolver uma vacina globalmente eficaz contra o vírus da hepatite C. As tabelas 2 e 3 representam um sumário agregado dos dados do banco.

Tabela 2 - Número de sequências submetidas, ocorrências de epítopos, correspondências únicas e total de epítopos exclusivos ao longo do tempo (2010-2019).

Ano	Total de sequências	Total de correspondências de epítopos	Total de correspondências únicas	Total de epítopos mapeados
2010	91.928	13.200.000	4.160.114	*3.014**
2015	164.473	42.600.001	14.757.840	*3.124**
2019	243.169	75.205.624	27.265.995	3.138

*no momento deste estudo, os dados disponibilizados do IEDB não possuíam informações temporais sobre a data de envio do epítopo; portanto, era impossível determinar com precisão o número de epítopos mapeados com sucesso para as sequências em 2010 e 2015

Tabela 3 - Número de ocorrências de epítopos (representados em valor absoluto e em %) de acordo com o subtipo de VHC ao longo do tempo (2010-2019)

Subtipo	2010	2015	2019
1a	2.131.656 - 52,53%	8.259.329 - 56,60%	16.890.179 - 62,86%
1b	1.492.966 - 36,79%	4.356.281 - 29,85%	6.500.419 - 24,19%
1c	8.607 - 0,21%	25.964 - 0,17%	37.879 - 0,14%

2a	22.839 - 0,56%	18.626 - 0,12%	300.188 - 1,11%
2b	23.631 - 0,58%	195.127 - 1,33%	25.497 - 0,09%
2c	1.133 - 0,02%	51.235 - 0,35%	66.357 - 0,24%
2k	7.763 - 0,19%	44.775 - 0,30%	74.928 - 0,27%
3a	190.075 - 4,68%	721.082 - 4,94%	1.509.956 - 5,62%
3b	16.476 - 0,40%	71.604 - 0,49%	18.982 - 0,07%
3k	2.528 - 0,06%	7.999 - 0,05%	8.724 - 0,03%
4a	7.826 - 0,19%	271.872 - 1,86%	548.889 - 2,04%
5a	26.908 - 0,66%	49.789 - 0,34%	72.701 - 0,27%
6a	45.586 - 1,12%	252.421 - 1,73%	385.726 - 1,43%
6b	4.558 - 0,11%	7.955 - 0,05%	11.356 - 0,04%
6d	44.519 - 1,09%	132.922 - 0,91%	212.102 - 0,78%
6g	5.713 - 0,14%	20.475 - 0,14%	31.146 - 0,11%
6h	2.424 - 0,05%	59.904 - 0,41%	95.711 - 0,35%

6k	22.459 - 0,55%	42.846 - 0,29%	74.944 - 0,27%

5.1.1 **Alinhamento e Subtipagem de Sequências**

O processo de alinhamento das sequências foi feito de forma exaustiva (se utilizando do método ótimo) que gerou no alinhamento global o resultado que pode ser visto no *front-end* (Tabela suplementar 1) onde é demonstrado em forma de tabela a quantidade de correspondências em cada um dos sítios da sequência (tanto os sítios estruturais quanto as não estruturais). O método ótimo utilizado, corresponde a um modelo exaustivo (não-heurístico), ou seja, garante que houve testes de todas as possibilidades existentes, o que faz a complexidade ter a grandeza de ordem *n* (aumenta de acordo com o aumento da quantidade de sequências). Modelos heurísticos podem diminuir o tempo e carga de processamento, porém, causando diminuição da acurácia. Logo, para garantir o melhor resultado possível sem chances de erros, todos os processos do HATsDB foram realizados de forma exaustiva.

O alinhamento mostra que no HATsDB existe quantitativamente muito mais sequências estruturais E1 e E2 que dos demais sítios. Entre os sítios Core, NS3, NS5A e NS5B é menor que os citados anteriormente, mas ainda superiores aos demais (Figura suplementar 1). Após o alinhamento, o processo de subtipagem foi realizado e o quantitativo pode ser visto na tabela suplementar 2.

No *front-end* desenvolvido há um gráfico com aceleração em *GPU* (utilizando *OpenGL* implementando *ThreeWebGL* através da biblioteca *WebGL Globe* desenvolvido pelo *Google Data Arts Team*) que possibilita ver a distribuição estatística das sequências de VHC ao redor do mundo, uma por vez como pode ser visto nas figuras suplementares 2 e 3, onde cada ponto no mapa representa uma cidade com pelo menos 10.000 habitantes, e a altura das barras representa a distribuição naquele país (que é expressa em percentagem do total no mundo).

5.1.2 **Mapeamento de Epítopos**

O mapeamento de epítopos teve um total de 289.967.640.312 de instâncias de alinhamentos locais, onde houveram cerca de 95 milhões de ocorrências dos epítopos em sequências de VHC.

Na tabela suplementar 3 extraída da base de dados que mostra a relação entre a quantidade de subtipos de VHC e a quantidade de sequências lineares únicas que mapearam naquela quantidade de subtipos.

De modo a consolidar uma análise mais fidedigna e facilitada em relação ao conjunto inteiro do mapeamento de epítopos, foi feita uma análise agrupando os dados, onde foram condensadas 95 milhões de registros pela sequência linear do epítopo, assim os dados foram agrupados em 3.120 registros. Desta forma foi possível criar uma interface que seja mais rápida e eficiente para visualização destes dados, como pode ser visto nas figuras suplementares 4 e 5.

No *front-end* foi implementado uma visualização de dados onde ao clicar na sequência linear do epítopo um gráfico é mostrado contendo os subtipos que aquele epítopo foi mapeado e a submissão das sequências relativas a este mapeamento por ano, com a possibilidade de filtrar subtipos e verificar a quantidade de ocorrências por ano. Uma demonstração deste gráfico pode ser visto nas figuras suplementares 6 e 7.

5.2 PLATAFORMA DIGITAL

Conforme proposto no escopo do projeto, foi feito um *front-end* que funciona como portal para informações e contém uma interface gráfica informativa que condensa os dados importados e processados.

Este *front-end* foi construído utilizando a *stack* web baseada nas seguintes tecnologias: *MySQL* na versão 5.7 para o banco de dados, PHP na versão 7.2.18 como linguagem de programação para a criação do *back-end* do *front-end* e para o processamento, *HTML* 5 como linguagem de marcação para a construção das páginas *WEB, CSS* na versão 3 para a estilização da página (que neste caso não é de caráter unicamente estético, é de extrema importância a estilização para que haja a precisão na interpretação dos dados e na visualização) e a linguagem *JavaScript* sob a versão 6, convencionado pela *ECMA* para as funções *client-side* para a criação das *interfaces* e demais utilidades.

Os resultados do desenvolvimento do *front-end* descrito ao longo deste texto pode ser acompanhado no portal online (http://pah.bahia.fiocruz.br/hat), segue em sequência nas figuras suplementares as imagens que exemplificam todo o trabalho feito.

5.3 *BENCHMARK* E VALIDAÇÃO DO PROCESSO

Após executar os processos de mapeamento e alinhamento das sequências e desenvolver o *front-end*, comparamos o desempenho da ferramenta HATsDB e analisamos os dados gerados.

Nossa ferramenta se mostrou precisa no processo de subtipagem quando foi feita a validação da análise. O processo de validação envolvendo mais de 29.000 sequências do LANL reforça a noção de que o algoritmo de subtipagem é fortemente preciso, permitindo, portanto, a classificação assertiva dentro de uma vasta quantidade de dados biológicos publicamente disponíveis. Esta alta precisão e alta velocidade de processamento permitiram a análise de todos os dados relacionados à VHC nos bancos de dados do *NCBI*. A tabela 4 mostra os resultados estatísticos resultantes da validação.

Tabela 4 - Dados da análise estatística da validação dos dados

Fonte: Autor.

Análise	Valor
Verdadeiro positivo	29883
verdadeiro negativo	508011
Falso positivo	66
Falso negativo	1122
Sensibilidade	0.9638
Especificidade	0.9999
Acurácia	0.9978
Valor preditivo positivo	0.9978
Valor preditivo negativo	0.9978
Taxa de verdadeiros positivos	0.9978
Taxa de falsos positivos	0.0001
Razão de probabilidade	0.0097

Precisão	0.9978
Revocação	0.9978

Esses resultados combinados com outras informações recuperadas do NCBI, como data de amostragem, local da amostra e fonte de isolamento fornece uma estimativa confiável de distribuição de genótipos em todo o mundo. Além disso, a distribuição geográfica observada em nosso banco de dados é semelhante ao descrito anteriormente na literatura (CHAN et al., 2014 e JONES, 2015). Essa semelhança em distribuição sugere que, apesar do viés intrínseco do uso de um banco de dados, esses dados poderiam refletir a frequência real do genótipo do VHC com precisão razoável. A genotipagem, informações geográficas e de tempo, com a recursividade do sistema, podem ainda fornecer vigilância epidemiológica molecular de genótipos de VHC nos países. Assim, a ferramenta poderia fornecer as frequências dos genótipos com demarcações temporais e identificar mudanças na distribuição ao longo dos anos ou mesmo a introdução de um novo genótipo, como relatamos com o ZIKV no Brasil (KASPRZYKOWSKI et al., 2020).

A resposta imunológica das células TCD8+ desempenha um papel importante no controle da infecção pelo VHC (PARK e REHERMANN, 2020). A eficiência destas células no controle de infecções a torna um importante alvo para o desenvolvimento de vacinas (SWADLING et al., 2014). No entanto, a maioria dos estudos com essas células têm como alvo o genótipo 1, negligenciando os demais (LUXENBURGER et al., 2020). Usando os epítopos IEDB mapeados nas sequências de VHC do NCBI, a ferramenta HATsDB fornece informações de todos os epítopos em todos os genótipos com informações de região geográfica. Como o banco de dados IEDB possui epítopos de uma variedade de patógenos, a ferramenta preenche a falta de informações sobre o epítopo de outros genótipos. Além disso, as informações do epítopo contribuirão para um possível desenvolvimento de vacinas e entenderão os diferentes padrões de doenças que podem ser lideradas por alterações na resposta imune CD8 +.

6 DISCUSSÃO

6.1 SOBRE O TRABALHO

6.1.1 **A discrepância entre número de sequências por subtipo**

Após o alinhamento, o processo de subtipagem que foi realizado evidenciou que existem subtipos com muito mais sequências que outros, como pode ser observado na tabela suplementar 2. Vista essa discrepância entre a quantidade de sequências por subtipo, aqui se abre uma discussão do porquê isso ocorre, uma vez que é possível que seja um dado real ou um viés de informação. Por conta disso, foram levantadas as seguintes hipóteses que ainda estão sendo discutidas:

- Os subtipos podem, de fato, serem os mais frequentes em todo o mundo.
- Os subtipos podem ser mais frequentes nos países que mais submetem sequências, como pode ser exemplificado na tabela suplementar 4 e nas figuras suplementares 2 e 3, o que não representa o valor global real.

Apenas com a informação de submissão que há nas sequências não é possível, no escopo deste trabalho, fazer uma checagem mais profunda de forma a tentar identificar o porquê isso ocorre, no entanto, outros estudos na literatura (MESSINA et al., 2015 e PETRUZZIELLO et al., 2016) já fizeram análises sobre essa distribuição e prevalência, indicando que há de maneira global uma presença maior destes genótipos e subtipos.

6.1.2 **Epítopos**

No que diz respeito aos epítopos, em uma primeira filtragem procurando por epítopos descritos com um rótulo para VHC, foram encontrados 3120 epítopos, no entanto, após o processo de mapeamento, observou-se que a quantidade de epítopos mapeados era maior que o esperado (12.955 epítopos), despertando curiosidade. Após algumas revisões, foi possível checar que o IEDB organiza de maneira descentralizada um único organismo, o que causou uma confusão. Após listar todos os organismos que tinham em seu registro o rótulo de “*HCV*” ou “*hepatitis C*” foi possível verificar 359 identificadores de organismos diferentes, entretanto, todos se referem ao VHC seja como acrônimos ou mesmo separando linhagens. Com estes identificadores diferentes do VHC em mãos, filtrando novamente todos os epítopos e estratificando-os, foi possível, então, verificar a existência de 13.133 epítopos. Portanto, o resultado deste processo de mapeamento ficou dentro do que era esperado, o número de epítopos descritos ficou próximo do que foi, de fato, mapeado.

6.2 LIMITAÇÕES

6.2.1 **O 7º genótipo**

O presente estudo apresentou algumas limitações em seu desenvolvimento, dentre elas, pode-se citar o sétimo genótipo existente de Hepatite C, que não foi utilizado apesar de estar presente na literatura. Isso se deu por não ter um quantitativo de sequências suficientes para alterar o resultado. Sendo assim, não é relevante refazer a análise adicionando o genótipo 7, visto que há registros de apenas 3 sequências, de genoma completo, de versões de acesso **KU861171.1**, **EF108306.2** e **NC_030791.1**, identificadas na literatura e no banco de dados *nucleotide* do NCBI.

6.2.2 **O congelamento do banco de dados**

Outra limitação presente neste estudo foi a manutenção ativa do banco de dados. Focando em analisar e consolidar os dados e melhorar as interfaces, foi decidido que o banco de dados do HATsDB deveria ser congelado em 2019, de forma que as alterações a nível de software pudessem ser feitas e os resultados fossem checados para evitar inconsistências.

Com este congelamento, apesar de recente, não é possível dizer que o banco de dados representa atualmente os dados globais em tempo real, sendo possível somente após todas as correções e implementações de software e interfaces descongelar o banco de dados e voltar a atualizar para que as informações sejam sempre as mais recentes.

6.2.3 **A “qualidade das sequências”**

O HATsDB tem em seu repositório as sequências que são enviadas ao banco de dados mundial, em que é assumido que estas tenham passado por um processo de curadoria e checagem da sua leitura e importação. Os algoritmos que compõem o HATsDB tentam então lidar com zonas de inserções e deleções para gerar uma pontuação de similaridade e verificar as regiões de maior similaridade para aquela sequência. Sendo assim, o HATsDB não faz uma verificação da qualidade das sequências importadas de maneira direta, ainda assim, o tamanho destas sequências e suas pontuações são indicadores que podem ser considerados para linhas de pesquisas futuras que queiram estratificar estes dados.

7 CONCLUSÃO

A base desenvolvida contém toda a informação das bases de dados primárias mundiais disponíveis, com a adição de informações relevantes sobre o HCV, como subtipo, localização genômica entre outras. Além disso, também está contido na base as informações de epítopos, relevantes para o desenvolvimento de vacinas e testes diagnósticos. Essa base especializada no HCV mais atualizada no momento.

A *interface* de usuário amigável permite acesso rápido a qualquer informação no HATsDB. Com os dados compilados, o usuário pode executar uma série temporal de genótipos de VHC, epítopos, ou ambos. Além disso, inferir a frequência de um epítopo específico em uma região geográfica ou uma fonte de isolamento de amostra. Toda essa informação combinada permite a consulta para geração de várias hipóteses sobre a infecção pelo VHC.

Em síntese a versão final do HATsDB permite aos usuários analisar com eficiência epítopos de VHC e dados de submissão de sequência. O esperado é que esta ferramenta e os dados (tanto os gerados quanto os adquiridos) possam ser úteis para estudos imunológicos e possam contribuir para o desenvolvimento de uma vacina globalmente eficaz contra o VHC. O banco de dados de sequências gerado por meio dessa ferramenta continuará sendo atualizado e disponível para acesso de toda a comunidade científica através de nosso *front-end* projetado.

De modo a consolidar e compartilhar internacionalmente os recursos, resultados e possibilidades do HATsDB, um artigo científico foi redigido em conjunto com todos os colaboradores do projeto e submetido em junho de 2020 no periódico PLOS *COMPUTATIONAL BIOLOGY*, e em setembro de 2020 continua sob revisão. O arquivo gerado pela submissão está anexado a esta dissertação.

8 REFERÊNCIAS

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

COMBE, Marine; SANJUÁN, Rafael. Variation in RNA Virus Mutation Rates across Host Cells. **Plos Pathogens**, v. 10, n. 1, p. e1003855, 23 jan. 2014. Public Library of Science (PLoS). Disponível em: http://dx.doi.org/10.1371/journal.ppat.1003855.

CHAKRABORTY, Angana; BANDYOPADHYAY, Sanghamitra. FOGSAA: Fast Optimal Global Sequence Alignment Algorithm. **Scientific Reports**, v. 3, n. 1, 29 abr. 2013. Semanal. Springer Science and Business Media LLC. Disponível em: http://dx.doi.org/10.1038/srep01746.

DESHMUKH, K. B., & KHARAT, M. U. Review on Retrieving Biological Sequence Alignment using Smith-Waterman Algorithm. **International Journal Of Innovative Research In Computer Science & Technology (Ijircst)**. P. 24-26. 7 jan. 2015. Disponível em https://www.ijircst.org/DOC/6_review_on_retrieving_biological_sequence_alignment_using_smith-waterman_algorithm.pdf

FARCI, P. *et al.* The Outcome of Acute Hepatitis C Predicted by the Evolution of the Viral Quasispecies. **Science**, American Association for the Advancement of Science (AAAS), v. 288, n. 5464, p. 339 - 344, 14 abr. 2000. ISSN 1095-9203. Disponível em: http://dx.doi.org/10.1126/science.288.5464.339.

FERREIRA, Cristina Targa; SILVEIRA, Themis Reverbel da. Hepatites virais: aspectos da epidemiologia e da prevenção. **Rev. Bras. Epidemiol.**, São Paulo , v. 7, n. 4, p. 473-487, dez. 2004 . Disponível em http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-790X2004000400010&lng=pt&nrm=iso.

GOWER, E. *et al.* Global epidemiology and genotype distribution of the hepatitis C virus infection. **Journal of Hepatology**, v. 61, n. 1, p. S45 - S57, 30 de Julho de 2014. Disponível em: http://dx.doi.org/10.1016/j.jhep.2014.07.027.

HANAFIAH, K. M. *et al*. Global epidemiology of hepatitis C virus infection: New estimates of age-specific antibody to HCV seroprevalence. **Hepatology**. AASLD, v. 57, n.4, p. 1333 - 1342, Abril 2013. Disponível em: http://dx.doi.org/10.1002/hep.26141.

HE, L.; ZHU, J. Computational tools for epitope vaccine design and evaluation. **ScienceDirect.** ELSEVIER, v. 11, p. 103 - 112. Abril, 2015. Disponível em: http://dx.doi.org/10.1016/j.coviro.2015.03.013.

JONES, L. H.. Recent advances in the molecular design of synthetic vaccines. **Nature Chemistry**, v. 7, n. 12, p. 952-960, 20 nov. 2015. Springer Science and Business Media LLC. Disponível em: http://dx.doi.org/10.1038/nchem.2396.

KASPRZYKOWSKI, J. I. *et al*. A recursive sub-typing screening surveillance system detects the appearance of the ZIKV African lineage in Brazil: Is there a risk of a new epidemic?. **International Journal Of Infectious Diseases**, v. 96, p. 579-581, jun. 2020. Elsevier BV. Disponível em: http://dx.doi.org/10.1016/j.ijid.2020.05.090. Acesso em: 20 de junho de 2020.

KING, A. M. *et al.* Family - Flaviviridae. In: KING, A. M. et al. (Ed.). **Virus Taxonomy**, Elsevier, 2012. p. 1003 - 1020. ISBN 9780123846846. Disponível em: http://www.sciencedirect.com/science/article/pii/B9780123846846000860

LAWITZ, E. *et al*. Simeprevir plus sofosbuvir, with or without ribavirin, to treat chronic infection with hepatitis C virus genotype 1 in non-responders to pegylated interferon and ribavirin and treatment-naive patients: the COSMOS randomised study. **The Lancet**, Elsevier, v. 384, n.9956, p. 1756 - 1765, Novembro 2014. ISSN 0140-6736. Disponível em: http://dx.doi.org/10.1016/s0140-6736(14)61036-9.

LUXENBURGER, H. *et al*. Differential virus-specific CD8+ T-cell epitope repertoire in hepatitis C virus genotype 1 versus 4. **Journal Of Viral Hepatitis**, v. 25, n. 7, p. 779-790, 27 fevereiro. 2018. Wiley. http://dx.doi.org/10.1111/jvh.12874. Acesso: 20 de junho. 2020.

LYONS, M. S. *et al.* Prevalence of Diagnosed and Undiagnosed Hepatitis C in a Midwestern Urban Emergency Department. **Clinical Infectious Diseases**, Oxford Academic. v. 62, n. 9, p. 1066-1071, 21 de Fevereiro 2016. Disponível em: http://doi.org/10.1093/cid/ciw073.

MCENTYRE, Jo. Linking up with entrez. **Trends In Genetics**, v. 14, n. 1, p. 39-40, jan. 1998. Elsevier BV. http://dx.doi.org/10.1016/s0168-9525(97)01325-5.

MESSINA, J.P. *et al.* Global distribution and prevalence of hepatitis C virus genotypes. **Hepatology**, AASLD, v. 61, n. 1, p. 77 - 87, Julho 2014. Disponível em: https://doi.org/10.1002/hep.27259 . Acesso: 20 jun. 2020.

NCBI. **Entrez Programming Utilities Help - NCBI Help Manual**. 2010. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK25501/. Acesso em: 27 de ago. 2016.

NCBI, **Hepatitis C virus genotype 1, complete genome**. Disponível em: https://www.ncbi.nlm.nih.gov/nuccore/NC_004102.1 . Acesso em: Junho de 2017.

NEEDLEMAN, S. B.; WUNSCH, C D.. A general method applicable to the search for

similarities in the amino acid sequence of two proteins. **Journal Of Molecular Biology**, Elsevier BV, v. 48, n. 3, p. 443-453, Março 1970. Disponível em: http://dx.doi.org/10.1016/0022-2836(70)90057-4.

PARK, Su-Hyung; REHERMANN, Barbara. Immune Responses to HCV and Other Hepatitis Viruses. **Immunity,** Elsevier BV. v. 40, n. 1, p. 13-24, Janeiro 2014. Disponível em: https://doi.org/10.1016/j.immuni.2013.12.010. Acesso: 20 jun. 2020.

PETRUZZIELLO, A. *et al.* Global epidemiology of hepatitis C virus infection: An up-date of the distribution and circulation of hepatitis C virus genotypes. **World Journal of Gastroenterology**, Copyright, v. 22. n. 34, p. 7824–7840. Setembro de 2016. Disponível em: https://doi.org/10.3748/wjg.v22.i34.7824. Acesso: 20 jun. 2020.

PRATI, D. *et al*. Influence of Different Hepatitis C Virus Genotypes on the Course of Asymptomatic Hepatitis C Virus Infection. **Gastroenterology**, Elsevier BV, v. 110, n. 1, p. 178-183, Janeiro. Disponível em: http://dx.doi.org/10.1053/gast.1996.v110.pm8536854.

SAYERS, Eric. **The E-utilities In-Depth: Parameters, Syntax and More**. 2009 May 29 [Updated 2014 Feb 14]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Disponível em: http://www.ncbi.nlm.nih.gov/books/NBK25499/.

SIMMONDS, P. *et al*. A proposed system for the nomenclature of hepatitis C viral genotypes. **Hepatology,** Baltimore, Md., v. 19, n. 5, p. 1321–1324. Maio de 1994. Disponível em: https://pubmed.ncbi.nlm.nih.gov/8175159/.

SIMMONDS, P. *et al.* Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. **Journal Of General Virology**, Society for General Microbiology, v. 74, n. 11, p. 2391-2399, Novembro de 1993. Disponível em: http://dx.doi.org/10.1099/0022-1317-74-11-2391
.

SMITH, Temple F. *et al*. Identification of common molecular subsequences. **Journal of molecular biology**, v. 147, n. 1, p. 195-197, 1981. Disponível em: https://doi.org/10.1016/0022-2836(81)90087-5.

STRAUSS, Edna. Hepatite C. **Rev. Soc. Bras. Med. Trop**., Uberaba , v. 34, n. 1, p. 69-82, fev. 2001. Disponível em http://dx.doi.org/10.1590/S0037-86822001000100011. Acesso em 01 agosto. 2020.

SWADLING, L*. et al.* A human vaccine strategy based on chimpanzee adenoviral and MVA vectors that primes, boosts, and sustains functional HCV-specific T cell memory. **Science Translational Medicine**, American Association for the Advancement of Science (AAAS), v. 6, n. 261, p. 153-261, 5 de Novembro de 2014. Disponível em: http://dx.doi.org/10.1126/scitranslmed.3009185. Acesso em: 20 jun. 2020.

VITA, R. *et al.* The immune epitope database (IEDB) 3.0. **Nucleic Acids Research**,

Oxford Academic, v. 43, n. 1, p. 405-412, 9 de Outubro de 2014. Disponível em: http://dx.doi.org/10.1093/nar/gku938.

WORLD HEALTH ORGANIZATION (org.). **Hepatitis C**. 27 July 2020. Disponível em: https://www.who.int/en/news-room/fact-sheets/detail/hepatitis-c#.Vw8RlQyFvio.mendeley. Acesso em: 01 ago. 2020.

WORLD HEALTH ORGANIZATION. **Global hepatitis report 2017**. World Health Organization. Disponível em: https://apps.who.int/iris/handle/10665/255016. ISBN: 9789241565455 . Acesso em: 5 janeiro 2021.

ZOU, D. *et al*. Biological Databases for Human Research. **Genomics, Proteomics & Bioinformatics**, Elsevier BV. v. 13, n. 1, p. 55-63, Fevereiro de 2015. Disponível em: http://dx.doi.org/10.1016/j.gpb.2015.01.006.

Material suplementar

Tabela suplementar 1 - Número de ocorrências de epítopos diferentes por quantidade de subtipos.

Fonte: Autor

Número de subtipos	Número de ocorrências
1	547
2	489
3	301
4	300
5	256
6	140
7	100
8	72
9	105
10	78
11	31
12	56
13	74
14	68
15	62
16	70
17	132

18	239

Tabela suplementar 2 - Número de sequências por subtipo no HATsDB.

Fonte: Autor

Subtipo	Número de sequências
1a	85.719 - 35.51%
1b	89.867 - 37.59%
1c	1.429 - 0.59%
2a	6.533 - 2.79%
2b	3.223 - 1.32%
2c	2.196 - 0.9%
2k	1.213 - 0.5%
3a	25.283 - 10.46%
3b	3.368 - 1.4%
3k	373 - 0.15%
4a	8.863 - 3.69%
5a	1.979 - 0.81%
6a	4.662 - 1.95%
6b	285 - 0.11%

6d	1.735 - 1.26%
6g	444 - 0.19%
6h	737 - 0.3%
6k	957 - 0.4%

Tabela suplementar 3 - Estratificação de dados do HATsDB

Fonte: Autor

Dado	valor
Quantidade total de epítopos	1.370.597
Epítopos marcados como sendo de VHC pelo IEDB	13.133
Quantidade de sequências	243169
Quantidade de ocorrências de mapeamentos de epítopos	75.205.624
Quantidade de epítopos únicos mapeados	12.955
Quantidades de sequências com ocorrência de ao menos 1 epítopo	209.269

Tabela suplementar 4 - Número de sequências por país dos 10 países com maior quantidade de submissões.

Fonte: Autor

País	Número de sequências
USA	36565 - 15.18%
Spain	24888 - 10.23%

China	18166 - 7.50%
Canada	14985 - 6.16%
France	10748 - 4.42%
Italy	9347 - 3.84%
Brazil	5575 - 2.29%
United Kingdom	5020 - 2.06%
Argentina	4924 - 2.02%
Ireland	4907 - 2.01%

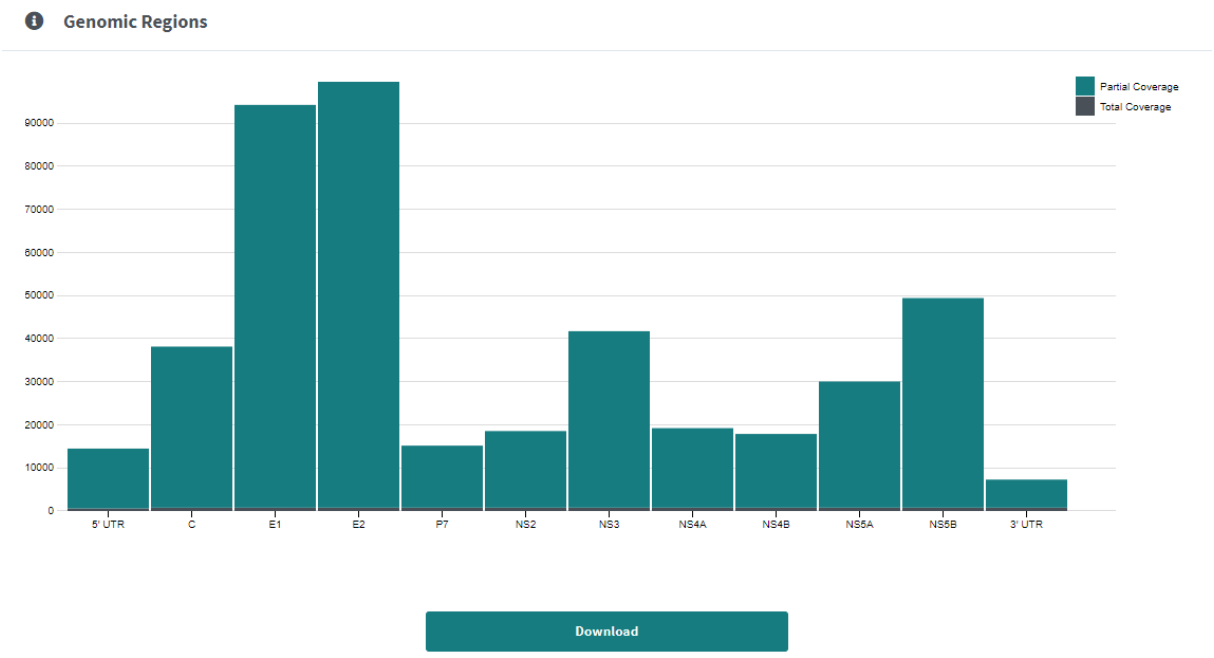


Figura suplementar 1 - Regiões genômicas das sequências. Fonte: Autor

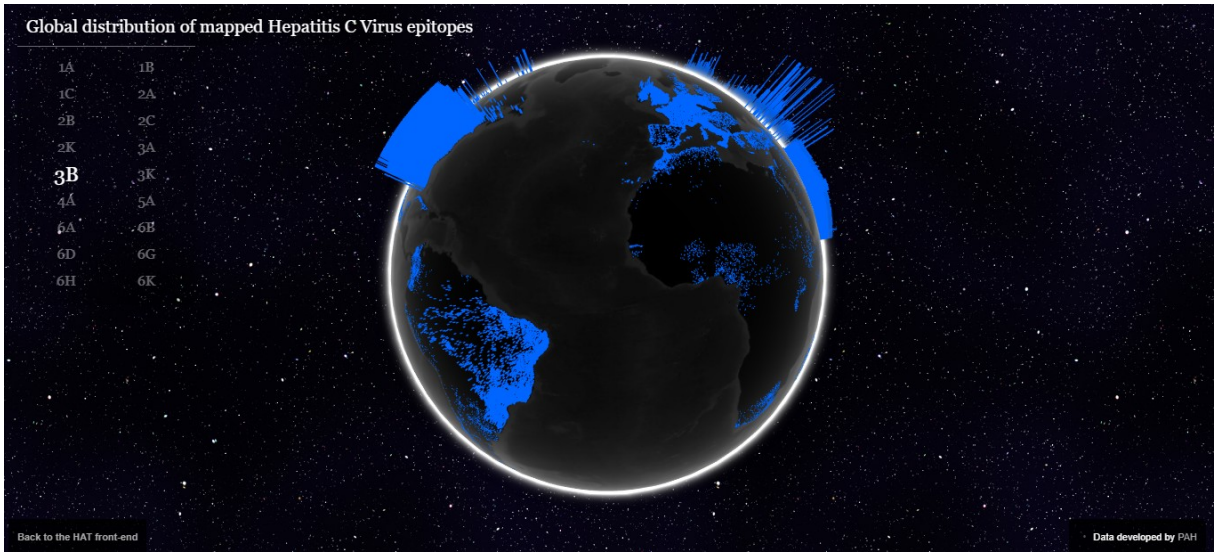


Figura suplementar 2 - Distribuição global das sequências mapeadas nos epítopos classificadas por subtipo. Fonte: Autor

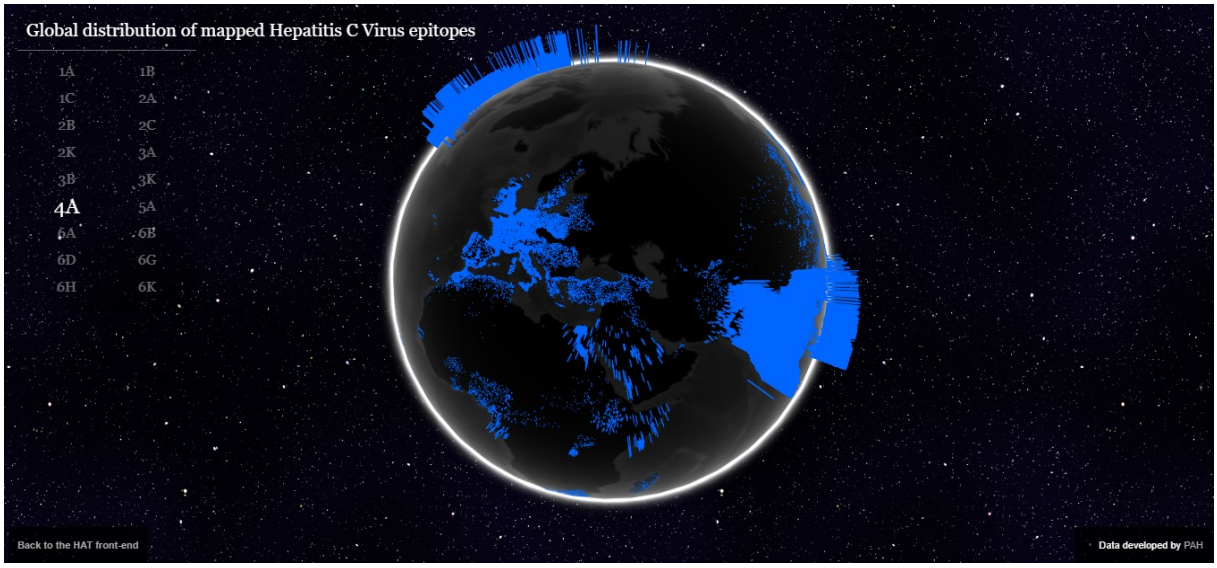
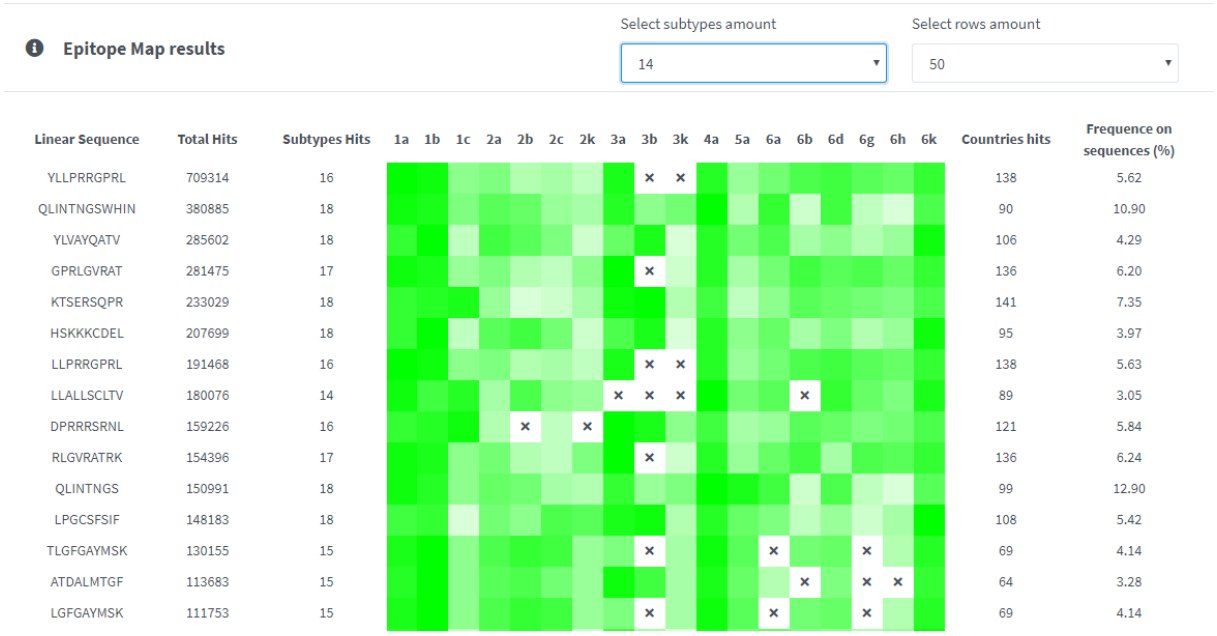


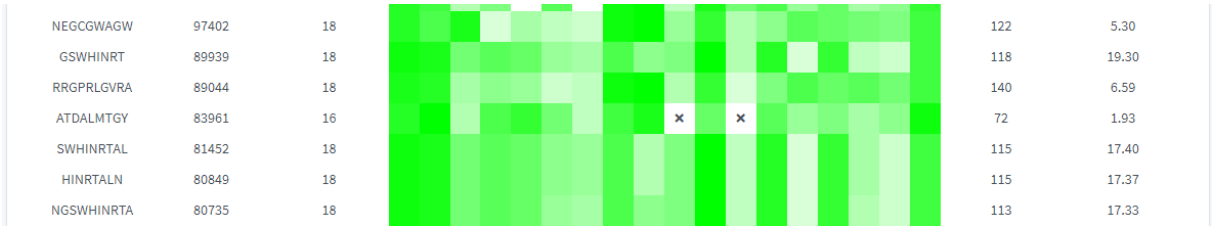
Figura suplementar 3 - Distribuição global das sequências mapeadas nos epítopos classificadas por subtipo. Fonte: Autor



Linear Sequence	Total Hits	Subtypes Hits	Countries hits	Frequence on sequences (%)
YLLPRRGPRL	709314	16	138	5.62
QLINTNGSWHIN	380885	18	90	10.90
YLVAYQATV	285602	18	106	4.29
GPRLGVRAT	281475	17	136	6.20
KTSERSQPR	233029	18	141	7.35
HSKKKCDEL	207699	18	95	3.97
LLPRRGPRL	191468	16	138	5.63
LLALLSCLTV	180076	14	89	3.05
DPRRRSRNL	159226	16	121	5.84
RLGVRATRK	154396	17	136	6.24
QLINTNGS	150991	18	99	12.90
LPGCSFSIF	148183	18	108	5.42
TLGFGAYMSK	130155	15	69	4.14
ATDALMTGF	113683	15	64	3.28
LGFGAYMSK	111753	15	69	4.14

Figura suplementar 4 - Mapeamento de epítopos. Fonte: Autor

Obs: os "x" representados no heatMap representam que não houve correspondência deste epítopo no subtipo representado na coluna.



NEGCGWAGW	97402	18	122	5.30
GSWHINRT	89939	18	118	19.30
RRGPRLGVRA	89044	18	140	6.59
ATDALMTGY	83961	16	72	1.93
SWHINRTAL	81452	18	115	17.40
HINRTALN	80849	18	115	17.37
NGSWHINRTA	80735	18	113	17.33

Figura suplementar 5 - Mapeamento de epítopos (foco nos epítopos com maior frequência no total de sequências). Fonte: Autor

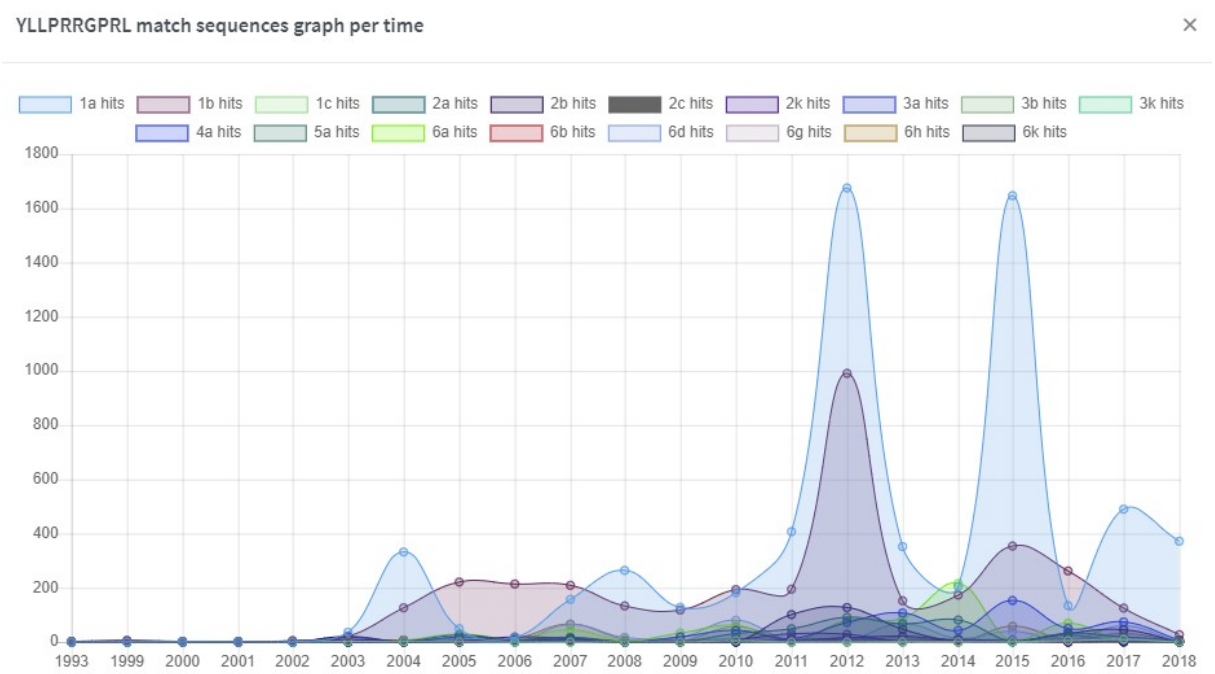


Figura suplementar 6 - Gráfico mostra a evolução da submissão de sequências nas quais o epítopo foi mapeado ao longo dos anos. O eixo Y representa a quantidade de ocorrências em sequências submetidas naquele ano, o eixo x representa o ano de submissão e cada cor representa um subtipo. Fonte: Autor

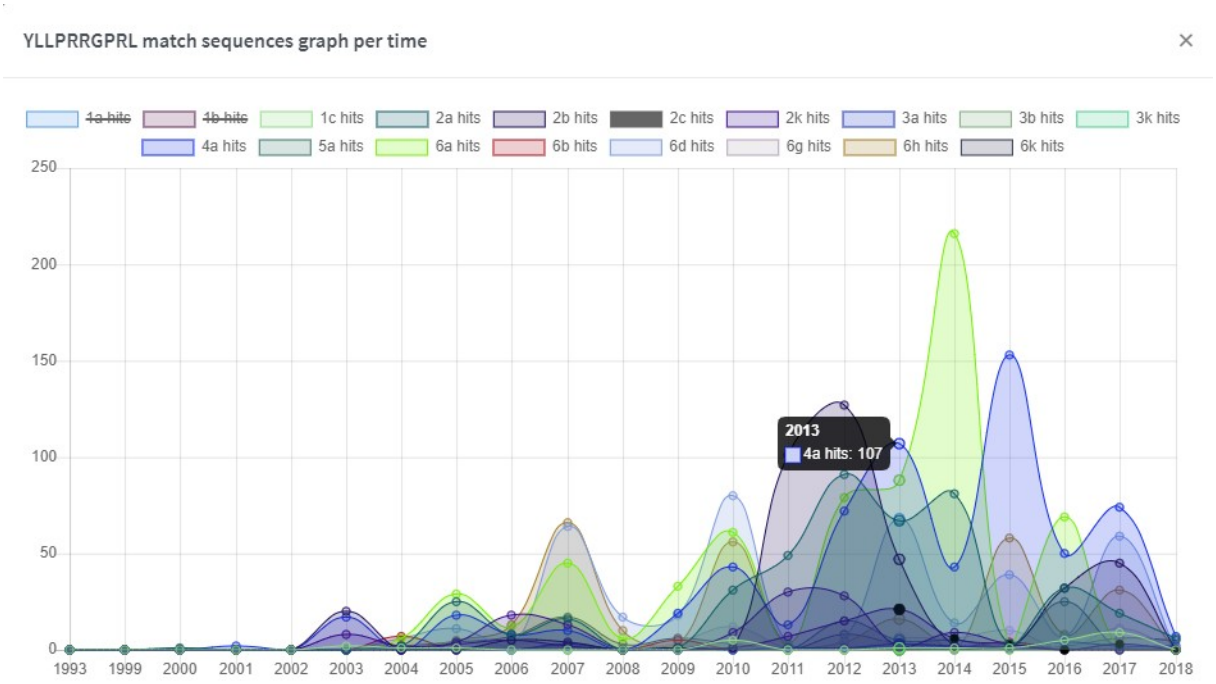
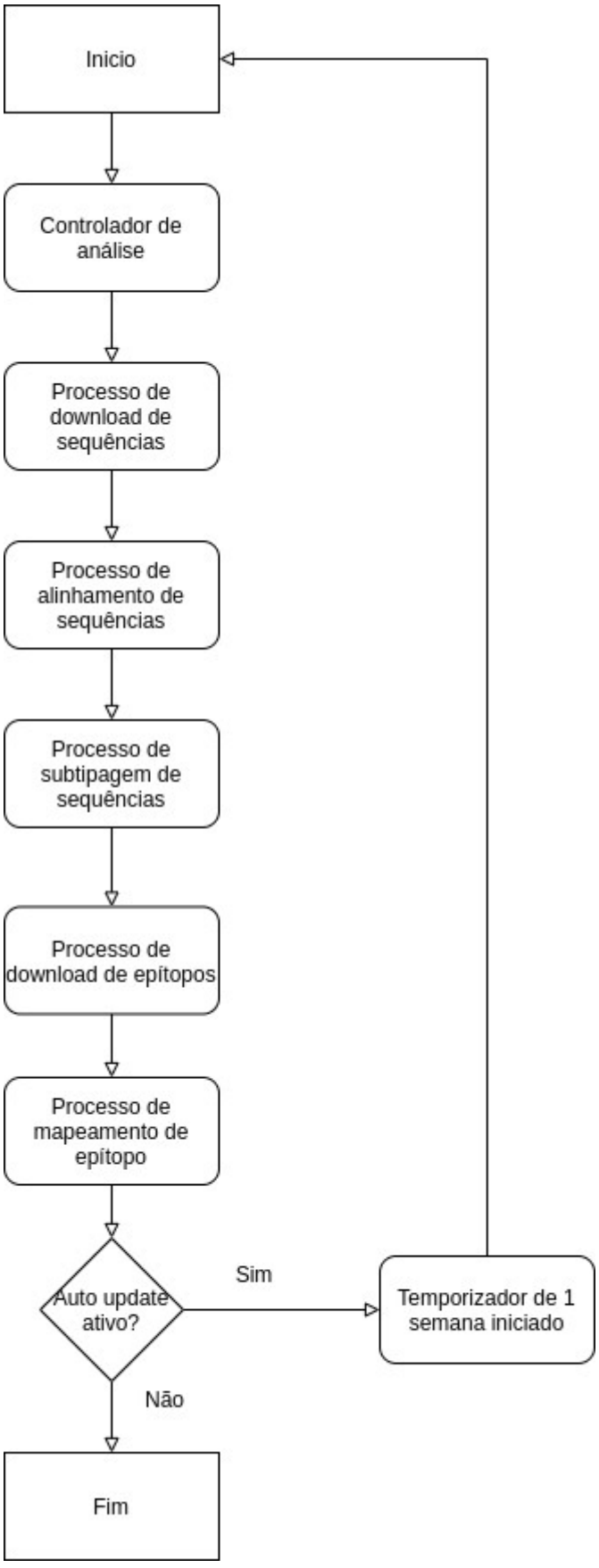
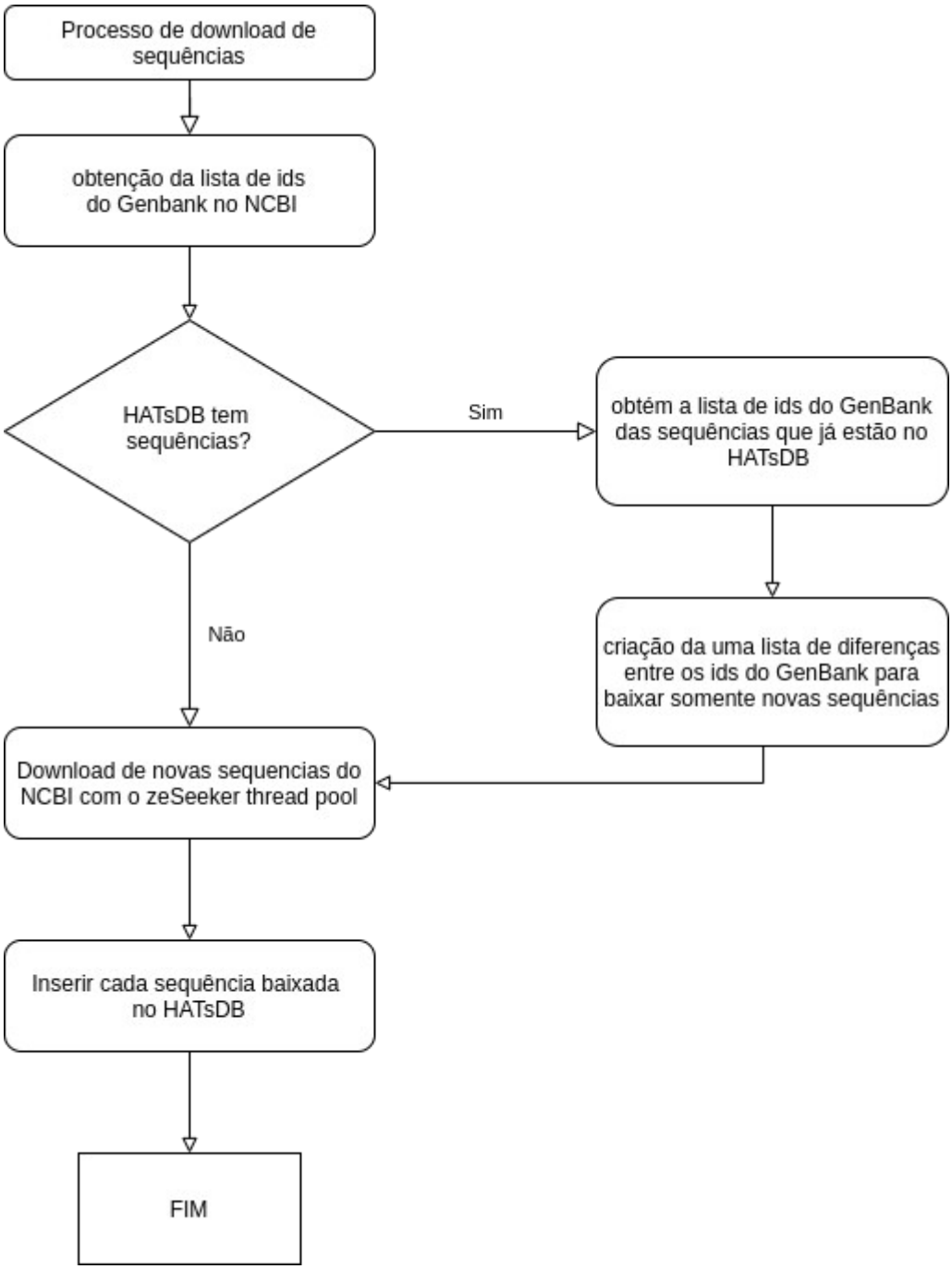


Figura suplementar 7 - Filtragem de subtipos e contagem por subtipo por ano. A figura representa o mesmo que a figura suplementar 6, porém exemplifica a filtragem de subtipos e exibe a quantidade de ocorrências por subtipo em um dado ano (o dado é apresentado ao passar o *mouse* sobre o círculo)

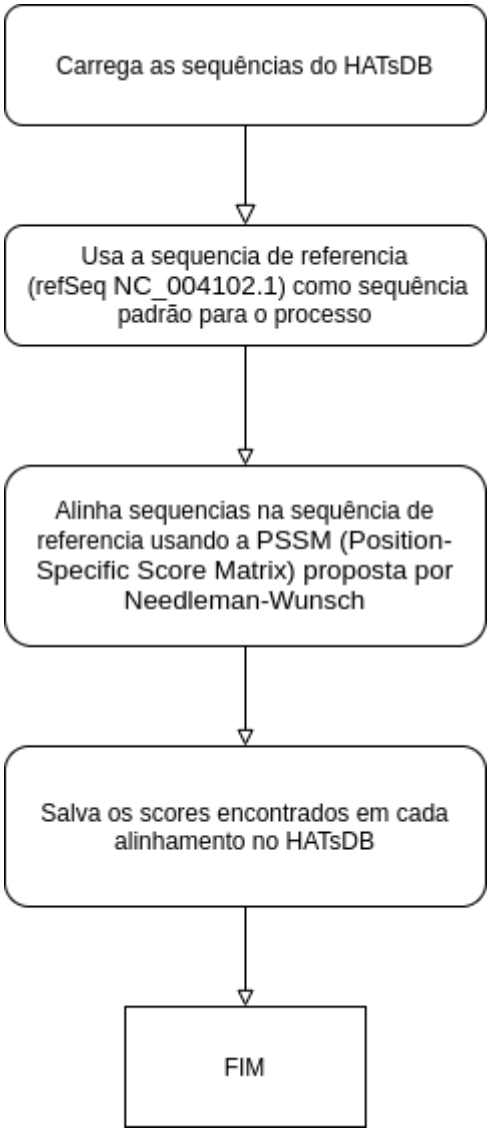
Fonte: Autor


Inicio
Controlador de análise
Processo de download de sequências
Processo de alinhamento de sequências
Processo de subtipagem de sequências
Processo de download de epítopos
Processo de mapeamento de epítopo
Auto update ativo?
Sim
Não
Temporizador de 1 semana iniciado
Fim

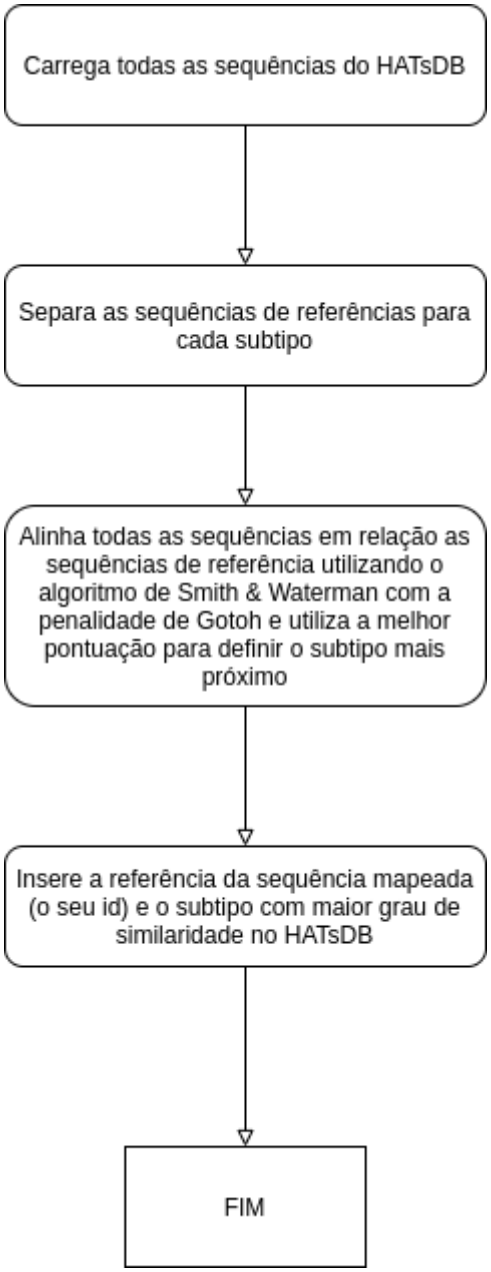
Fluxograma 1 - Fluxo da aplicação geral, Fonte: Autor



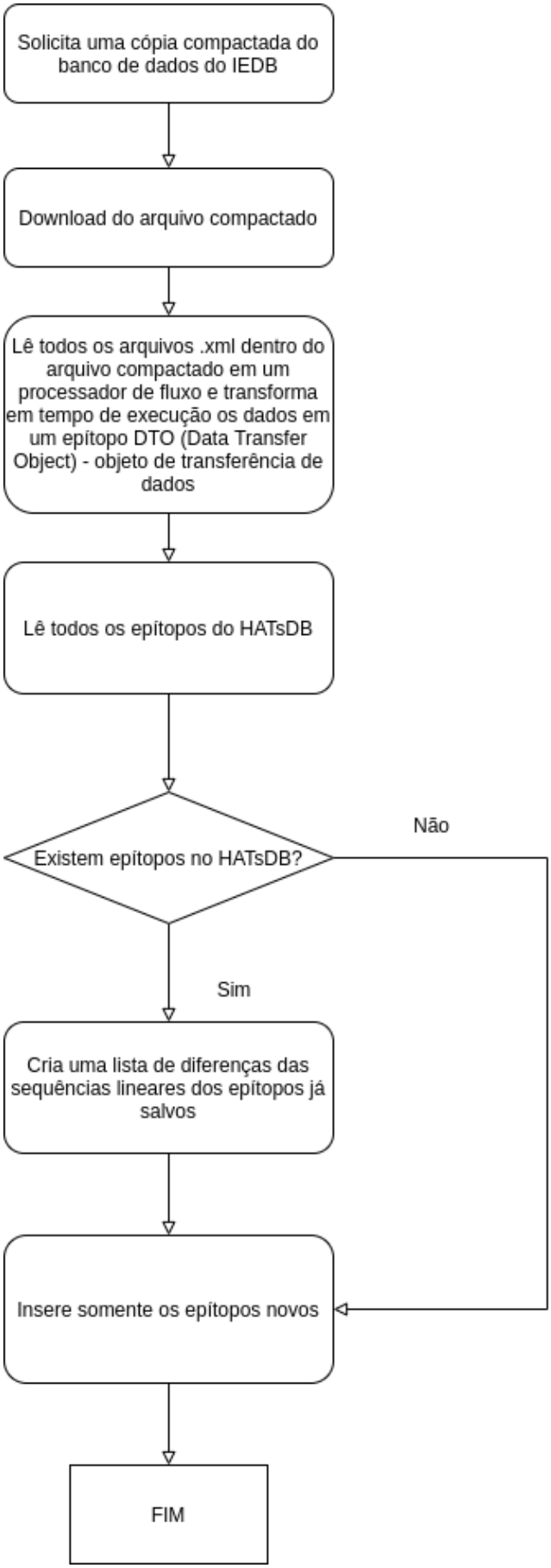
Fluxograma 2 - Download de sequências, Fonte: Autor



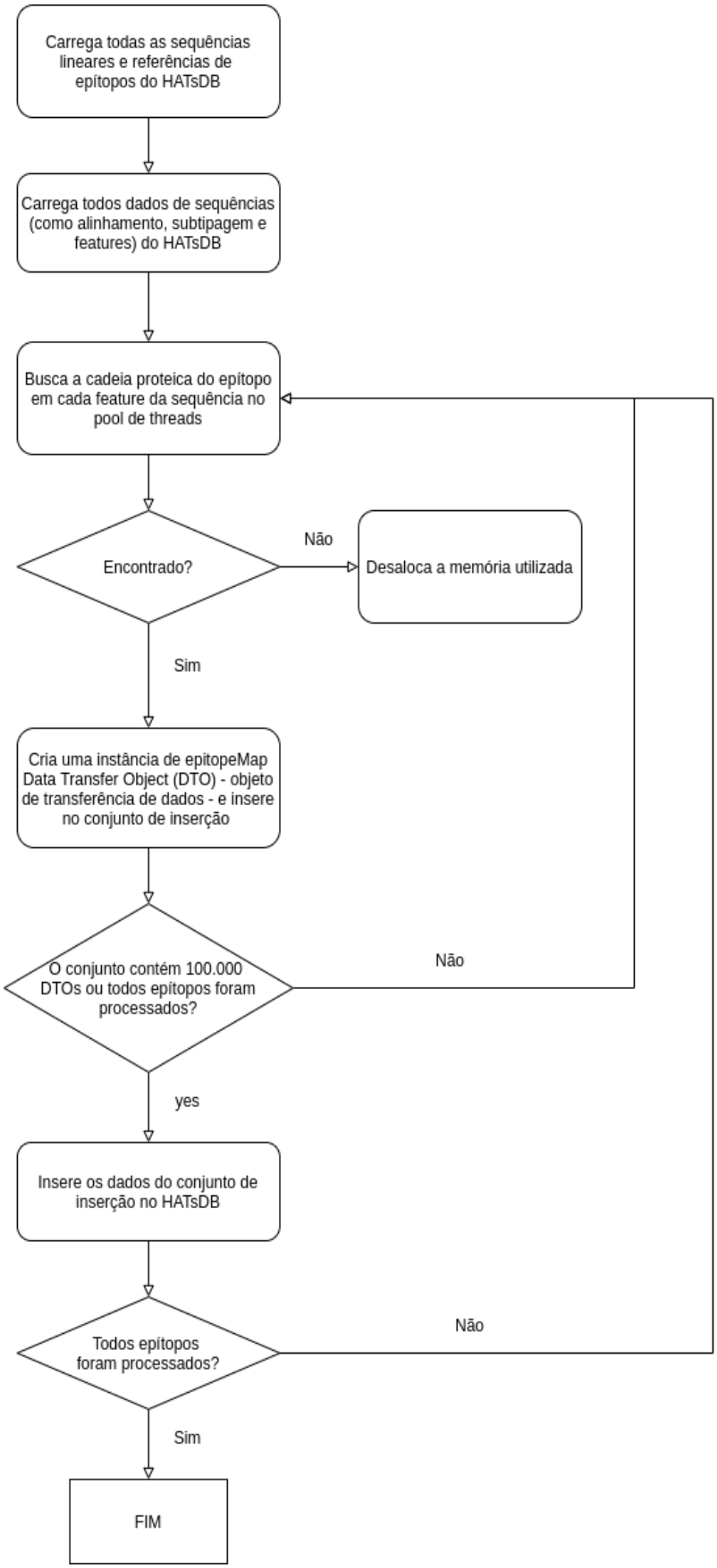
Fluxograma 3 - Mapeamento de sequências, Fonte: Autor



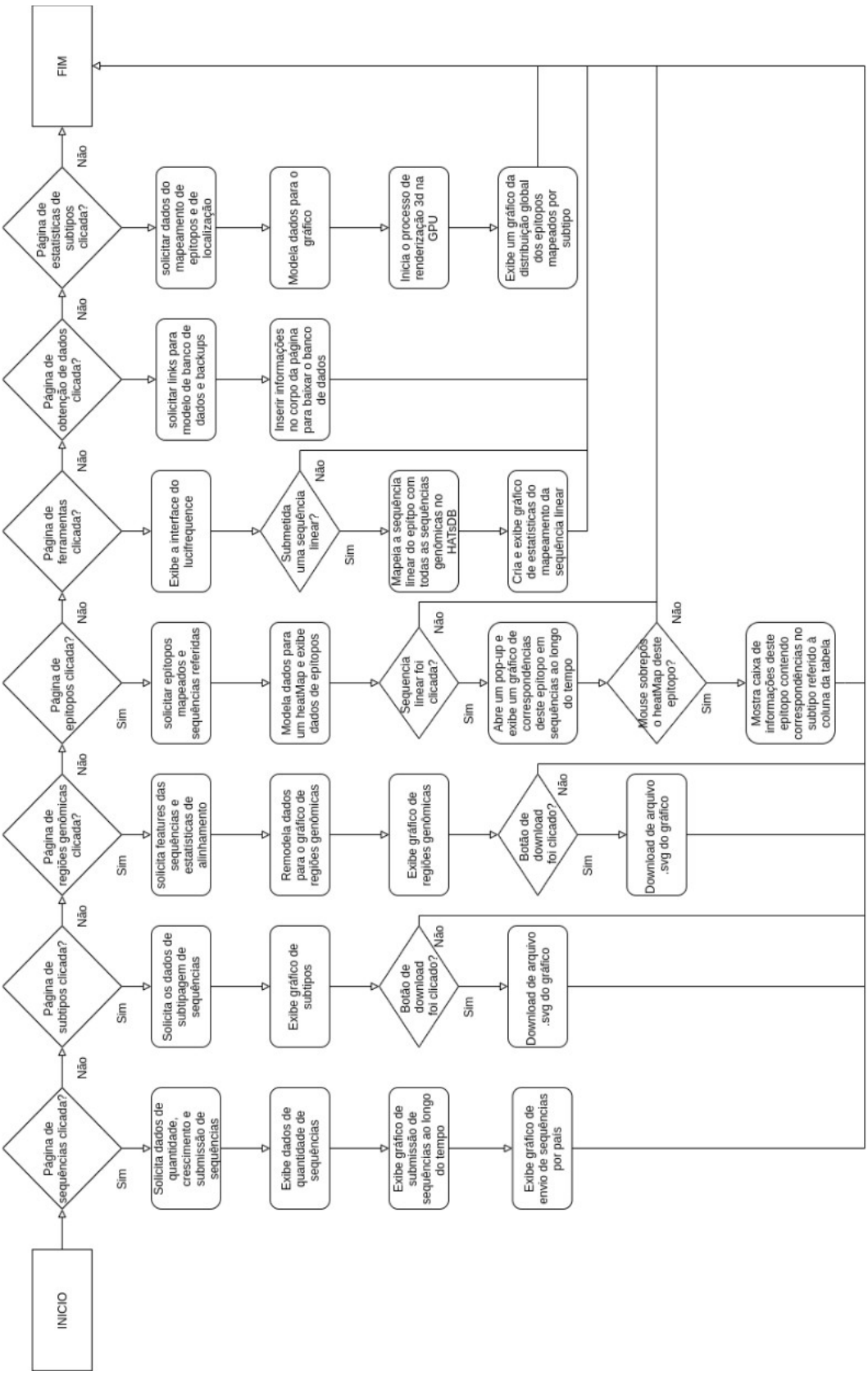
Fluxograma 4 - Subtipagem de sequências, Fonte: Autor


Solicita uma cópia compactada do banco de dados do IEDB
Download do arquivo compactado
Lê todos os arquivos .xml dentro do arquivo compactado em um processador de fluxo e transforma em tempo de execução os dados em um epítopo DTO (Data Transfer Object) - objeto de transferência de dados
Lê todos os epítopos do HATsDB
Existem epítopos no HATsDB?
Não
Sim
Cria uma lista de diferenças das sequências lineares dos epítopos já salvos
Insere somente os epítopos novos
FIM

Fluxograma 5 - *Download* de epítopos, Fonte: Autor



Fluxograma 6 - Mapeamento de epítopo, Fonte: Autor



Fluxograma 7 - Front-end da aplicação, Fonte: Autor