Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

Sistema para análise de sequências nucleotídicas do HIV disponíveis no GenBank

José Irahe Kasprzykowski Gonçalves

Feira de Santana

2016

Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

José Irahe Kasprzykowski Gonçalves

Sistema para análise de sequências nucleotídicas do HIV disponíveis no GenBank

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Orientador: Prof. Dr. Artur Trancoso Lopo de Queiroz

Feira de Santana

2016

José Irahe Kasprzykowski Gonçalves

Sistema para análise de sequências nucleotídicas do HIV disponíveis no GenBank

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Feira de Santana, 15 de dezembro de 2015

BANCA EXAMINADORA

Prof. Dr. Artur Trancoso Lopo de Queiróz (Orientador)
Centro de Pesquisas Gonçalo Moniz da Fundação Oswaldo Cruz

Prof. Dr. Iranderly Fernandes de Fernandes
Departamento de Física da Universidade Estadual de Feira de Santana

Prof. Dr. Filipe Ferreira de Almeida Rego
Departamento de Ciências Biológicas e da Saúde da Universidade Católica do Salvador

Ficha Catalográfica – Biblioteca Central Julieta Carteado

Gonçalves, José Irahe Kasprzykowski
G626s Sistema para análise de sequências nucleotídicas do HIV disponíveis no GenBank / José Irahe Kasprzykowski Gonçalves. – Feira de Santana, 2016.
59 f. : il.

Orientador: Artur Trancoso Lopo de Queiroz.

Dissertação (mestrado) – Universidade Estadual de Feira de Santana, Programa de Pós-Graduação em Computação Aplicada, 2016.

1. Computação – Genética. 2. HIV – Genoma – Banco de dados. I. Queiroz, Artur Trancoso Lopo de, orient. II. Universidade Estadual de Feira de Santana. III. Título.

CDU: 004.65:616.9

Abstract

HIV infects over 40 million people worldwide and is considered by the World Health Organization a large scale pandemic. Which the associated disease has no cure. New data and analysis can help new treatment and vaccine development. However, the dataset is vast, with over 500,000 sequences available on GenBank. This data still lacks essential information such as subtyping and genome location. To help minimize these problems we developed a system for automated analysis from GenBank data. The tool performs sequence map according to HXB2 and subtyping by comparison with subtype reference sequences. This process uses Needleman-Wusch and Smith-Waterman respectively. All 582,678 sequences were mapped in 5 days and 14 hours and subtyped in 1 day and 7 hours with our algorithm, while the original approach was estimated to finish in 36 and 97 years respectively. Our tool was able to analyse the massive data in a reliable time. No current subtyping tool can analyse this high-throughput data. Our results showed that pol and gag genes were the most prevalent genes on the dataset, and could be explained because treatment and subtyping are based on these genes. Moreover, the structural genes were most prevalent, with 66.41%. This highlighted the low representation of regulatory genes on available data. The subtyping results showed that the subtype B was most frequent, with 45.96%. The recombinants together represent 43.37%. Furthermore, subtype C presented only 4.12% and the other pure subtypes less than 4%. Also, the geographical data was recovered from database and USA presented higher frequency, with 24.50%, showing a significant country bias. Our results present a new HIV subtype distribution with the most complete and recent dataset.Herein, we presented a new user friendly software for massive data analysis of viruses. This software is able to analyse highly mutational virus data, such as HCV and HIV in reliable time. Further, severe country bias raises questions regarding world subtype distribution. The analysis of all sequences from HIV provides new epidemy insights about subtypes and country distribution.

Keywords: HIV, Nucleotide Sequenqces, Subtypes, Genotypes, Genetics

Resumo

O HIV infecta mais de 40 milhões de pessoas no mundo e é considerado pela Organização Mundial de Saúde como uma pandemia. A doença associada não possui cura clínica. Novas análises e informações podem ajudar no desenvolvimento de novos tratamentos e vacinas. No entanto, o conjunto de dados sobre o agente etiológico disponível é vasto, contando com mais de 500 mil sequências no GenBank. Este conjunto de dados ainda carece de informações essenciais, como subtipo viral e localização no genoma de referência. Para auxiliar na minimização destes problemas, desenvolvemos um sistema para análise dos dados disponíveis no GenBank. A ferramenta realiza o mapeamento de acordo com o genoma referencia HXB2 e a subtipagem comparando as sequências de referência dos subtipos. Estes processos utilizam os algoritmos de Needleman-Wusch e Smith-Waterman respectivamente. Todas as 582.678 sequências foram mapeadas em 5 dias e 14 horas, e subtipadas em 1 dia e 7 horas com nosso algoritmo. Enquanto a abordagem original estima terminar em 36 e 97 anos respectivamente. Nenhuma ferramenta de subtipagem disponível atualmente é capaz de analisar esta quantidade de dados. Nossos resultados mostraram que os genes *gag* e *pol* são mais prevalentes no conjunto de dados. O que pode ser explicado pelo fato de técnicas de avaliação de resistência aos antirretrovirais e subtipagem serem baseadas nesses genes. Além disso, os genes estruturais exibiram uma prevalência absoluta de 66.41%. Isto evidencia a pouca representatividade de genes regulatórios no conjunto de dados. Os resultados da subtipagem mostram que o subtipo B é o mais frequente com 45,96% de prevalência. Os recombinantes, combinados, representam 43.37%. Ademais, o subtipo C apresentou apenas 4,12% de prevalência absoluta e outros subtipos puros menos de 4%. Além disso, dados geográficos foram recuperados do banco de dados. Os Estados Unidos representam a maior frequência de sequências submetidas, com 24,5% de todos os dados disponíveis. Nossos resultados apresentam uma nova distribuição genotípica do HIV, com o conjunto de dados mais recente e completo. Neste trabalho apresentamos um novo software para análise das sequências nucleotídicas do HIV disponíveis no GenBank. Este software é capaz de analisar dados de vírus com elevado comportamento mutacional como HIV e HCV em um curto espaço de tempo. A análise de todas as sequências do HIV disponíveis no GenBank oferece um novo ponto de vista sobre a epidemia, distribuição de subtipos e geográfica.

Palavras-chave: HIV, Sequências Nucleotídicas, Subtipo, Genótipo, Genética.

Prefácio

Esta dissertação de mestrado foi submetida a Universidade Estadual de Feira de Santana (UEFS) como requisito parcial para obtenção do grau de Mestre em Computação Aplicada.

A dissertação foi desenvolvida dentro do Programa de Pós-Graduação em Computação Aplicada (PGCA) tendo como orientador o Dr. **Artur Trancoso Lopo de Queiroz**.

Agradecimentos

Agradeço inicialmente a Deus, meu amigo, companheiro e pai eterno. Este que sempre esteve do meu lado, nas noites sem dormir, nas viagens intermináveis e desafios intransponíveis. Gostaria ainda de agradecer à minha mãe, Márcia Elizabeth Kasprzykowski por ter me colocado no mundo, e me ensinado que um desafio só é desafio até que você o supere. E que é tudo uma questão de manter a mente quieta, a espinha ereta e o coração tranquilo.

Dedico este trabalho à minha família que me ensinou a ousar, a tentar, a persistir e a nunca desistir. Em especial, gostaria de dedicar à meu irmão, Mario Cauhe Kasprzykowski Gonçalves, meu melhor amigo, meu ídolo, meu porto seguro. À minha *Vokita* que todos os dias me ensina o significado de superação. Dedico ainda este trabalho à Ana Caroline Guimarães Silva, sem mais, o amor da minha vida.

Não posso deixar de mencionar meu agradecimento aos amigos que participaram ativamente durante toda essa jornada: Mateus Oliveira Malaquias, Felipe Guimarães Torres e Leonardo Melo. Obrigado pelos momentos de compreensão, companheirismo, perseverança e apoio.

Gostaria ainda de agradecer a todos os amigos que colaboraram de forma direta ou indireta para que este pudesse tomar forma. Dedico especialmente aos amigos da galera do "Rei do Patinete". Assim como aos meus colegas de trabalho e meus amigos da Van que sempre tem uma palavra de carinho e consolo.

Em especial, gostaria de agradecer ao meu orientador Prof. Dr. Artur Trancoso Lopo de Queiróz. Pela paciência, pelos ensinamentos, pela boa vontade de transformar um *"sambarilove"* num projeto apresentável. Obrigado por ser um PAI, e ajudar a construir o profissional, cientista, professor e ser humano que sou.

Sumário

Lista de Figuras

Lista de Abreviações

Abreviação	**Descrição**
AIDS	Síndrome da Imunodeficiência Adquirida (Acquired Immunodeficiency Syndrome)
CRF	Forma Recombinante Circulante (Circulating Recombinnant Form)
HCV	Vírus da Hepatite C (Hepatitis C Virus)
HIV	Vírus da Imunodeficiência Humana (Human Immunodeficiency virus)
HIV-1	Vírus da Imunodeficiência Humana tipo 1 (Human Imunodeficiency Virus type 1)
HMM	Modelo Oculto de Markov (Hidden Markov Models)
MHC	Complexo Principal de Histocompatibilidade (Main Histocompatibility Complex)
NCBI	National Center for Biotechnology Information
ORF	Quadro Aberto de Leitura(Open Read Frame)
SGBD	Sistema Gerenciador de Banco de Dados
SIV	Vírus da Imunodeficiência em Símios (Simian Immunodeficiency virus)
VSDBM	Gerenciados de Bases de Dados Virais (Viral Sequence Data Base Manager)
SOAP	Protocolo de Acesso a Objetos Simples (Simple Object Access Protocol)
URF	Forma Recombinante Única (Unique Recombinant Form)
XML	Linguagem de Marcação Extensível (eXtensible Markup Language)

Capítulo 1

Introdução

"Tudo é uma questão de manter a mente quieta, a espinha ereta e o coração tranquilo. "

– Walter Franco

O agente etiológico da síndrome da imunodeficiência adquirida, o HIV, infecta cerca de 2,3 milhões de pessoas por ano, num total de 40 milhões de pessoas infectadas no mundo. Já a síndrome, mata cerca de 2 milhões por ano [UNAIDS 2013]. Apesar da prevalência da patologia e do grande número de novas infecções, o tratamento para esta síndrome ainda apresenta desafios e não existe uma cura clínica. Grande parte do motivo desta dificuldade de tratamento se dá pela taxa mutacional e a consequente heterogeneidade genotípica associada. O HIV-1 apresenta uma taxa mutacional maior que os vírus e micro-organismos de DNA [Combe e Sanjuán 2014].

Essa característica adaptativa apresentada por este vírus, associada a mutações nas regiões alvo de fármacos conhecidas como "mutações de escape", reduzem as chances de descoberta de novas regiões alvo para intervenções terapêuticas [UK Colaborative Group on HIV Droug Resistance 2013]. Este comportamento mutacional facilita a geração de diversas quasiespécies do vírus, que por não sofrerem ação dos fármacos, se proliferam, e consequentemente criam uma heterogeneidade entre os genótipos do vírus [Li et al. 2015].

Esta heterogeneidade genotípica, aliada a elevada taxa reprodutiva, permite a rápida adaptação da população viral, o que facilita a ocorrência de mutações nas regiões reconhecidas pelo sistema imune (epítopos). Estas mutações geralmente causam o não pareamento dos epítopos com as moléculas do complexo principal de histocompatibilidade (MHC). Desta forma, o sistema imune falha em identificar o organismo invasor, o que causa a ausência relevante de resposta imune. Isso permite que o vírus seja poupado da influência do sistema imunológico e haja a persistência deste genótipo viral [Neher e Leitner 2010]).

Para amenizar estes problemas, diversos estudos sobre a estrutura genética do HIV vem sendo realizados para identificar informações que levem a formas eficazes de tratamento [Johannessen et al. 2011, Crous et al. 2012]. A identificação de regiões imunogênicas evolutivamente estáveis no HIV seria o ponto chave para o desenvolvimento de um tratamento eficaz. A existência de regiões desta natureza é altamente provável, se levadas em consideração as restrições funcionais e estruturais do genoma [Snoeck et al. 2011].

Para identificar tais áreas, é necessário compreender a diversidade genética dos organismos, suas características e comportamento [de Queiróz et al. 2011]. Sendo assim, os genomas são sequenciados e formatados em arquivos ou em bancos de dados biológicos, para posteriormente aplicar modelos de análise matemáticos e computacionais. Assim, serão geradas informações sobre as características genotípicas do conjunto de dados, e assim desenvolvidas intervenções mais eficazes para a doença [Chan et al. 2014, Crous et al. 2012, McGovern et al. 2010].

Na última década as tecnologias de sequenciamento vem crescendo exponencialmente para atender às demandas de compreensão genética. Partindo da identificação do genoma de referência humano, este desenvolvimento tecnológico possibilita o aumento na geração de dados sobre a estrutura genética. O crescimento no conjunto de dados, por sua vez, possibilita estudos em escalas antes infactíveis. Na abrangência das tecnologias de nova(ou próxima) geração (*next generation*) a disponibilidade de cepas sequenciadas nos bancos de dados biológicos cresceu na última década [Mardis 2011].

Deste modo, o HIV, como um agente etiológico largamente estudado, apresenta um considerável conjunto de dados disponível. Apenas no GenBank, são disponibilizadas mais de 550 mil sequências distribuídas em mais de 60 grupos, entre subtipos puros e formas recombinantes únicas e circulantes. Desta forma, com um conjunto de dados tão grande como o disponível no GenBank, os procedimentos de obtenção de informações como mapeamento e subtipagem tornam-se relativamente árduos [Hemelaar 2012].

Seguindo a estratégia serial utilizada atualmente seriam necessários cerca 40 milhões de alinhamentos para que se localize o subtipo mais provável de cada sequência o que seria inviável computacionalmente. Percebe-se então a necessidade de implementação de um modelo computacional que obtenha de forma recursiva o conjunto de dados disponível no GenBank, e realize o alinhamento, mapeamento e subtipagem das sequências disponíveis neste, armazenando os dados gerados pelos procedimentos para futuras análises.

1.1 Considerações Preliminares

A heterogeneidade genotípica e a elevada taxa mutacional apresentada pelo HIV representam fatores importantes que contribuem para a dificuldade do combate a

Síndrome da Imunodeficiência Adquirida (AIDS). Para melhor compreender esses fatores é necessário o estudo extensivo da estrutura genética do vírus e da distribuição dos subtipos virais. Desta forma será possível identificar principalmente, os fatores relacionados com essa elevada taxa mutacional e suas consequências. A identificação de áreas imunogênicas evolutivamente estáveis entre os diversos subtipos e elementos virais seria o ponto chave para o desenvolvimento de tratamentos eficazes para as doenças cujos agentes etiológicos possuam comportamento mutacional similar ao do vírus em questão [de Queiróz et al. 2011].

Dessa maneira, é necessário obter constantemente a informação genética disponível nos bancos de dados, indexar e então mapear e subtipar as sequências nucleotídicas. Os dados resultantes deste processo vão determinar as frequências e os quantitativos de variação genotípica viral destes indivíduos facilitando a compreensão de sua dinâmica evolutiva e apresentando um quadro claro do estado atual do organismo [Hemelaar 2012]. Este quadro de informações sobre a dinâmica evolutiva é relevante para o desenvolvimento de tratamentos eficazes para a patologia associada. Uma vez que a elevada taxa mutacional é a principal causa da ausência de intervenções eficazes [Castro-Nallar et al. 2012]. Entretanto, o grande volume de dados referentes ao HIV e o elevado custo computacional para analisar essas sequências tornam essa tarefa laboriosa e complexa.

A avaliação de grandes volumes de dados requer algorítimos específicos, otimizados e direcionados à aquisição, mapeamento e alinhamento de sequências virais. Para suprir tal necessidade, a plataforma de gerenciamento de sequências nucleotídicas virais (VSDBM) oferece um *workbech* para obtenção, inserção e indexação das sequências nucleotídicas disponíveis. Além disso, oferece um extenso *framework* para desenvolvimento e aplicação de análises nas sequências inseridas [Irahe Kasprzykowski 2013].

1.2 Objetivos

1.2.1 Objetivo Geral

Um software de obtenção recursiva e análise das sequências nucleotídicas do HIV disponíveis no GenBank.

1.2.2 Objetivos Específicos

- Desenvolver um modelo de dados adaptado às características genotípicas do organismo.
- Criar um banco de dados biológico com as sequências provenientes do GenBank.

- Modelar, Desenvolver e aplicar o mapeamento das sequências no genoma completo.
- Modelar, Desenvolver e aplicar a classificação das sequências nos subtipos e formas recombinantes conhecidas atualmente.
- Disponibilizar publicamente a aplicação e o conjunto de dados.

1.3 Contribuições

Visando melhorar o *workflow*, este trabalho pretende implementar um modelo de dados capaz de armazenar o conjunto de dados obtidos, além de permitir sua indexação. Este modelo deve ainda apresentar uma flexibilidade que permita a posterior inserção e indexação das informações obtidas com as análises. Além de permitir a busca recursiva diária nos principais bancos de dados por novas sequências.

Serão implementados novos protocolos de análise de sequências para identificar novas características e adiciona-las às *features* da sequência. Assim, a necessidade do desenvolvimento de um modelo computacional capaz de alinhar, mapear e subtipar as sequências inseridas nesta plataforma fica evidente. Isto possibilita a identificação de novas informações relevantes sobre a estrutura genética viral, reconhecimento imunológico e aspectos evolutivos, auxiliando o desenvolvimento de técnicas eficazes de combate às doenças causadas pelos mesmos.

1.4 Organização do Trabalho

Este trabalho esta organizado em seis capítulos: Introdução, Revisão da Bibliografia, Metodologia, Resultados, considerações finais e Referências.Na Introdução o leitor entra em contato com o problema de estudo deste trabalho e é apresentado uma breve contextualização na temática que será mais profundamente abordada na próxima seção. Foi feita uma revisão bibliográfica sobre os principais temas abordados neste trabalho.

Na Metodologia, é descrito o processo utilizado para alcançar o objetivo esperado. Na seção de Resultados Parciais, estão descritos todos os resultados obtidos até o momento com a aplicação dessa metodologia. As considerações finais apontam as contribuições e impactos deste trabalho aliados uma breve discussão sobre os resultados. A seção Referências é onde podem ser encontradas os trabalhos e fontes que embasaram esta pesquisa.

Capítulo 2

Revisão da Bibliografia

2.1 HIV

O *Human Imunodeficiency Virus type 1* (HIV-1), pertencente à família *Retroviridae*, gênero *Lentivirus*, foi identificado como agente etiológico da Síndrome da Imunodeficiência Humana Adquirida (AIDS)[Barré-Sinoussi et al. 2013, Barre-Sinoussi, F., J. C. Chermann 1983, Gallo et al. 1983]. Esta patologia atingiu no Brasil cerca de 700 mil casos entre 1980 e 2012 [MINISTÉRIO DA SAÚDE 2012].

Nas primeiras duas semanas de infecção ocorre a disseminação viral pelo organismo do hospedeiro, porém sem apresentação de sintomas. Até a quarta semana de infecção é apresentada uma elevação na viremia, uma alta taxa de infecção de células T CD4+ e linfonodos acompanhados de sintomas similares a uma gripe viral. Após esta fase, existem apenas mais duas, uma que pode chegar a vinte anos de duração, onde se observa um crescimento acentuado da viremia e por via de regra não há apresentação de sintomas. A última fase de infecção se caracteriza pelo decaimento na contagem das células T CD4+ o que ocasiona a falha imune no organismo infectado abrindo caminho para infecções oportunistas podendo levar à morte do indivíduo [Coffin e Swanstrom 2013].

A principal característica da infecção por HIV é a disfunção progressiva do sistema imune, avançando para a síndrome da imunodeficiência adquirida, na maioria dos casos observados. A dinâmica evolutiva apresentada pelo agente etiológico em questão durante a infecção é guiada inicialmente pelo sistema imune do paciente, e posteriormente pela pressão seletiva inferida pelos fármacos. Normalmente, um período não sintomático é apresentado pelo paciente durante os primeiros anos de infecção onde os níveis virais são baixos, porém os níveis de replicação e variabilidade chegam a 10^{10} por dia [Neher e Leitner 2010].

Este organismo apresenta em seu genoma completo cerca de 9 mil pares de base. Seu genoma é flanqueados por regiões não traduzidas, que participam diretamente

na reprodução viral e na integração com o material genético do hospedeiro. Estas regiões estão divididas em LTR-3' e LTR-5' [van der Kuyl e Berkhout 2012].

Os genes apresentados no genoma do HIV são o gag que codifica as proteínas *p1*, *p2* e *p6*, além das das estruturas do capsídeo (*p24*), nucleocapsídio (*p7*) e matriz (*p17*), o *env* que codifica as proteínas do envelope viral (*gp120* e *gp41*), e o *pol* que codifica as enzimas responsáveis pela protease, transcriptase reversa, e integrase. Outras regiões com quadros abertos de leitura menores, codificam proteínas adicionais como *Vpr*, *Vpu*, *Vif*, *Tat*, *Nef*, *Rev*, possuem função regulatória [van der Kuyl e Berkhout 2012].

As polimerases responsáveis pela replicação viral do organismo, não possuem atividades de correção possuem taxas de erro elevadas se comparadas às das polimerases codificadas por eucariotos. Este fator então, juntamente com a elevada taxa de replicação viral e a recombinação genética apresentada por este organismo, causa inserções e deleções no genoma viral, além de produzir diversos tipos de mutações. Isto posto, estas mutações contribuem diretamente para a resistência apresentada por este vírus em relação aos fármacos [Coffin e Swanstrom 2013].

Estas modificações no genoma viral do HIV ocasionam uma elevada heterogeneidade genotípica. Esta característica contribui significativamente para a adaptabilidade dos organismos virais em questão. Uma vez que mutações nas áreas reconhecidas pelo sistema imune ou nas áreas de ação dos fármacos, causam não pareamento com o MHC e a consequente ausência de resposta imune. Estes fatores contribuem para a sobrevivência desta quasiespécie, proliferando assim este genótipo [Rouzine et al. 2014]. Além de facilitar a sobrevivência dos patógenos no organismo, mutações em determinadas áreas do genoma geram uma elevada diversidade genotípica viral, o que dificulta a identificação de novas regiões alvo para intervenções terapêuticas [UK Colaborative Group on HIV Droug Resistance 2013].

2.2 Variabilidade Genética do HIV

A variabilidade genética do vírus em questão é uma das principais causas da ausência de tratamentos eficazes para as doenças causadas pelos mesmos. Estudos foram realizados sobre a variabilidade genotípica do HIV onde se inferiu que este é dividido filogeneticamente em três grupos principais: O (atípico), N (novo, não-N, não-O) e M (principal). Destes grupos, é possível destacar o grupo M, pois os vírus que pertencem a este grupo são os principais responsáveis pela pandemia da doença. Estes estão subdivididos em 9 subtipos distintos, além de apresentar mais de 70 formas recombinantes circulantes e únicas (CRF's e URF's) [UK Colaborative Group on HIV Droug Resistance 2013, Hemelaar 2012].

A variação viral responsável pela infecção em seres humanos, se originou na África central, onde ocorreu infecção de seres humanos por variações do vírus da imunodeficiência símia (SIV). Estre processo de infecção ocorreu provavelmente no processo

de caça e corte da carne para alimentação, além da venda e distribuição de símios como animais de estumação [Hahn et al. 2000]. Desta forma, eventos independentes de transmissão do vírus de primatas não humanos para humanos ocasionou a criação de algumas linhagens virais agrupadas em dois tipos principais, o HIV-1 e o HIV-2. O tipo 1 é apontado como mais prevalente na pandemia, por conter o grupo M, que infecta cerca de 33 milhões de pessoas[Hemelaar 2012].

Um dos principais fatores que contribuem para a dispersão da pandemia no mundo, é a variabilidade genética apresentada pelo agente etiológico, associada à rápida evolução viral. Esta variabilidade é causada pela elevada taxa de replicação viral aliada a elevada taxa mutacional e de recombinação da enzima transcriptase reversa, que não possui mecanismo de auto-correção [Roberts et al. 1988, Ho et al. 1995] .

Tendo em vista a elevada taxa mutacional, a identificação de regiões imunogênicas evolutivamente estáveis no HIV seria o ponto chave para o desenvolvimento de tratamentos eficazes. Para identificar estas regiões, é necessário obter informações acerca da estrutura genética, genotípica e comportamento mutacional. Como demonstrado para o HCV (*Hepatitis C Virus*), que possui comportamento mutacional semelhante ao do HIV [de Queiróz et al. 2011].

Para obter informações desta natureza a respeito do vírus em questão, é necessário aplicar alguns tratamentos ao conjunto de dados disponível. Sendo assim, amostras destes organismos são sequenciadas e armazenadas formando um conjunto de dados. Posteriormente são aplicados a estes, técnicas de bioinformática como alinhamento, mapeamento e subtipagem que resultarão em informações a respeito da relação genotípica das sequências com as estruturas mapeadas anteriormente.

Desta forma é possível fornecer um conjunto de dados mais completo e confiável. Um dos pontos importantes no desenvolvimento de estratégias de tratamentos eficazes para a etiologia em questão, é o acompanhamento de variantes virais emergentes. É necessário então conhecer a variabilidade apresentada no cenário atual levando em consideração o conjunto de dados completo, e não só uma secção restrita [Chan et al. 2014].

Para observar o comportamento mutacional do organismo em questão e a ocorrências de regiões codificantes é necessário analisar as sequências disponíveis nos bancos de dados primários como o GenBank. Para obter informações como a localização de uma determinada sequência no genoma completo, é realizado o processo de alinhamento global da sequência de referência com a sequência *query* (assumindo sua homologia, ou seja, que ambos partem de um ancestral comum) para identificar o posicionamento dos nucleotídeos de uma em relação à outra.

O processo de alinhamento geralmente é feito através de algoritmos matemáticos, modelados computacionalmente de forma a implementar técnicas de matriz de score. Por exemplo, o algoritmo de Needleman e Wunsch utiliza conceitos de programação dinâmica para buscar o alinhamento global ótimo [Needleman e Wunsch 1970, Day 2010].

2.3 Alinhamento de Sequências

Para comparar duas sequências nucleotídicas é necessário alinha-las. O alinhamento de sequências, se trata do processo de pareamento entre duas ou mais sequências. Esse processo leva em consideração a similaridade entre determinadas regiões destas sequências, assim como *indels*, que são inserções ou deleção de nucleotídeos. O termo é originado, pois não é possível determinar, sem análises filogenéticas, se houve inserção ou deleção no sítio em questão.

Por exemplo, assumindo a comparação entre duas sequências, é possível representá-las em duas dimensões para identificar as possibilidades de pareamento entre as sequências em questão. O resultado do processo de alinhamento trata-se do coeficiente de similaridade entre as duas sequências, além de uma sequência que representa o pareamento entre as regiões [Deshmukh e Kharat 2015].

Em diversos estudos, os alinhamentos são descritos em dois formatos funcionais que são local e global. O alinhamento global, pressupõe a homologia ou ancestralidade comum entre as duas sequências e constrói um alinhamento partindo da suposição de que as sequências possuem o mesmo tamanho, enquanto o alinhamento local, busca o local de maior nível de similaridade [Deshmukh e Kharat 2015]. No alinhamento global, o algoritmo considera 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 sítio n da sequência B.

O alinhamento ótimo é obtido como resultado de algoritmos exaustivos(ou exatos) de alinhamento, e é bastante utilizado no processo 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. Já a utilização de heurísticas no processo de alinhamento, surgiu pela necessidade de tratamento de bases de dados grandes. Uma vez que a estratégia utilizada no encadeamento de algoritmos exatos não satisfazia as necessidades de otimização de recursos [Chakraborty e Bandyopadhyay 2013].

Os alinhamentos ótimos de sequências como os propostos por Needleman e Wunsch, e Smith e Waterman, [Needleman e Wunsch 1970, Smith, T. F.; Waterman 1981] possuem uma ordem de complexidade computacional proporcionais ao tamanho das sequências trabalhadas ou seja, de ordem quadrática de tempo e recurso de máquina. Uma vez que ambos utilizam matrizes de computação dinâmica, onde cada possibilidade de alinhamento é calculada, e posteriormente anotada, para que este tipo de algoritmo possa garantir sempre o melhor alinhamento possível entre estas duas sequências.

Um algoritmo de alinhamento exaustivo, utilizado para realizar alinhamentos do tipo global, é o Needleman e Wunsch [Needleman e Wunsch 1970]. Este algoritmo utiliza técnicas de programação dinâmica para construir e percorrer uma matriz, que se refere à relação entre os nucleotídeos das sequências em questão. Durante o processo de construção da matriz (Figura 2.2), o algoritmo toma uma série de decisões, que podem ser uma das representadas na Figura 2.1.

$$u_k = \begin{cases} \searrow & \text{Diagonal direita inferior,} \\ \rightarrow & \text{Direita,} \\ \downarrow & \text{Inferior.} \end{cases}$$

Figura 2.1: Decisões tomadas pelo algoritmo ao percorrer a matriz. Fonte: [Polanski e Kimmel 2007]

	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

Figura 2.2: Matriz de escores para o alinhamento utilizando o algoritmo de Needleman e Wunsch. Fonte: [Polanski e Kimmel 2007]

Cada decisão tomada em uma posição da matriz gera um estado, construindo uma matriz de estados, onde o próximo estado é definido pelo estado atual, dada as posições nas sequência e a decisão atual. Esta matriz de estados é construída como definido na Figura 2.3, e posteriormente percorrida construindo a sequência final de alinhamento.

	s_2	a	c	g	t	g	a	g	a	g	t
s_1	↘→	↘→	↘	↓	↓	↓↘	↓	↓↘	↓↘	↓	↓
t	↘	↓→	→	↘	↓	↓↘	↓	↓↘	↓	↘	↓
t	→	↘	↘→	↘	↓	↓↘	↓	↓↘	↓	↓↘	↓
c	→	→	↘→	→	↘	↘	↓	↘	↓	↓↘	↓
g	→	→	↘→	→	↘	→	↘	→	↘	↓↘	↓
g	→	→	→	→	→	→	→	↘	↓↘	↘	↓
a	→	→	→	→	→	→	→	→	→	→	STOP

Figura 2.3: Matriz de decisões otimizadas do algoritmo de Needleman e Wunsch. Fonte: [Polanski e Kimmel 2007]

O coeficiente de similaridade entre as sequências é calculado de forma cumulativa à matriz de estados atuais, de forma que é levado em consideração o valor resultante de cada pareamento, e em seu somatório é obtido o escore final de similaridade entre as sequências como exibido na seguinte fórmula:

$$S = \max_{u_1, u_2, \ldots, u_L} \sum_{k=1}^{L} f(x_{k+1}, u_k).$$

Outro algoritmo exato largamente utilizado no processo de análise de sequências nucleotídicas é o Smith e Waterman. Este algoritmo não pressupõe homologia entre as sequências, por tanto oferece um alinhamento local entre as sequências(local de maior similaridade). Para tanto, uma matriz é criada, posteriormente preenchida, e finalmente percorrida. Desta forma é gerado o alinhamento entre as regiões com maior similaridade. Para identificar tais regiões, o algoritmo atribui uma penalidade às áreas não pareadas. Estas penalidades permitem que a alinhamento final não possua necessariamente o mesmo tamanho das sequências alinhadas [Deshmukh e Kharat 2015].

Outras técnicas de alinhamento de sequências largamente utilizadas são as consideradas Heurísticas [Pruesse et al. 2012, Chakraborty e Bandyopadhyay 2013]. Estas técnicas não trabalham com a totalidade das possibilidades de alinhamento de duas determinadas sequências para inferir o melhor alinhamento. Este tipo de algoritmo determina dinamicamente as possibilidades de alinhamento menos prováveis e as elimina do processo de alinhamento.

Entretanto a utilização de técnicas heurísticas como *Hidden Markov Models* (HMM), alinhamento progressivo, técnicas iterativas determinísticas e computação evolutiva (Algoritmos Genéticos) reduz a acurácia do alinhamento. Além de depender diretamente da forma como os dados estão dispostos [Chakraborty e Bandyopadhyay 2013].

Desta forma, o tratamento heurístico em bases de dados extremamente dinâmicas e evolutivas não seria recomendável. Uma vez que dados são inseridos constantemente, modificando a configuração do conjunto de dados. Desta forma é modificada consequentemente a dinâmica da análise heurística. O que impacta diretamente em sua parametrização e consequentemente o resultado. Sendo assim é ratificada a necessidade de um sistema adaptativo, que utilize algoritmos pré-parametrizados que se adaptem à um conjunto de dados dinâmico.

2.4 Mapeamento de Sequências

O procedimento de mapeamento de sequências no genoma completo é considerado essencial para a análise de dados provenientes sequenciamento de alto desempenho. Este procedimento permite determinar quais genes a sequência compreende. Além

disso, caso alguma mutação tenha ocorrido é possível rastrear o gene em que aquela mutação ocorreu, e assim melhorar a abordagem de tratamentos antirretrovirais [Combe e Sanjuán 2014].

No conjunto de dados disponibilizado pelo GenBank por exemplo, a média de tamanho de sequências do HIV é de apenas 1000 pares de base. Enquanto o genoma completo do vírus possui mais de 10 vezes esse tamanho(10700pb). Isso ocorre devido a variação das técnicas de sequenciamento e abordagens de montagem dos *contigs*. Desta forma, o processo de identificação da correlação entre uma determinada sequência, e o genoma completo torna a informação contida no fragmento mais relevante.[Vrancken et al. 2016]

Essencialmente, o processo de mapeamento de uma sequência nucleotídica parte da construção de um mapa de características. Este mapa é criado a partir de uma sequência de referência ou genoma completo onde informações funcionais e posicionais são levadas em consideração. A partir do alinhamento entre a sequência em questão e a sequência de referência, é possível identificar quais áreas da sequência inicial se referem em relação à sequência de referência do organismo, desta forma é possível identificar quais características são comuns. Este processo possui uma ordem de complexidade proporcional ao tamanho das sequências alinhadas.

Durante o processo de mapeamento, é realizado o alinhamento global entre a sequência em questão e a sequência de referência do organismo. Este processo de alinhamento indicará quais regiões do genoma completo são "cobertas"pelo genoma em questão. A partir desta "cobertura"é possível identificar quais características presentes no genoma completo são completamente expressas na sequência em questão, e quais são apenas parcialmente expressas.

A "cobertura"de uma determinada sequência é determinada pela posição de início e fim do alinhamento. A posição de início do alinhamento é determinada pela primeira posição pareada em relação ao genoma completo. Enquanto a última posição pareada do alinhamento corresponde ao final da cobertura.

Ao analisar cada característica mapeada do genoma completo, é necessário verificar a presença de *indels* na região observada. Desta forma é possível qualificar a "cobertura"como parcial ou total, indicando respectivamente se a sequência em questão codifica totalmente ou apenas parcialmente a característica observada.

Com este mapa de "coberturas", é possível identificar conjuntos de dados alvo para determinados estudos. É possível ainda selecionar sequências que codifiquem 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 novos tratamentos, e qualificação dos existentes.

2.5 Subtipagem de Sequências

Além de informações sobre localização da sequência *query*, no genoma completo, é necessária a identificação do subtipo viral para identificar as restrições e características deste organismo em específico, a partir de um perfil já traçado de seu grupo [Chan et al. 2014].

O processo de subtipagem de uma determinada sequência, se trata de selecionar entre os subtipos já identificados do organismo, aquele em que a sequência melhor se encaixa. É necessário então, determinar todas as sequências de referência (que representam as características do subtipo) do subtipo.

As sequências de referência são alinhadas localmente com a sequência em questão, obtendo-se um escore (coeficiente de similaridade) para cada alinhamento. Com este escore é possível identificar o nível de similaridade da sequência *query* com os subtipos, permitindo assim, que o subtipo mais similar seja selecionado.

Como o processo de subtipagem demanda um ou mais alinhamentos, a complexidade deste é proporcional ao tamanho das sequências alinhadas em relação a quantidade de subtipos identificados no organismo em questão. Sendo assim, se um organismo possui dez subtipos com apenas uma sequência de referência cada um, é necessário realizar dez alinhamentos da ordem dos tamanhos das referências de cada um deles.

O processo de subtipagem exaustivo é essencial para a investigação da resistência viral e das diferenças na patogênese entre os subtipos [Pineda-Peña et al. 2013]. Por consequência são necessárias as sequências de referência de subtipos e recombinantes deste organismo. Estas sequências devem ser alinhadas localmente com a sequência da cepa alvo. Desta forma, os valores de similaridade entre a cepa alvo e os subtipos informam a qual subtipo aquela determinada sequência mais se parece [Chan et al. 2014].

2.6 Gerenciamento do Conjunto de Dados

Devido ao avanço tecnológico e do baixo custo do sequenciamento de alta demanda, os conjuntos de dados biológicos cresce exponencialmente. O desenvolvimento de dispositivos para o gerenciamento de grandes quantidades de dados biológicos é considerado fundamental em bioinformática [Zou et al. 2015]. Em 2014 foi reportada a existência de 1552 bases de dados de acesso público [Fernández-Suárez et al. 2014].

Em termos de classificação de bases de dados biológicas, se pode dividir em três principais: Escopo da cobertura dos dados, tipo de dados gerenciados, método de curagem dos dados. A respeito do processo de curagem, os bancos de dados biológicos podem der divididos em primários, secundários e especializados.

Os bancos de dados considerados primários contém dados brutos, geralmente não curados, como o GenBank. Já os bancos de dados secundários possuem um certo

nível de curagem. Os bancos de dados especializados, possuem por sua vez dados de um determinado organismo, com um certo nível de curagem [Zou et al. 2015].

O conjunto de dados disponível no Genbank sobre o vírus da imunodeficiência humana é muito grande, o que torna o processo de aplicação de modelos matemáticos e computacionais uma tarefa árdua. Este conjunto de dados é composto por mais de meio milhão de sequências nucleotídicas disponíveis no Genbank.

Estas sequências são distribuídas em 4 grupos, dos quais o M se destaca como principal e apresenta 9 subtipos ou tipos subordinados puros, e mais de 70 formas recombinantes que são quimerizações entre mais de um subtipo puro. Atualmente são necessários cerca de 40 milhões de alinhamentos apenas para que se obtenha o subtipo no qual cada sequência está classificada [Crous et al. 2012], o que seria um processo laborioso, complexo e computacionalmente inviável.

Além da extensão do conjunto de dados, ao utilizar matrizes de escore como base para a comparação de sequências, seja local ou global, o modelo estratégico atual, encontra a restrição do tamanho das sequências. Uma vez que computacionalmente, existem limites para a quantidade registros de uma matriz. Este limite é associado diretamente à quantidade de endereços que podem ser criados e gerenciados para um mesmo objeto. Desta forma é criado mais um grande desafio no tratamento de dados.

Recentemente nosso grupo demonstrou a possibilidade de obtenção e indexação de sequências nucleotídicas a partir de diversas técnicas de busca e padrões de formatação. Um exemplo de trabalho neste sentido, é a plataforma VSDBM [Irahe Kasprzykowski 2013]. Plataforma esta que disponibiliza um extenso *framework* de desenvolvimento de técnicas de bioinformática. Este *framework* possibilita o desenvolvimento de um modelo computacional capaz de unir a inserção otimizada e a indexação de sequências nucleotídicas virais à novas técnicas avançadas de tratamento de dados biológicos em um software que possua os principais procedimentos de bioinformática como alinhamento, mapeamento e subtipagem de sequências.

2.7 Cenário Atual

No cenário atual, a base de dados utilizada como padrão de consulta de informações sobre o HIV é a "*HIV Databases*", disponibilizada pelo Laboratório Nacional de Los Alamos. Apesar de existirem outros bancos de dados com informações sobre o agente etiológico, este conjunto de dados é considerado um banco de dados biológico especializado, com informações a respeito de sequências genéticas, epítopos, mutações associadas a resistência a fármacos e testes de vacinas [Brander et al. 2014].

Esta base de dados representa o padrão adotado pela maioria dos pesquisadores da área, contendo dados curados do organismo em questão. Estes dados referem-se não só ao HIV-1, mas aos tipos 2, 3 e ao SIV (*simian imunodeficiency virus*).

Com informações sobre sua estrutura genética e resposta imune, além de diversas ferramentas de gerenciamento e auxílio à análise [Brander et al. 2014].

Das sequências disponíveis no *HIV Databases*, é possível destacar a presença de 580.490 sequências curadas. Tais sequências representam uma população proporcional aos três organismos representados por este banco de dados biológico. É possível realizar o download de cerca de 350.000 sequências nucleotídicas do HIV-1. Dentre estas sequências, temos ainda sequências que representam o grupo N, O e P em menor número [Foley et al. 2012].

A classificação das sequências no *HIV Databases* é geralmente realizada pelo autor original. Portanto, os métodos de classificação variam de acordo com o que foi abordado pelo autor do sequenciamento original. Apenas sequências já classificadas são inseridas no conjunto de dados, uma vez que não existe uma técnica centralizada de classificação deste dados [Los Alamos National Laboratory 2015b]).

Este processo de classificação não leva em consideração a atualização das classes disponíveis, uma vez que os autores que identificaram sequências mais antigas só classificaram suas sequências utilizando como base os subtipos e formas recombinantes disponíveis no momento da classificação. Este processo vem ocorrendo desde a década de 80, enquanto as formas recombinantes circulantes mais recentes foram identificadas ainda em 2015.

No GenBank existem cerca de 580.000 sequências relacionadas ao HIV-1. Estas sequências estão disponíveis. Em termos de cobertura, o genoma do HIV se encontra de forma fragmentada no GenBank, principalmente devido à variação nas técnicas de sequenciamento. É possível constatar este fato ao observar que apesar de o genoma completo deste organismo possuir cerca de 10.000 pares de base, o tamanho médio das sequências disponíveis é de apenas 1.000 pares de dados.

Capítulo 3

Metodologia

O trabalho consistiu na modelagem e construção de um software capaz de obter, indexar, e aplicar técnicas de bioinformática nas sequências de HIV-1 disponíveis no GenBank. Para isso, foi necessária a obtenção e análise das sequências disponíveis permitindo a identificação das características específicas de genes, subtipos e genótipos. Estas características foram essenciais para o desenvolvimento de um modelo relacional que pudesse oferecer o suporte necessário ao desenvolvimento do módulo de análise das informações genômicas do agente etiológico.

A partir das informações genotípicas das estruturas virais, foram desenvolvidos dois modelos de armazenamento de dados: o modelo de armazenamento de dados de apoio à análise, e o modelo de dados de armazenamento final das informações. Ambos os modelos foram adaptados ao modelo de gerência de organismos virais já disponível no pacote do VSDBM e finalmente fundidos em um único modelo.

Com o modelo de dados pronto, foi necessário modelar um software que realize a obtenção, modulação e importação das sequências nucleotídicas disponíveis no GenBank. Após o processo de modelagem do software, o processo de importação foi executado e completado para que o conjunto de dados de apoio à análise possa fornecer as informações necessárias.

Em posse do conjunto de dados organizado e indexado, foi então desenvolvida a estratégia de mapeamento. Esta estratégia consiste na interação de fatores como computação dinâmica, relacionamento direto entre objetos de baixo nível, matemática computacional, computação paralela e etc. Estes fatores foram dispostos de forma que o processo de mapeamento das sequências seja realizado no menor tempo possível com a máxima acurácia oferecida pelos algoritmos de alinhamento exaustivos.

A partir do mapeamento foi desenvolvida a estratégia de subtipagem das sequências, que conceitualmente é o processo mais árduo. Isso se dá pois o conjunto de dados disponível é grande e em relação a um número relativamente grande de subtipos conhecidos, gera uma quantidade de comparações elevada. A estratégia de subtipagem

contemplou esta característica do conjunto de dados, objetivando reduzir a quantidade de comparações necessárias. Desta forma foram utilizadas principalmente estratégias de agrupamento por derivação de recombinação, desistência sumária, pré-comparação, clusterização e pré-estruturação dos dados.

Com o processo de subtipagem dos dados pronto, foi realizada uma validação dos dados gerados com o *Dataset*. Desta forma foi possível identificar o nível de similaridade do cenário em escala global, com o cenário disponível no HIV Database. Além deste processo de validação, foi realizada a subtipagem dos dados utilizando a técnica de comparação todos contra todos. Onde todos os alinhamentos com as sequências de referência são realizados, no intuito de identificar divergências nas estratégias de subtipagem aplicadas.

Após o processo de análise e armazenamento dos dados, as bases de dados foram disponibilizadas online para download, no portal do VSDBM. Além dos dados obtidos na análise, o aplicativo desenvolvido será disponibilizado no mesmo portal, juntamente com seu código fonte para posteriores implementações e análises.

Visando a dinamização do conjunto de dados e a melhor vigilância da pandemia causada pelo HIV, foi construído um módulo de obtenção e análise recursiva, que estará constantemente atualizando o conjunto de dados, para mante-lo sempre atualizado. Obtendo assim um conjunto de dados sempre mais completo acerca do agente etiológico em questão.

3.1 Adaptação do modelo de dados

O modelo de dados proposto foi modelado e desenvolvido levando em consideração a heterogeneidade apresentada pelo HIV. Onde a taxa de substituição anual por loco é de ∼0,002 (10^{-3} mutações por sítio por ano), e a taxa de mutação no genoma por geração é de 0,2. Estas taxas representam um comportamento mutacional elevado, se levado em consideração o fato de que o HIV-1 apresenta alta reprodutibilidade em um curto espaço de tempo [Castro-Nallar et al. 2012].

Por tanto, o modelo de dados foi implementado alinhado às peculiaridades genotípicas do HIV, de forma a contemplar inclusive as formas recombinantes circulantes. Essas formas são recombinações geradas a partir da coinfecção de dois ou mais subtipos e/ou formas recombinantes. Já as formas recombinantes únicas são recombinações entre cepas do mesmo subtipo com histórico evolutivo diferente [UK Colaborative Group on HIV Droug Resistance 2013].

O modelo desenvolvido permite ainda a identificação de quadros abertos de leitura (*ORF's*). Desta forma possibilita a posterior identificação de códons de início e parada. Esta identificação facilita o mapeamento de regiões codificantes e não codificante nas sequências virais disponíveis no conjunto de dados. Assim, o modelo proposto pela plataforma VSDBM atendeu apenas parcialmente ao funcionamento

deste sistema, pois contempla apenas de forma generalizada a análise de subtipos de um determinado organismo.

O modelo sintetizado teve ainda que tratar as subdivisões e inter-conexões apresentadas pela recombinação viral, onde uma determinada forma recombinante circulante pode ter derivado de dois ou mais subtipos considerados puros. Estas subdivisões e inter-conexões foram completamente satisfeitas durante o processo de cadastro dos subtipos e facilitaram o processo de estruturação dos dados para comparação.

Além dos dados necessários para a aplicação dos procedimentos de bioinformática, o modelo ainda dispõe de vertentes que possibilitam a inserção dos resultados do mapeamento e subtipagem. No processo de subtipagem das sequência é possível armazenar todos os alinhamentos realizados, ou apenas o melhor alinhamento.

Este formato de indexação dos dados permite a aplicação posterior de novas análises, de forma a considerar os procedimentos já realizados e assim evitar a repetição dos procedimentos. Caso novos subtipos sejam identificados, só será necessário comparar as sequências disponíveis na base com este novo subtipo, comparando o coeficiente de similaridade obtido com aquele já disponível no modelo. Isso evitará a repetição do processo de subtipagem, fazendo com que a eficácia do sistema seja exponencialmente elevada a medida que novos dados sejam adicionados.

Para criar o modelo de dados foi utilizado a ferramenta gráfica de modelagem de bases de dados MySQL *Worckbench*, onde o modelo foi sintetizado e posteriormente aplicado ao servidor de banco de dados. A versão 5.6 do referido SGBD foi utilizada em sua licença *Community*.

3.2 Obtenção das sequências

A obtenção de sequências será feita de 2 formas diferentes, de forma que seja oferecida uma maior flexibilidade da base de dados, montando um conjunto de dados mais especificado. As formas de obtenção e inserção de sequências são:

3.2.1 Inserção via arquivos de texto

Esta técnica de inserção é baseada na inserção de arquivos de sequências em massa no formato GenBank(gbk). O sistema irá executar a leitura gradual deste arquivo para extrair neste processo, as sequências. Devido à natureza dos arquivos gbk, será necessário criar uma estratégia de modulação dos dados, adaptando-os aos modelos de suporte a análise.

Observando que várias sequências podem ser armazenadas no mesmo arquivo Genbank, é necessário que este arquivo não seja completamente carregado em memória. A abertura e leitura gradual do *stream* de dados foi implementada como solução

para o tamanho do arquivo e a marcação de finalização das sequências deve ser respeitada como final da sequência em questão.

3.2.2 Obtenção Automática

Cadastrado o organismo, o sistema realiza uma busca utilizando o cadastro do organismo no serviço externo do GenBank pelos id's das sequências disponíveis para download através da utilização da ferramenta ENTREZ através do protocolo SOAP. Os dados são então transferidos no formato intermediário em XML. Durante este processo é realizada uma verificação na base de dados local e a construção de uma lista de sequências a obter, através de técnicas de busca em profundidade, evitando a repetição de sequências e o download desnecessário.

Após gerada a lista de id's exclusivos das sequências que ainda não estão disponíveis na base de dados, o sistema deve utilizar esta como parâmetro para contabilização e obtenção. O processo de obtenção ocorre através do consumo de um serviço web ou *webservice* disponibilizado pelo próprio NCBI. Após este passo os downloads sumários são realizados, discriminando seu progresso. A medida que os downloads das sequências são finalizados, um processo paralelo deve, antes de armazenar estas sequências na base de dados, modular os dados lendo o modelo padrão de obtenção e traduzindo o para o formato reconhecido pela plataforma.

3.3 Mapeamento das sequências do HIV

A interação do vírus com o hospedeiro, assim como a dinâmica da infecção é influenciada diretamente pelas características do vírus. Estas características são determinadas pro estruturas codificas em regiões diferentes no genoma. Assim, a determinação e mapeamento destas regiões no conjunto de dados seria um ponto chave para o desenvolvimento de tratamentos eficazes.

A inibição de resposta imune por parte do hospedeiro é associada diretamente à alta variabilidade apresentada pelo HIV. Desta forma, o estudo do processo de interação entre o agente etiológico e o sistema imune do hospedeiro se torna um ponto relevante para o desenvolvimento de uma vacina eficaz [Henn et al. 2012].

Desta forma, é necessário conhecer a interação e a dinâmica viral, principalmente os epítopos que são reconhecidos pelo sistema imune e sua prevalência mutacional. Apesar de ser um processo complexo e laborioso, a identificação de epítopos imunologicamente relevantes é criticamente importante para o controle da infecção. Uma vez que se observa que a ausência de resposta imune por parte do hospedeiro em relação ao vírus, se dá principalmente por mutações de escape. Estas mutações são mutações em regiões do genoma que codificam epítopos [Roider et al. 2014] .

Considerando a existência de uma gama de epítopos já identificados, é necessário identificar sua prevalência. Então mapear suas regiões codificantes dentro do conjunto de dados, para que se possa validar um possível preditor se torna essencial. Este preditor possibilitará a indicação melhores estratégias de combate à patologia associada ao agente etiológico em questão [Abidi et al. 2014].

Entretanto, além da interação com o sistema imune do hospedeiro, é necessário identificar regiões alvo para ação de fármacos que inibem determinadas funções do vírus. Estas funções são codificadas em regiões do genoma. Assim, o mapeamento destas regiões representa um relevante avanço para a criação de conjuntos de dados para estudos nesse sentido.

Além dos epítopos, os dados disponíveis no NCBI dispõem de características pré identificadas nas sequências. Porém não necessariamente relacionadas ao genoma completo, de forma que diversas sequências podem não representar em sua estrutura, toda a sequência do genoma completo, representando apenas alguns de seus aspectos. Onde esta representação pode ser total ou parcial, indicando uma área de elevada taxa mutacional. Obter tais informações de mapeamento dos traços das sequências seria relevante no processo de organização das informações(sequências).

O mapeamento de sequências, ocorre para organizar as sequências disponíveis no banco de dados por áreas relacionadas ao genoma completo do organismo. De forma a agrupar aquelas que possuem ”cobertura”completa ou parcial das propriedades já disponíveis e mapeadas no genoma completo.

Sendo assim, um mapa baseado nas peculiaridades pré-mapeadas da sequência de referência HXB2 [Li et al. 2015] foi criado. Este mapa contém informações sobre as faculdades pré identificadas, além da informação da localização desta no genoma completo. De forma a considerar os atributos genotípicos apresentadas pelo vírus, é necessário realizar o processo de alinhamento entre a sequência *query* e a sequência de referência, para selecionar identificar as regiões pareáveis e sua cobertura.

A técnica de alinhamento exaustivo descrita inicialmente por Needleman e Wunsch em 1970 [Needleman e Wunsch 1970] foi implementada no software para utilizar as sequências disponíveis na base de dados. De forma a considerar o tamanho médio das sequências em relação à utilização de recursos de máquina, o que influencia diretamente no desempenho da técnica.

Desta forma, no processo de alinhamento apenas uma quantidade de sequências *query* são obtidas do banco de dados por vez. Assim o processo fica livre de uma sobrecarga de informações desnecessária. O processo é baseado em uma busca paginada, onde toda a tabela no banco é bloqueada, evitando que o processo sofra interferências externas.

Todo o processo ocorre de forma que mais de um alinhamento possa acontecer simultaneamente reduzindo assim o tempo total do mapeamento. Para isso, é necessário dividir as tarefas de alinhamento em *threads*. Porém, um limite na quantidade de

threads em execução deve ser aplicado, resguardado a capacidade física da máquina servidora.

Desta forma, foi utilizado o conceito de pool de *threads* que limita e controla as tarefas simultâneas em execução, além de manter, atualizar e controlar uma fila de execução. Este controlador de execução foi implementado de forma a limitar a quantidade de tarefas a um valor menor que o tamanho da "página"de sequências obtidas na base de dados. Desta forma, o desempenho da aplicação durante o processo se adapta aos recursos oferecidos pela máquina servidora, variando o tempo de execução de acordo com os recursos disponíveis.

O processo de preparação e mapeamento das sequências a partir da sequência de referência do organismo está descrita em forma de fluxograma na Figura 3.1. Onde o sub fluxo "Executa Tarefas de Alinhamento" representa o a tarefa mapeamento de uma única sequência.

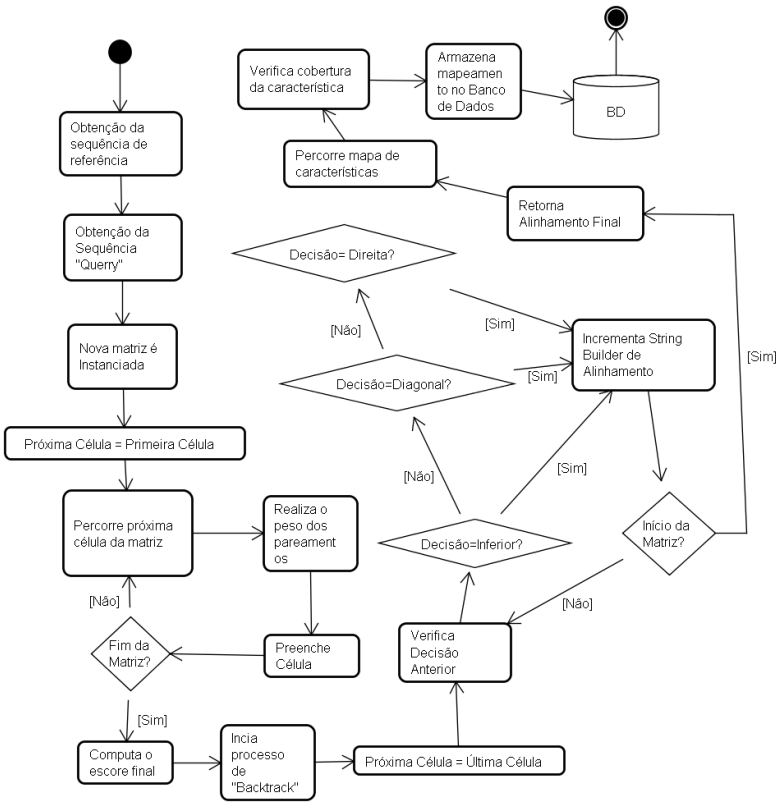

Figura 3.1: Sub fluxo de mapeamento das sequências do HIV. Fonte: Próprio Autor

Cada *thread* criada e inseria do pool representa uma tarefa de mapeamento. Esta tarefa é responsável pelo processamento do alinhamento entre a sequência não mapeada e a sequência de referência do organismo. Ademais de ser responsável pelo processo de armazenamento do resultado final na base de dados. Estas tarefas devem ocorrer até que não hajam mais sequências não mapeadas na base de dados.

A abordagem utilizada na implementação do algoritmo para o mapeamento das sequências foi baseada em desempenho e acurácia. A acurácia foi mantida em 100% enquanto o desempenho foi variado durante o processo de desenvolvimento. Para tanto, o algoritmo foi intensamente testado e modificado, utilizando um modelo de desenvolvimento incremental.

Por fim, o processo de alinhamento foi levemente alterado para que a técnica utilize menos recursos de máquina, e implemente alguns padrões de projeto. Isso facilita a reutilização de espaço em memória e reduzindo o tempo de processamento para cada alinhamento. Além de aumentar a manutenibilidade do software em questão.

Ao observar o processo de alinhamento global, é possível constatar que o processo de alinhamento dura em torno de 47% do tempo do mapeamento. Enquanto o processo de *"backtrack"*, que consiste na construção das *strings* de alinhamento a partir da matriz gerada pelo alinhamento, consome os 53% do tempo restante. O processo de construção das *strings* foi otimizado com alguns recursos disponíveis na própria linguagem como a utilização de *"StringBuilders"*, Iteração paralela e etc. Reduzindo assim o tempo total do alinhamento.

A tarefa responsável pelo alinhamento da sequência *query* e do posterior armazenamento do resultado está representado na Figura 3.2 de forma que cada sequência não mapeada da base de dados deve passar pro este procedimento. O resultado foi então armazenado na base de dados de forma a respeitar o limite de conexões ativas e a concorrência entre transações.

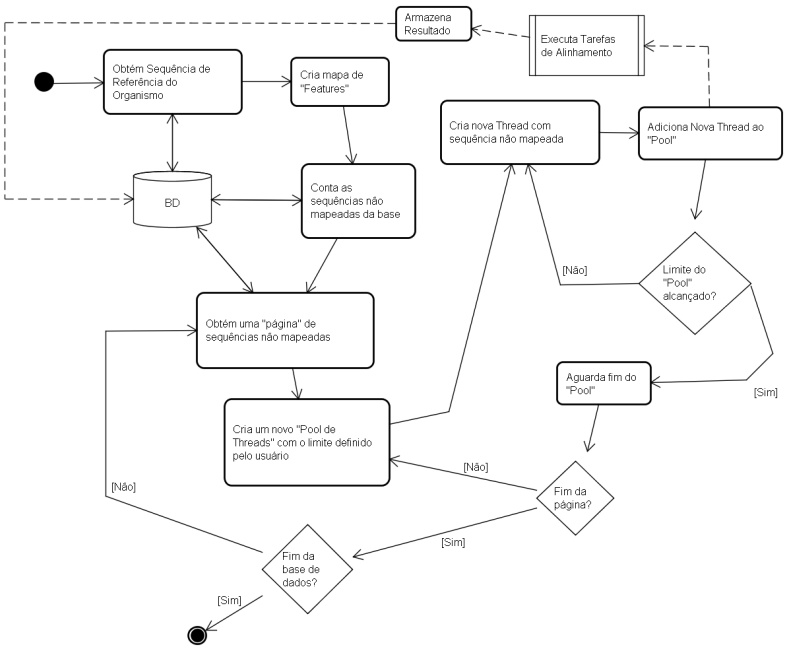

Figura 3.2: Processo principal de mapeamento das sequências do HIV. Fonte: Próprio Autor

Por se tratar de uma técnica de alinhamento não heurística, é necessária a criação de uma matriz de caminhos(possíveis alinhamentos). Essa matriz é baseada nos nucleotídeos disponíveis das sequências a serem alinhadas. Desta forma, o tempo de processamento e a utilização de memória permanecem proporcionais à quantidade de nucleotídeos das sequências. Por tanto a complexidade do pareamento permanece de ordem quadrática.

Durante o processo de alinhamento, a similaridade entre as duas sequências em questão é calculada a partir dos *matches*, dos *mismatches* e dos *gaps*. Estes componentes são representados em uma matriz de similaridade gerada a partir das sequências. Nesta matriz são atribuídos a cada um, valores e computado no resultado final. Este escore possibilita uma futura curagem dos dados a partir da similaridade entre as sequências descritas como pertencendo ao organismo e aquelas que realmente pertencem.

O escore gerado representa a maximização das semelhanças entre as sequências em questão. Estas sequências são representadas pelos possíveis caminhos que percorrem a matriz e assim formam alinhamentos. O total então é calculado através do somatório do peso de cada decisão. Estas decisões são geradas a partir de um pareamento de duas posições da matriz e a consecutiva função de pareamento. As decisões podem ser representadas da seguinte forma:

$$f(x_{k+1}, u_k) = \begin{cases} d(s_1(x^r_{k+1}), s_2(x^c_{k+1})) \text{ if } u_k = \searrow \\ d(-, s_2(x^c_{k+1})) \text{ if } u_k = \rightarrow \\ d(s_1(x^r_{k+1}), -) \text{ if } u_k = \downarrow \end{cases}$$

Após implementado, o algoritmo foi adaptado para utilizar os dados já obtidos de forma que diversos alinhamentos pudessem ocorrer ao mesmo tempo. Por se tratarem de processos que tratam de *I/O bound's*, as leituras e escritas no banco de dados tiveram que ser limitadas por um processo de paginação da informação.

Este procedimento foi então aplicado para criar um conjunto de informações relevantes a respeito da posição das sequências em relação ao genoma completo. Isso possibilita a posterior utilização deste mapa para a montagem de um conjunto de treinamento para um possível preditor de epítopos.

3.4 Subtipagem do Conjunto de Dados

Para obter informações consistentes sobre o comportamento mutacional e a heterogeneidade genotípica do organismo em questão, é necessário conhecer suas características genotípicas e classificá-las entre os grupos pré determinados. O HIV-1 possui 9 subtipos puros e cerca de 70 subtipos relacionados, fruto de recombinações [Hemelaar 2012, Los Alamos National Laboratory 2015a].

Tal variabilidade genotípica contribui diretamente para a prevalência do agente etiológico [Abidi et al. 2014], uma vez que o organismo apresenta uma adaptabilidade considerável. Isso dificulta a identificação de uma estratégia de combate eficaz [Cohen e Dolin 2013]. Estudos demonstram que as cepas podem variar no nível de aminoácidos em ate 42% [Hemelaar 2012].

Para obter informações sobre a disposição genotípica do agente etiológico, é necessário classificar as sequências disponíveis no conjunto de dados entre os grupos previamente determinados pela comunidade científica que podem ser observados juntamente com suas respectivas interconexões e derivações na Figura 3.3 . Desta forma, é necessário realizar o processo de subtipagem das sequências.

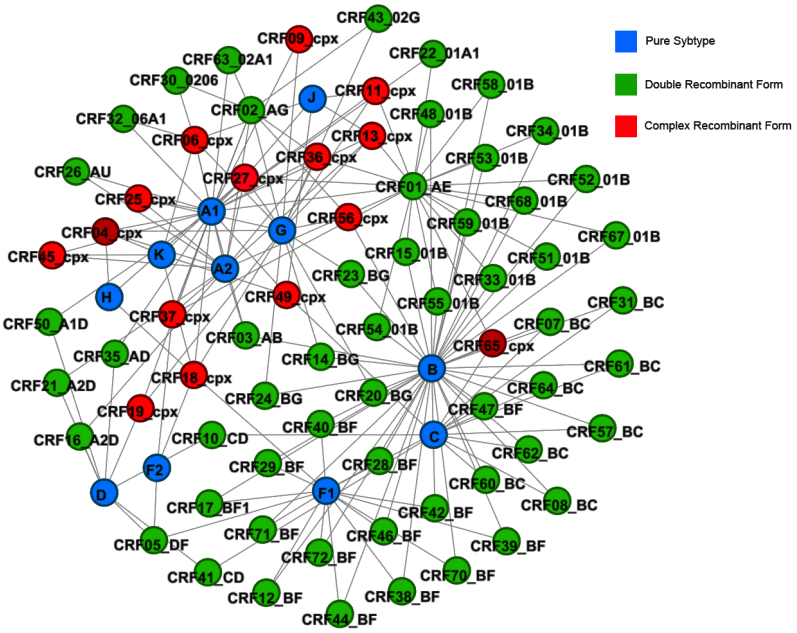

Figura 3.3: Subtipos do HIV-1 e suas respectivas derivações. Fonte: Próprio Autor

Para o processo de subtipagem das sequências são comparadas as sequências disponíveis no conjunto de dados com as sequências de referência dos subtipos identificados pela comunidade cientifica. Não obstante, é necessário observar o tamanho do conjunto de dados disponível (cerca de 580.000 sequências) em relação a quantidade de subtipos e formas recombinantes. Por conseguinte, para que se possa realizar a análise e classificação dos dados em tempo hábil, é necessário o desenvolvimento de uma estratégia de comparações que evite alinhamentos desnecessários e portanto, o desperdício de tempo de análise.

A estratégia desenvolvida leva em consideração a estrutura genotípica do organismo, suas formas recombinantes e seus híbridos. Assim, é levado em consideração o tamanho das sequências de referência de cada subtipo, em relação ao score máximo possível, aplicando a estratégia de desistência sumária caso este valor não possa ser alcançado.

Além de organizar os dados por tamanho durante o processo de estruturação dos dados, as inter-relações entre os diversos grupos são levados em consideração, no intuito de reduzir a quantidade de alinhamentos por sequência da base de dados, acelerando assim o processo de análise.

Este processo de preparação dos dados antes do processo de análise é feito baseado nas derivações de recombinação, onde um determinado recombinante é derivado de um ou mais subtipos puros a partir da coinfecção dos mesmos. Se considerarmos o subtipo puro como a centroide de um agrupamento é possível identificar os alinhamentos mais prováveis, os menos prováveis e os dispensáveis. Estes dispensáveis provém à análise uma maior agilidade e praticidade, evitando alinhamentos desnecessários.

Ademais, é necessário observar que os procedimentos de alinhamento e classificação das diversas sequências disponíveis no conjunto de dados não são dependentes entre si. Dessa maneira é possível a paralelização dos procedimentos, conferindo agilidade no processo de classificação mesmo com um conjunto de dados extenso. Assim, fluxo de trabalho da classificação das sequências nos nichos genéticos pode ser visualizado na Figura 3.4 em forma de fluxograma.

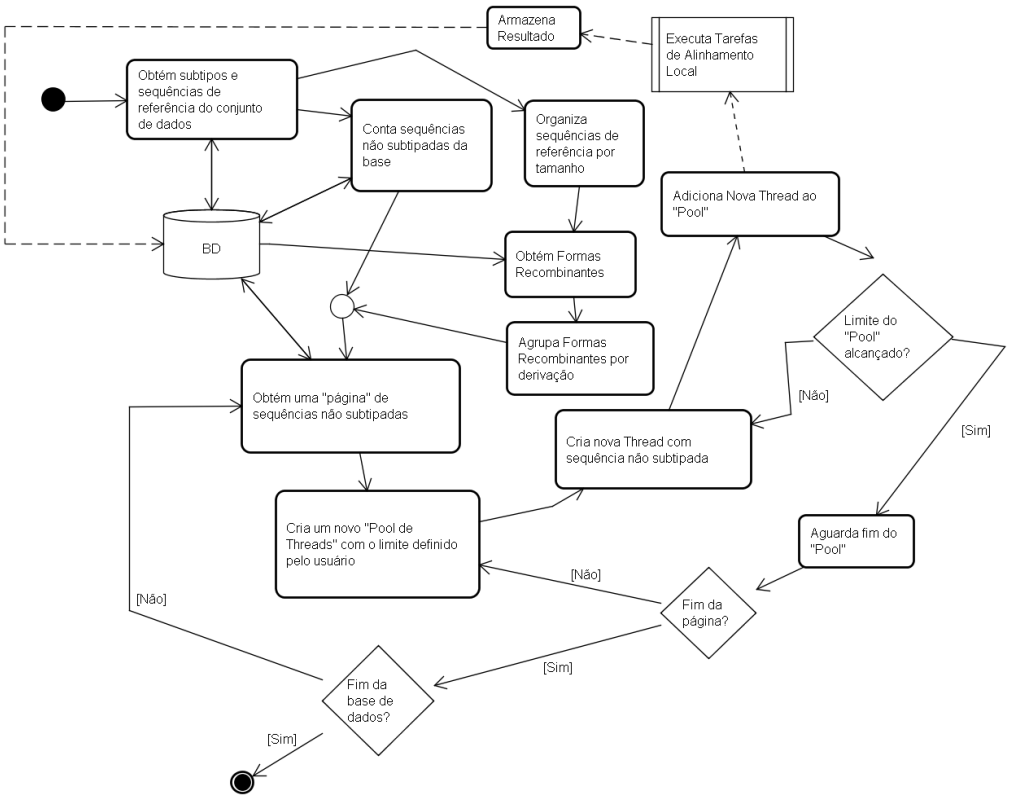

Figura 3.4: Fluxo de subtipagem de sequências Fonte: Próprio Autor

Para criar um conjunto de dados dinâmico que melhor represente o quadro do organismo no banco de dados, se faz necessário utilizar algoritmos ótimos (ou exaustivos)

para o alinhamento. Os algoritmos considerados ótimos são aqueles que retornam o melhor alinhamento possível entre duas sequências nucleotídicas enquanto os algoritmos heurísticos retornam um dos melhores. A abordagem heurística, apesar de mais rápida, é diretamente dependente da disposição dos dados e pode vir a comprometer a qualidade do alinhamento [Chakraborty e Bandyopadhyay 2013] além do conjunto final de dados.

Esta abordagem é necessária pois o organismo é heterogêneo genotipicamente. Dessa forma, quanto mais acurácia o alinhamento tiver, com mais certeza é possível organizar os dados nos nichos genéticos apresentados pelo organismo. Assim, maior será a qualidade do conjunto de dados e as informações geradas a partir deste.

O Algoritmo aplicado no processo de alinhamento local das sequências, baseado no algoritmo proposto por Smith Waterman [Smith, T. F.; Waterman 1981] pode ser visualizado em forma de fluxograma na Figura 3.5.

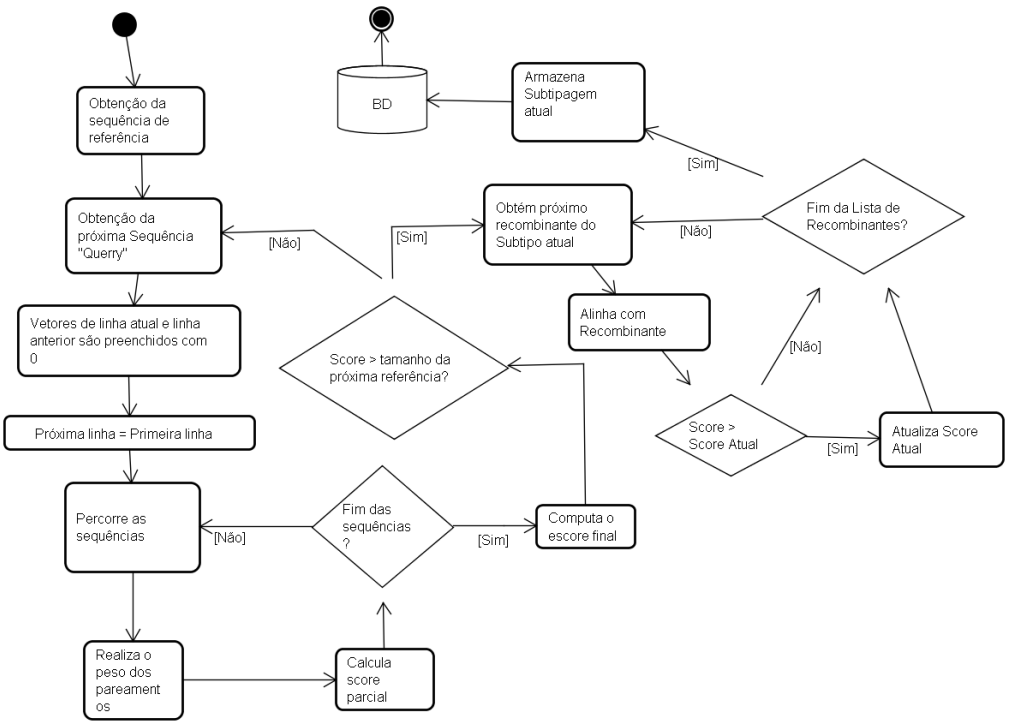

Figura 3.5: Sub fluxo de subtipagem de sequências Fonte: Próprio Autor

O processo de subtipagem ocorre quando a sequência *query* é plotada em uma matriz juntamente com sequência de referência. Nesta matriz serão preenchidos os escores para cada possibilidade de alinhamento. Posteriormente esta matriz é completamente percorrida no intuito de identificar a área de melhor pareamento entre as duas sequências.

Por fim, é realizado o processo de *backtracking* para identificar que o alinhamento seja construído em forma de cadeia de caracteres. Cada um destes caracteres representa

um pareamento positivo (*match*) negativo (*missmatch*) ou um *gap*.

Porém, o algoritmo como definido inicialmente pode vir a atrasar o processo de alinhamento, pois as técnicas exaustivas são mais demoradas e necessitam de mais recursos de máquina. Sendo assim, novas estratégias foram introduzidas no algoritmo. Foram removidos procedimentos desnecessários à subtipagem. Além disso, a utilização dos recursos de máquina e tempo de processamento foram otimizados.

Levando em consideração que o processo de *backtracking* não fornece informações relevantes para a classificação da sequência, e este representa cerca de 57% do tempo de processamento do algoritmo, este foi removido do *workflow* para fornecer mais performance. Outro processo moroso é o processo de alocação e leitura da matriz.

Portanto, este processo foi analisado e modificado, no intuito de otimizar os recursos de máquina e o tempo de processamento do algoritmo. O processo de montagem do score de similaridade entre as duas sequências ocorre então em um único laço, onde dois vetores dinâmicos representando a linha atual da matriz e a linha anterior, necessárias para o cálculo do score atual. Cada interação do laço a linha anterior passa a ser a atual. A atual é limpa e o score parcial é computado. Ao final do laço principal, o score do alinhamento local exaustivo já foi calculado e armazenado.

Após o cálculo de todos os escores parciais para a sequência em questão, é selecionado o maior entre estes. O maior escore representa a maior similaridade entre um determinado grupo de derivação. A partir deste escore são realizados os alinhamentos com todos os membros do grupo de maior similaridade. Por fim, é possível identificar o mair coeficiente de similaridade entre a sequência em questão e a sequência de referência do subtipo ao qual esta é dita a pertencer.

3.5 Disponibilização Pública dos Dados

Os dados resultantes da análise serão disponibilizados publicamente no portal do VSDBM, hospedado em http://vsdbm.tk e posteriormente em http://vsdbmsa.tk, juntamente com o código fonte e executável do software implementado. Além disto, um acesso remoto aos dados vai ser garantido através de um usuário público no banco de dados, que poderá acessar os dados com permissão apenas de leitura.

A criação do usuário de acesso avançado ao conjunto de dados será feito mediante contato prévio com a equipe responsável pela manutenção dos servidores, respeitando as regras e os prazos da fundação que hospeda os servidores. O banco de dados de suporte à análise foi extraído, comprimido e disponibilizado para download no portal, possibilitando assim que novas análises possam ser feitas, e novos dados possam ser adicionados a este conjunto, tornando o mesmo ainda mais rico e completo.

Capítulo 4

Resultados

4.1 Software

Ao final do processo de desenvolvimento, é apresentado um software otimizado para análise de sequências nucleotídicas virais. Este software é otimizado e preparado para tratar de organismos com comportamento mutacional e heterogeneidade genotípica similares ao HIV, a exemplo do HCV (*Hepatitis C Virus*).

Além do software, foi disponibilizado no portal da plataforma VSDBM, o modelo final de dados, contendo o script de criação de um novo banco de dados, além do backup inicial. Para que novos estudos possam ser realizados acerca do HIV-1 ou de outros organismos.

4.2 Conjunto de dados

O conjunto de dados disponibilizado ao final dos processos, contém informações relevantes e atualizadas sobre do agente etiológico em questão. Este conjunto de dados possui no total 255GB de informação, o banco de dados conta com informações acerca do mapeamento das sequências em relação ao genoma completo. Além de duas estruturas de classificação, a subtipagem clássica e a subtipagem utilizando a técnica de agrupamento por derivação de recombinação.

As informações contidas neste conjunto de dados poderão auxiliar diretamente no desenvolvimento de tratamentos eficazes para a patologia associada ao organismo em questão. Pois estas agregam informações sobre a dinâmica viral, disposição genotípica e ainda validam a produção de analisadores de padrão, classificadores e preditores de epítopos.

4.3 Mapeamento de Sequências

Para conhecer melhor a interação e a dinâmica deste agente etiológico, uma vez que se observa que a ausência de resposta imune do hospedeiro, se dá principalmente por mutações de escape. Estas mutações são mutações em regiões que codificam epítopos [Roider et al. 2014]. Por tanto, seria relevante identificar principalmente os epítopos que são reconhecidos pelo sistema imune, e sua prevalência mutacional.

No intuito de identificar a prevalência de tais estruturas é necessário observar a densidade da representação das regiões do genoma do agente etiológico no conjunto de dados. Uma vez que cada estrutura possui uma função diferente que influi diretamente na dinâmica da infecção. Por conseguinte, será possível identificar a prevalência das estruturas pré-identificadas do agente etiológico no conjunto de dados.

A identificação da prevalência das regiões do genoma do vírus é relevante para o aprestamento de diversos estudos acerca da pandemia. Visto que é necessário identificar um *dataset* para a realização dos estudos e este processo parte da obtenção das sequências que codificam as áreas ”alvo”dos estudos.

Assim, o processo de mapeamento ocorreu de forma que as coordenadas de cada alinhamento foram armazenadas, para que posteriormente a prevalência de cada região possa ser obtida e observada. Não obstante, nem todo alinhamento tem ”cobertura”completa sobre a região observada, sendo assim necessária uma análise qualitativa da ”cobertura”do alinhamento em relação à região.

Os dados foram então coletados de forma agrupada, onde os alinhamentos que representam de forma completa a região observada foram classificados como total e aqueles que representam apenas uma fração da área como parcial. Podemo observar na Figura 4.1 a densidade de cobertura em cada região do genoma completo, levando em consideração o mapa criado usando as coordenadas do genoma completo.

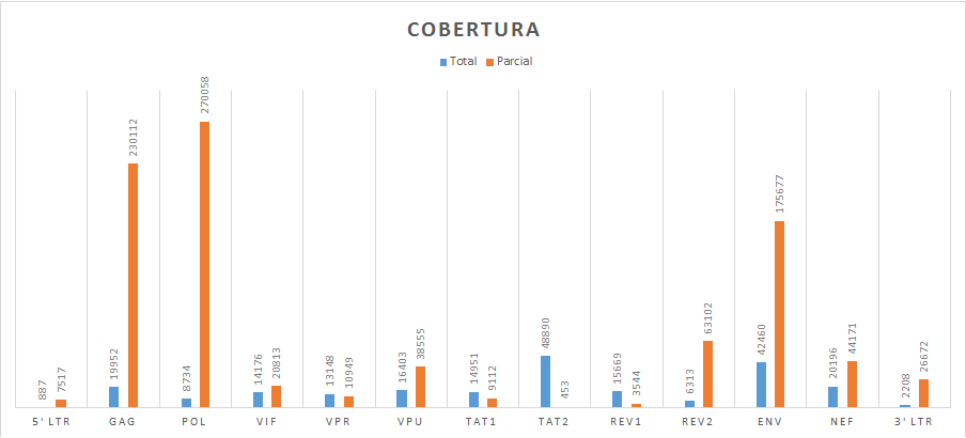

Figura 4.1: Mapa de densidade das regiões do genoma completo do HIV. Fonte: Próprio Autor

4.4 Subtipagem das Sequências

Para obter informações relevantes a respeito da heterogeneidade genotípica do agente etiológico presente no conjunto de dados, foram aplicadas técnicas de alinhamento exaustivo em duas metodologias de classificação. Levando ou não em consideração a derivação de recombinação, no intuito de reduzir a quantidade de alinhamentos necessários e assim o custo computacional do processo.

A metodologia clássica de classificação compara todas as sequências com todos os subtipos disponíveis, levou cerca de dois dias e cinco horas. Foram realizados em torno de 40 milhões de alinhamentos durante esse período. Já a metodologia otimizada reduziu esse tempo e quantidade de alinhamentos em cerca de 50%. Desta forma, terminando todo o processo de classificação dos dados em apenas um dia e sete horas de processamento.

Entretanto, foi necessário validar a acurácia das técnicas de agrupamento por derivação de recombinação e desistência sumária por tamanho da referência aplicadas à metodologia otimizada. Assim, foram criadas duas bases de dados cada uma contendo os resultados de uma metodologia de mapeamento, e posteriormente foi realizado um teste de semelhança sequência por sequência.

Este teste de semelhança gerou a matriz de confusão que pode ser visualizada na Figura 4.2, onde pode ser observado que independente do subtipo ou forma recombinante. Todas as sequências comparadas foram classificadas da mesma forma nas duas metodologias. O que significa que o nível de similaridade da técnica de agrupamento e desistência sumária é de 100% para os subtipos observados.

	A1	A2	B	C	D	F1	F2	G	H	J	K	CRFs
A1	15116	0	0	0	0	0	0	0	0	0	0	0
A2	0	1282	0	0	0	0	0	0	0	0	0	0
B	0	0	258942	0	0	0	0	0	0	0	0	0
C	0	0	0	21813	0	0	0	0	0	0	0	0
D	0	0	0	0	8792	0	0	0	0	0	0	0
F1	0	0	0	0	0	3286	0	0	0	0	0	0
F2	0	0	0	0	0	0	1945	0	0	0	0	0
G	0	0	0	0	0	0	0	1465	0	0	0	0
H	0	0	0	0	0	0	0	0	987	0	0	0
J	0	0	0	0	0	0	0	0	0	569	0	0
K	0	0	0	0	0	0	0	0	0	0	1184	0
CRFs	0	0	0	0	0	0	0	0	0	0	0	226023

Figura 4.2: Matriz de confusão entre a metodologia clássica de subtipagem, e a metodologia de agrupamento por derivação de recombinação. Fonte: Próprio Autor

Assim, é possível validar a técnica de agrupamento, uma vez que esta representa com exatidão a técnica exaustiva clássica de classificação, porém em um tempo 50%

menor de processamento. Desta forma, é possível preparar o conjunto de dados e o software para o crescimento recursivo do mesmo, uma vez que novas sequências vão sendo sequenciadas e disponibilizadas.

A partir da validação da metodologia adotada para a classificação das sequências, é necessário comparar os resultados obtidos com o cenário disponibilizado pelos bancos de dados biológicos especializados. É possível observar na Figura 4.3 o cenário apresentado pelo *dataset* em questão, gerado pelo software em relação ao *dataset* fornecido pelo *HIV Databases* onde é possível observar uma prevalência já esperada do subtipo B em ambos os casos, cerca de 50%, enquanto que os recombinantes representam uma disparidade de cerca de 20%.

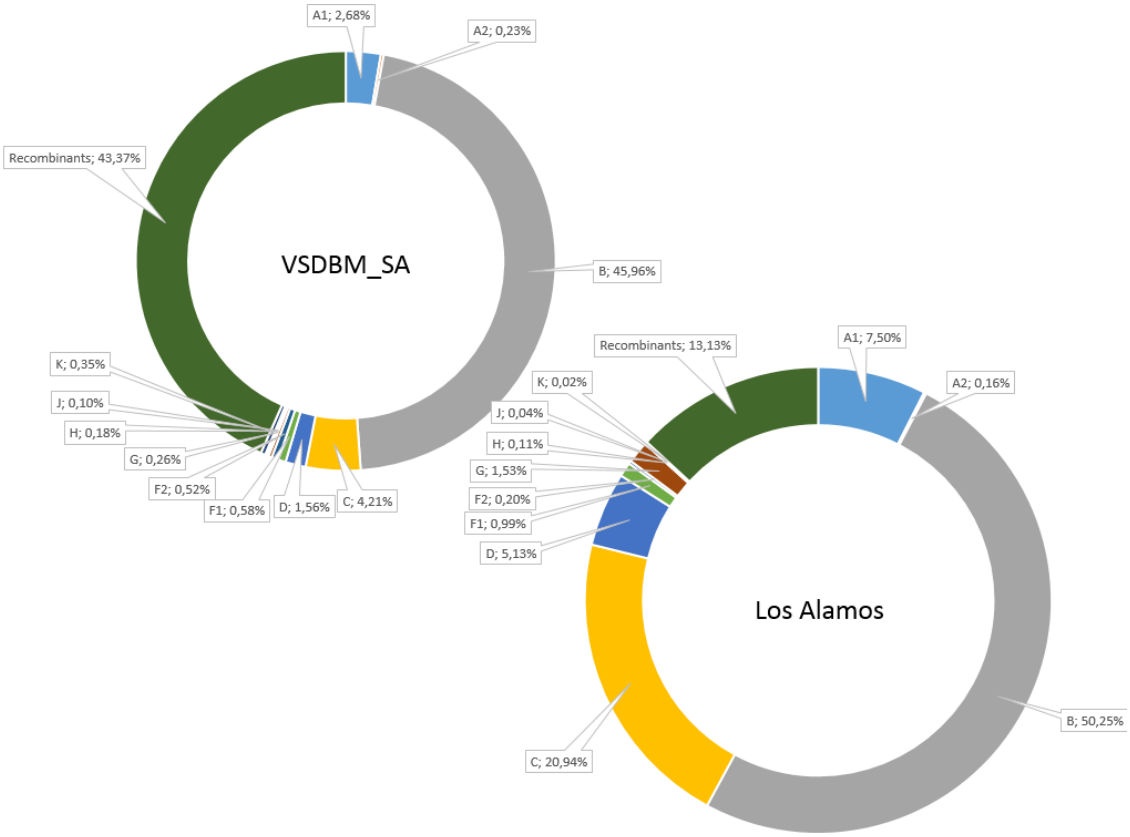

Figura 4.3: Prevalência dos subtipos no conjunto de dados gerado pelo *software*, e no conjunto de dados fornecido pelo *Los Alamos National Laboratory*. Fonte: Próprio Autor

Esta variação na similaridade pode ser explicada a partir da observação do processo de classificação aplicado pelo HIV *Databases*, onde este utiliza a classificação realizada pelo autor que sequenciou a amostra (LOS ALAMOS NATIONAL LABORATORY, 2015c). Além disso, o banco de dados disponibilizado pelo Los Alamos possui cerca de 200.000 sequências a menos que o conjunto de dados utilizado pelo *software*.

Assim, pode-se constatar uma instabilidade no processo de classificação do banco, uma vez que autores diferentes, utilizam técnicas diferentes que variam o nível de acurácia entre si. Desta forma, na Figura 4.4 é plotada uma matriz de confusão que compara a classificação entre os dois conjuntos de dados classificados, neste caso considerando apenas subtipos puros. Desta comparação, podemos observar que 97% das sequências foram classificadas de forma semelhante nos dois *datasets*.

	A1	A2	B	C	D	F1	F2	G	H	J	K
A1	10058	10	4	23	13	0	0	2	2	1	0
A2	104	328	16	10	13	0	1	11	2	0	1
B	61	1	133783	232	1169	13	0	36	3	2	1
C	51	0	858	13593	16	2	0	18	1	0	0
D	22	0	53	23	6000	7	0	14	1	3	0
F1	11	0	50	18	1	1568	16	3	2	0	0
F2	4	0	241	7	70	34	570	17	3	0	0
G	8	0	44	8	18	0	1	555	0	0	0
H	5	0	75	117	6	0	0	23	262	1	0
J	12	0	6	165	7	0	0	5	1	79	0
K	11	1	437	23	50	11	0	10	1	0	38

Semelhança:	97,47%
Diferença:	2,53%

Figura 4.4: Matriz de confusão entre os dois *dataset's* em questão para subtipos puros. Fonte: Próprio Autor

Foi observado ainda que o coeficiente de semelhança entre os resultados tende a subir a medida que são removidos os fragmentos de sequência, que podem causar uma falha de identificação. Aplicando um limite de 500 pares de base, é possível verificar que a semelhança entre as sequências sobe para 99,61%.

Já na Figura 4.5 pode-se observar a mesma matriz de confusão incluindo as formas recombinantes. nesta matriz, a similaridade entre os resultados das classificações é de 61%. Isso ocorre devido a uma confusão já esperada entre os subtipos puros e algumas formas recombinantes, como o subtipo CRF57_BC. Este fato pode ser explicado se observadas as datas de identificação dos subtipos. O subtipo considerado puro "C"possui 28 anos de diferença em relação ao CRF57_BC, o que significa que todas as sequências classificadas neste período não poderiam ter sido classificadas como recombinantes, pois esta classe não existia.

Desta forma, se observa que o conjunto de dados gerado pelo software se trata de um conjunto de dados mais atualizado e concreto a respeito do agente etiológico em questão. O processo de atualização das informações acerca do organismo é relevante para a geração de conhecimento sobre a dinâmica viral. Possibilitando assim a realização de estudos mais precisos e a identificação de possíveis curas clínicas.

Ainda sobre a classificação realizada no conjunto de dados, pode-se observar na Figura 4.7 a relação entre todos os subtipos considerados puros e todas as formas recombinantes circulantes, separadas por *dataset*. Estes representam quase metade dos dados analisados no caso do conjunto de dados gerado pelo software, e apenas

	A1	A2	B	C	D	F1	F2	G	H	J	K	CRF´s
A1	10058	10	4	23	13	0	0	2	2	1	0	33
A2	104	328	16	10	13	0	1	11	2	0	1	33
B	61	1	133783	232	1169	13	0	36	3	2	1	546
C	51	0	858	13593	16	2	0	18	1	0	0	1458
D	22	0	53	23	6000	7	0	14	1	3	0	245
F1	11	0	50	18	1	1568	16	3	2	0	0	427
F2	4	0	241	7	70	34	570	17	3	0	0	98
G	8	0	44	8	18	0	1	555	0	0	0	285
H	5	0	75	117	6	0	0	23	262	1	0	59
J	12	0	6	165	7	0	0	5	1	79	0	41
K	11	1	437	23	50	11	0	10	1	0	38	162
CRF´s	14743	185	32349	55789	9717	1712	87	4416	66	50	10	33155

Semelhança:	61,19%
Diferença:	38,82%

Figura 4.5: Matriz de confusão entre os dois *dataset's* em questão incluindo subtipos recombinantes. Fonte: Próprio Autor

13% no conjunto de dados disponibilizado pelo Los Alamos. O que mostra uma tendência de recombinação, o que é esperado de um vírus com um característica mutacional tão ativa, e de heterogeneidade genotípica elevada.

Com a análise realizada pelo software, é possível ainda observar as prevalências das formas recombinantes de forma separada. É possível observar que ha prevalência dos subtipos derivados dos subtipos puros ”A”, ”B”, ”C”e ”F1”. Como pode ser observado na Figura 4.6, o subtipo CRF57_BC representa cerca de 10% de todo o montante de recombinantes no conjunto de dados. Enquanto o CRF03_AB sendo o mais representativo, apresenta prevalência de 13,61%.

Assim, é possível observar que o conjunto de dados gerado pelo software em questão, se encontra mais atualizado e, por tanto, mais completo. Além disso, o software conta com uma técnica unificada de análise e classificação dos dados. Esta estabilidade na análise torna os dados mais concretos e por conseguinte mais precisos sobre o organismo.

Os subtipos considerados recombinantes podem ser divididos em dois grupos principais, os recombinantes duplos ou *double*, os quais derivam da recombinação a partir da coinfecção de dois subtipos diferentes e os complexos ou *cpx*, que derivam de três ou mais subtipos circulantes. A relação de proporção de prevalência entre estes grupos de recombinantes pode ser observada ainda na Figura 4.7, onde os recombinantes complexos representam apenas cerca de 10% do montante total, por consequência da raridade deste tipo de coinfecção.

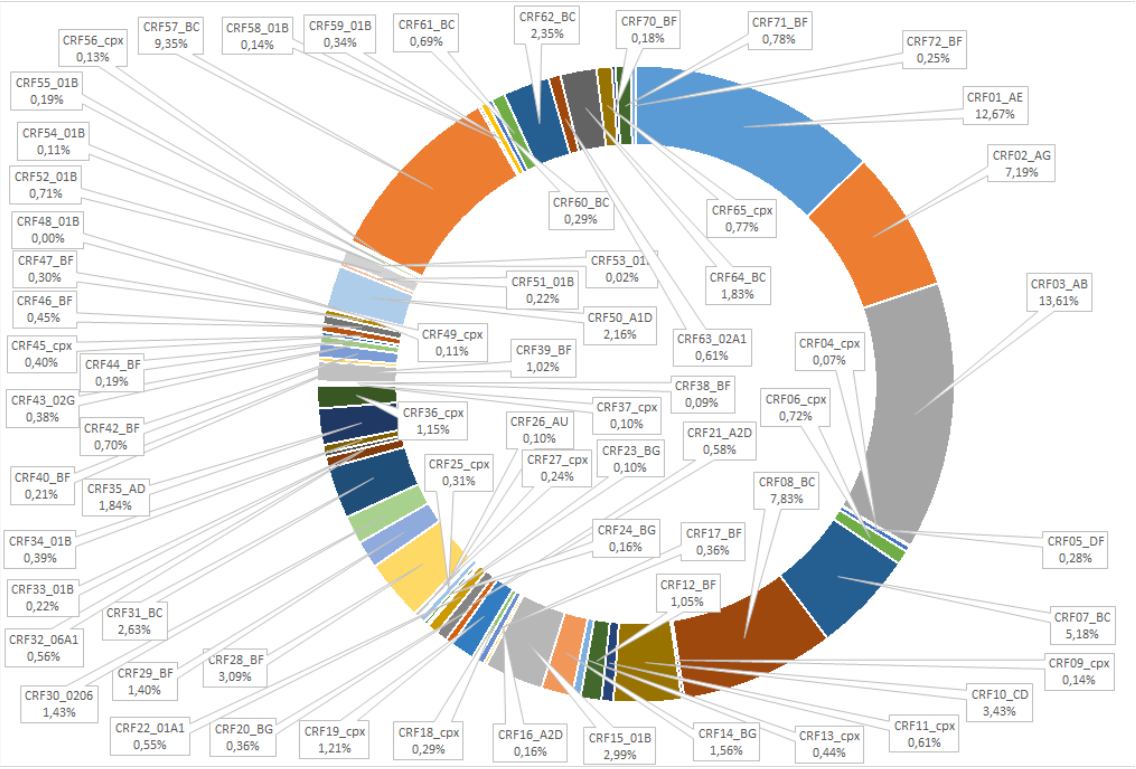

Figura 4.6: Distribuição da prevalência de cada forma recombinante presente no conjunto de dados. Fonte: Próprio Autor

4.5 Distribuição por país das sequências e análise dos subtipos

Ao obter todo o conjunto de dados disponível no GenBank, foi possível identificar os dados relativos ao país de coleta e submissão das sequências disponíveis. Este dado permite que seja observada a densidade de sequências em relação ao país de origem. Como mostra a Figura 4.8. Os Estados Unidos é o país com maior representatividade, sendo o *outlier*. Onde a maioria das sequências disponíveis no conjunto de dados não provém dos países que contém mais pessoas infectadas. Essa informação permite uma nova visão sobre a pandemia, permitindo novas inferências e análises.

Observando a densidade de sequências em relação ao país de origem, podem ser destacados os 10 países com maior número de sequências submetidas. Os Estados Unidos com 142.784 sequências, seguido pela África do sul, Uganda e Quênia com cerca de 21.000 sequências cada. A China vem representada com 17.859 sequências, seguida pelo Brasil(15.875), Canadá(14.193), Tailândia(14.098), França(13.717) e Japão(12.511). Destes países é notória a participação dos Estados Unidos e África do Sul. Onde os Estados unidos representa um notório polo de estudos em HIV. E a África do Sul, onde há uma pandemia generalizada [Cristina et al. 2012].

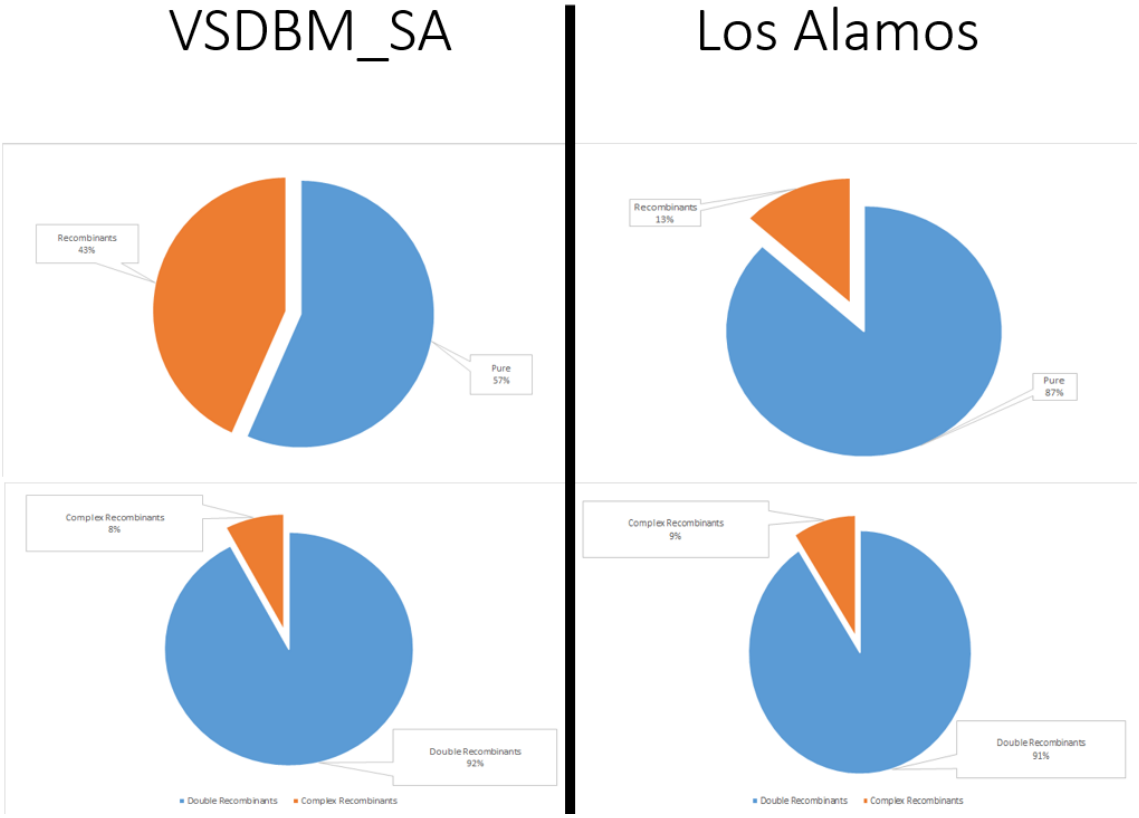

Figura 4.7: Relação entre subtipos puros e recombinantes, e tipos de recombinantes para ambos os *dataset's*. Fonte: Próprio Autor

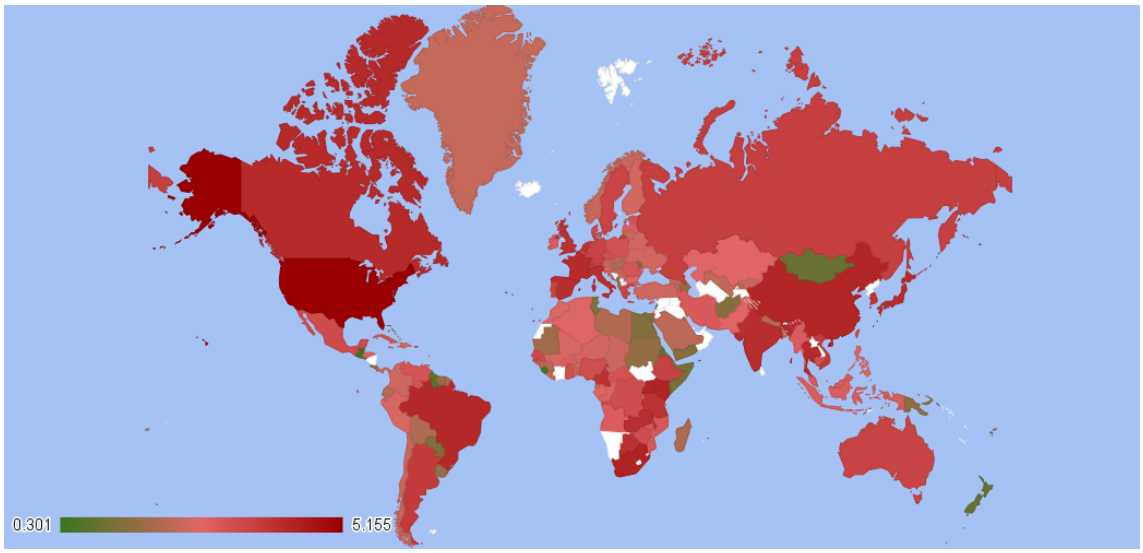

Figura 4.8: Representação do montante de submissão de sequências por país. Fonte: Próprio Autor

O software desenvolvido, permite ainda a análise da distribuição de subtipos filtrada pelo país de origem das sequências. Esta análise retorna a prevalência de cada subtipo no conjunto de sequências submetidas em um determinado país. Com isso é possível obter uma perspectiva diferente da pandemia para cada país. Esta perspectiva facilita o processo de manejo da epidemia.

Considerando a distribuição de subtipos aliada às sequência geo-referenciadas, é possível identificar diferenças na distribuição de subtipos nos países, assim como a preponderância relativa das formas recombinantes. Como observado na Figura 4.9 que representa a distribuição de subtipos das sequências associadas aos Estados Unidos.

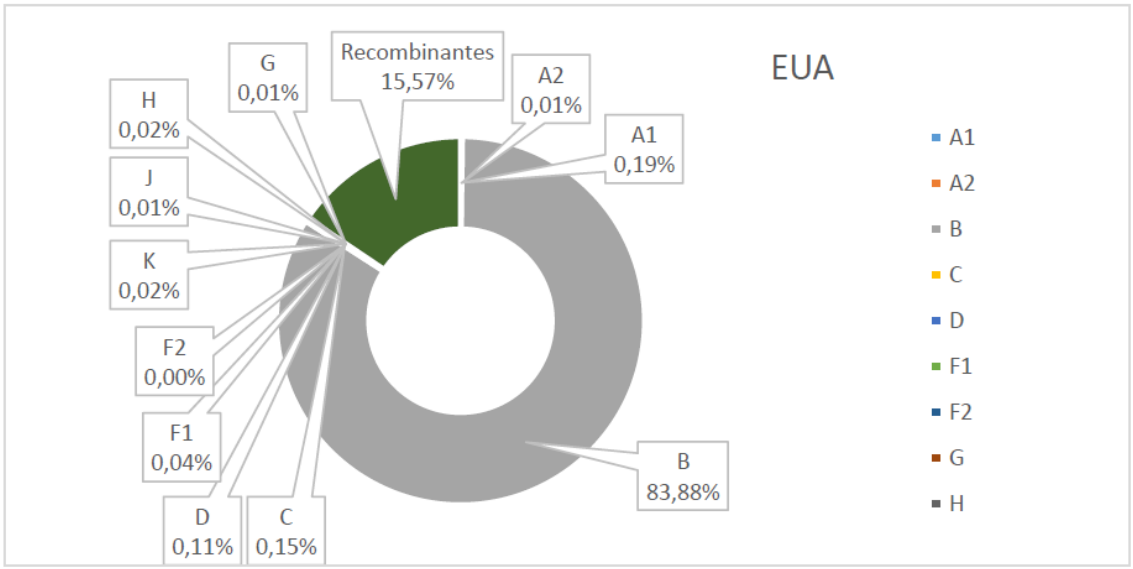

Figura 4.9: Distribuição de subtipos nos Estados Unidos. Fonte: Próprio Autor

Observa-se que neste caso, o subtipo B é mais prevalente assim como na distribuição global anteriormente discutida. Entretanto, este subtipo se mostra ainda mais predominante. Com 83,88% das sequências submetidas por este país sendo identificadas como pertencentes ao subtipo B.

Ainda sobre a distribuição de subtipos apresentada pelos Estados Unidos, é possível observar a prevalência de formas recombinantes nas sequências analisadas. Esta prevalência representa um montante menor que o identificado na distribuição global. Apresentando 15,57% de prevalência para formas recombinantes.

Não obstante, a distribuição apresentada pelos países do continente africano apresentam um cenário diferente. A Figura 4.10, representa a distribuição das sequências originadas da África do Sul. Nesta figura é possível observar que a distribuição de subtipos apresenta um predomínio de sequências associadas à formas recombinantes com 87,09

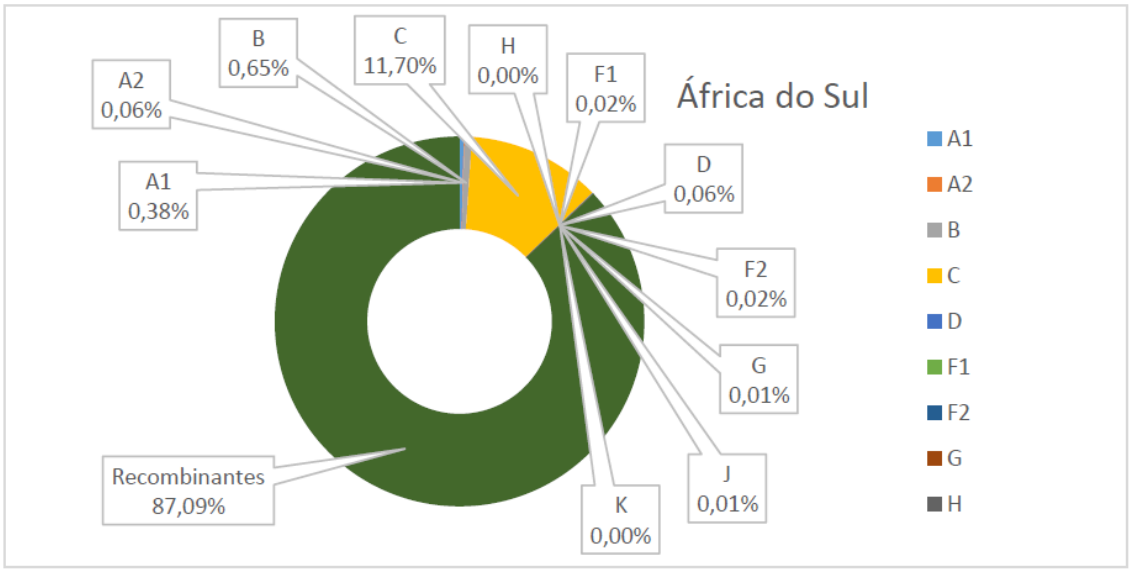

Figura 4.10: Distribuição de subtipos nos África do Sul. Fonte: Próprio Autor

Esta predominância de sequências classificadas como formas recombinantes pode ser causada pela densidade da população infectada. Neste país, a prevalência de infeção é tão alta que chegava a cerca de 6 milhões de infectados em 2012 [UNAIDS 2012]. Com isso, o processo de transmissão e consequente coinfecção do agente etiológico de subtipos diferentes se torna mais recorrente. Esta coinfecção gera novos genótipos virais, o que dificulta ainda mais o manejo da epidemia.

Entretanto, ainda sobre o continente africano, Uganda, representado na Figura 4.11 apresenta um panorama diferente. Este país também apresenta um predomínio de sequências associadas com formas recombinantes de 55,56%. Entretanto o cenário demonstra que os subtipos D e A1 também possuem relativo predomínio de 25,50% e 15,92% respectivamente.Assim como no Quênia (Figura 4.12), onde o subtipo A1 chega a ser 29,57% frequente.

O continente asiático é representado entre os 10 países com maior número de sequências no conjunto de dados por China e Japão. A china(Figura 4.13) apresenta um cenário semelhante em relação aos recombinantes. Onde as sequências associadas a estes representam 79,09% das sequências originadas deste país.

Diferente da China, o Japão, representado na Figura 4.14 exibe uma predominância de sequências associadas ao subtipo B de 73,27%. Demonstrando que apesar de possuírem características geográficas semelhantes, ambos demonstram distribuições inversamente distintas. Assim como o Japão, o Canadá(Figura 4.15) apresenta um panorama similar. Com predominância de sequências de subtipo B de 77,56%

Já a França, representada na Figura 4.16 apresenta um cenário mais proporcional. Nesta, as sequências associadas ao subtipo B apresentam predominância de 57,18%. Enquanto as sequências identificadas como formas recombinantes represen-

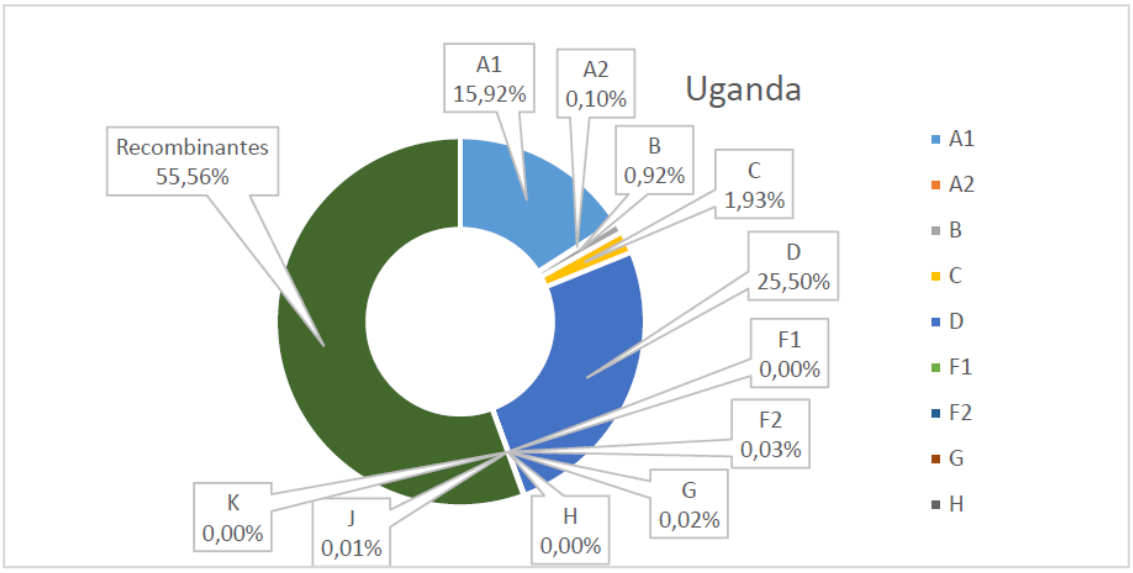

Figura 4.11: Distribuição de subtipos em Uganda. Fonte: Próprio Autor

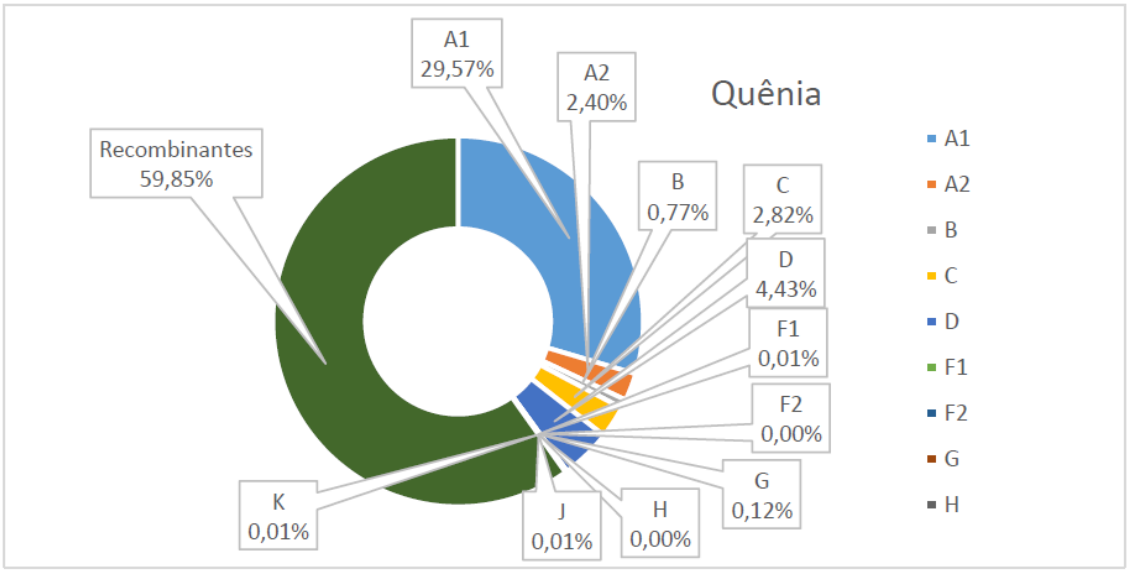

Figura 4.12: Distribuição de subtipos no Quênia. Fonte: Próprio Autor

tam 40,21%. Já a prevalência dos outros subtipos neste país não representa mais de 0.5% cada.

Representando a América do Sul, o Brasil, possui sua distribuição demonstrada na Figura 4.17. Nesta pode-se observar uma predominância de 59% de sequências associadas ao subtipo B. Enquanto as sequências consideradas formas recombinantes somam um total de 28%. Entretanto, os subtipos C e F1 são representados de forma consistente neste país com respectivamente 9% e 4% dos dados associados.

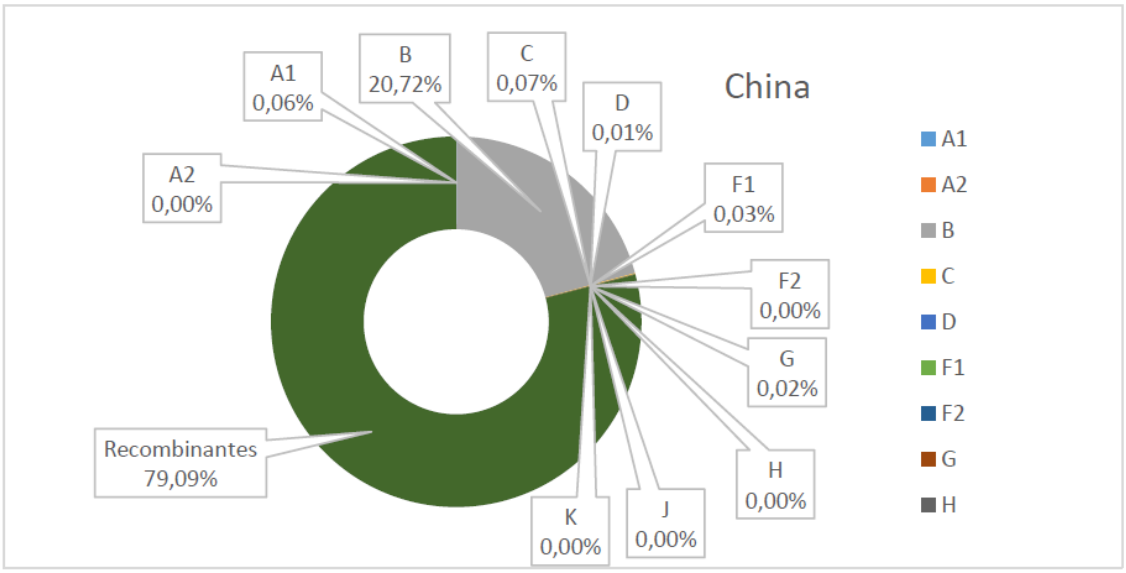

Figura 4.13: Distribuição de subtipos na China. Fonte: Próprio Autor

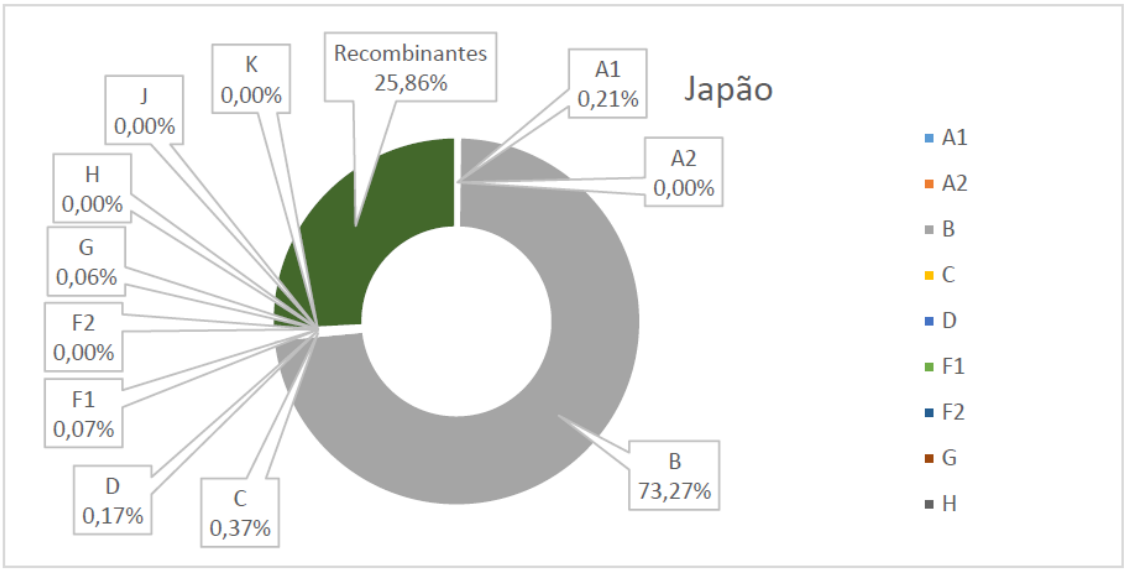

Figura 4.14: Distribuição de subtipos no Japão. Fonte: Próprio Autor

Entre 10 países que mais contribuíram com sequência, o cenário mais claro de predominância de recombinantes é na Tailândia (Figura 4.18), onde 94,1% dos dados subtipados são relacionados com formas recombinantes. Enquanto o subtipo B representa apenas cerca de 5%. Isso representa uma disparidade em relação ao cenário global.

Entretanto, é preciso levar em consideração a variável tempo. Alguns conjuntos de dados possuem um espaço de tempo maior. A exemplo dos Estados Unidos, onde existem sequências que datam desde a década de 80. Enquanto a sequência mais

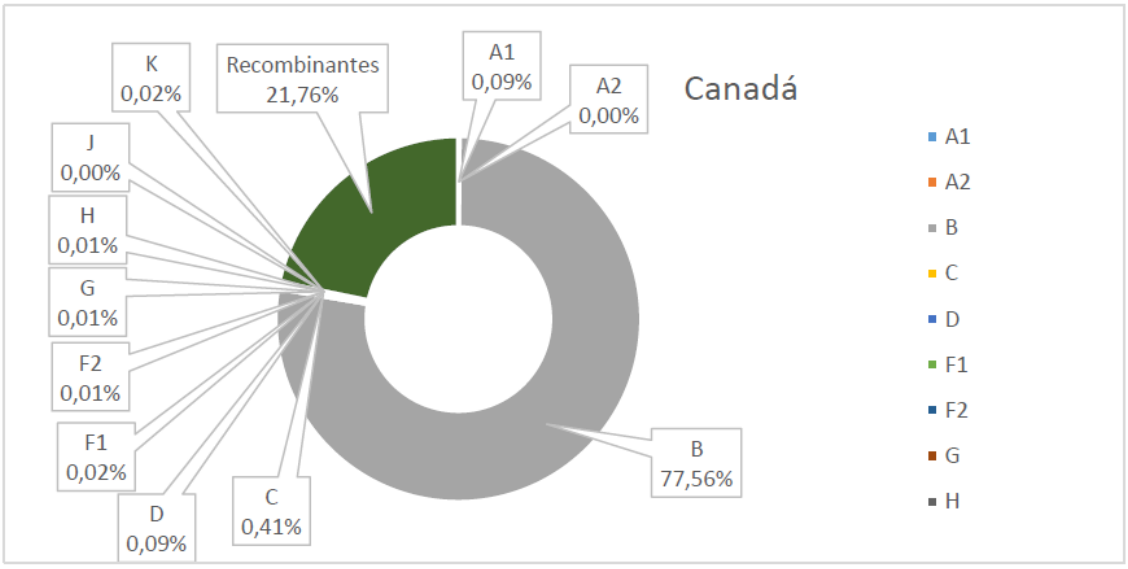

Figura 4.15: Distribuição de subtipos no Canadá. Fonte: Próprio Autor

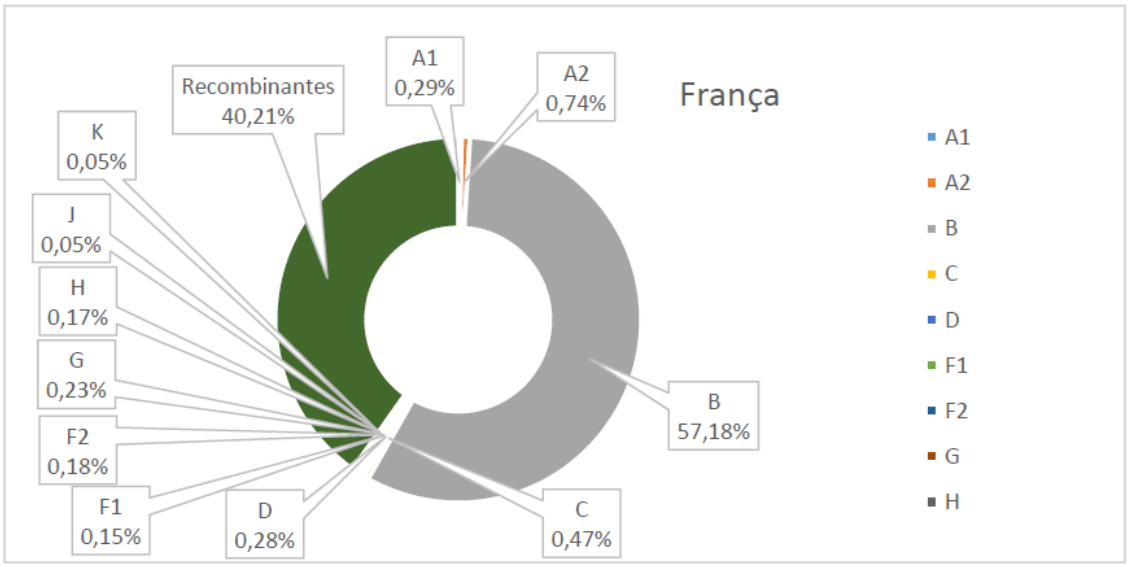

Figura 4.16: Distribuição de subtipos na França. Fonte: Próprio Autor

antiga associada ao Brasil data de 1994. Ou seja, tem muito mais tempo de estudo nos EUA do que no Brasil.

Desta forma, o cenário apresentado varia de acordo não só com a distribuição dos subtipos no conjunto de dados, mas da dinâmica da pandemia no país durante o período representado. Além da variação em relação a porcentagem da população infectada, o que influi diretamente para o crescimento da frequência das sequências associadas à formas recombinantes.

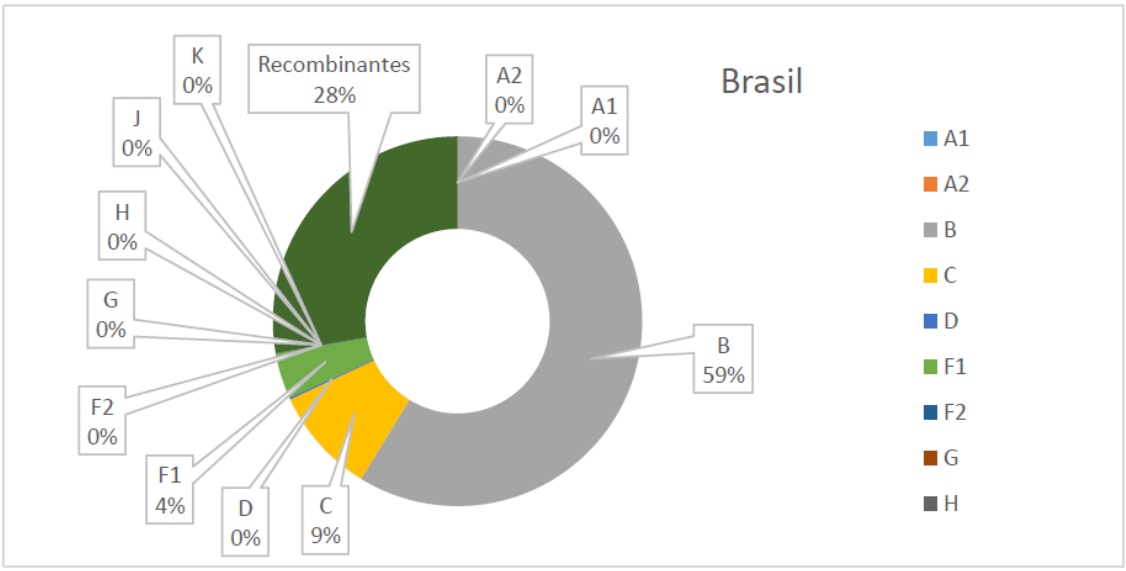

Figura 4.17: Distribuição de subtipos no Brasil. Fonte: Próprio Autor

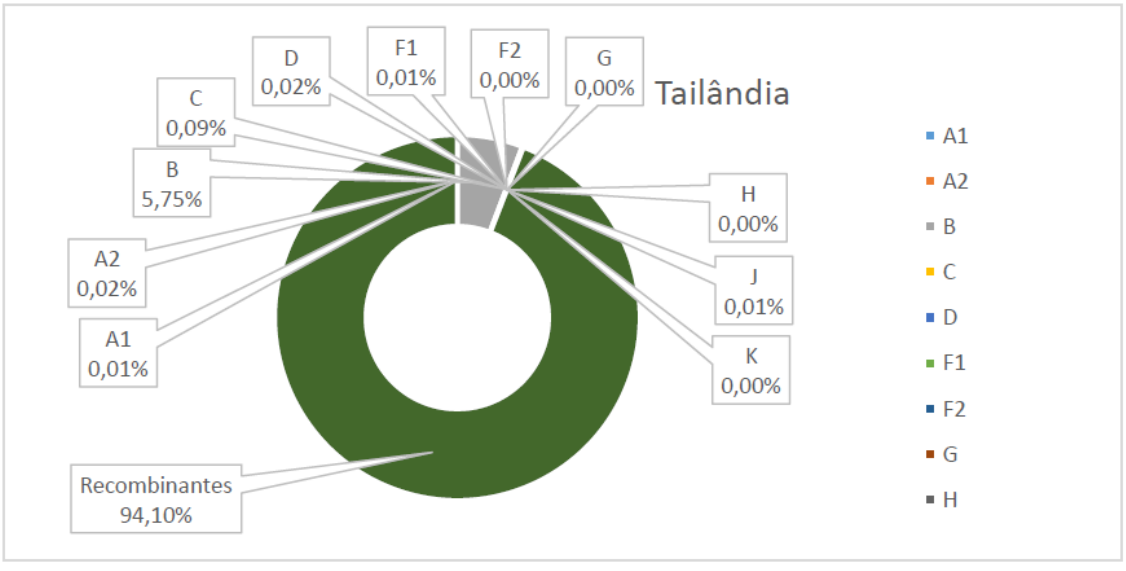

Figura 4.18: Distribuição de subtipos na Tailândia. Fonte: Próprio Autor

Capítulo 5

Considerações Finais

O software apresentado neste trabalho contribui diretamente para o avanço nos estudos genéticos e epidemiológicos do HIV. Este apresenta ferramentas intuitivas, simples e otimizadas para análise de vírus com taxas mutacionais elevadas. Sendo assim, está pronto para realizar análises em vírus como, por exemplo, o HCV. As análises realizadas pelo software foram desenvolvidas de forma que se adaptem às características do organismo submetido e às características e recursos computacionais da máquina servidora.

Além disso, as análises foram feitas de forma a manter o máximo de acurácia, mesmo com a constante modificação do conjunto de dados. Neste sentido, nenhuma ferramenta disponível hoje, realiza análise de um conjunto de dados tão grande, em tão pouco tempo com o alto nível de acurácia da ferramenta desenvolvida. Com as informações geradas, é possível verificar a evolução da epidemia até o cenário atual. De forma a permitir novas abordagens para a gestão e controle da doença.

Os resultados gerados auxiliam ainda na identificação de estruturas no genoma completo, de subtipos e de regiões genômicas imunogênicas. Estes fatores capacitam a construção de novos estudos. Desta forma, facilitam o desenvolvimento de vacinas eficazes e a possível cura. Além disso, é gerado com os resultados obtidos, um novo panorama da pandemia, organizando as informações de forma que a facilitar a vigilância da mesma. Com as informações geográficas é possível verificar a distribuição de subtipos em cada país, facilitando a restrição de escopo de estudos nestes. Desta forma, facilita o combate à doença associada ao agente etiológico.

Entretanto, com o constante crescimento do conjunto de dados, é necessário manter a ciclicidade das análises. Assim, o conjunto de dados gerado estará sempre o mais atualizado possível. Possibilitando novas inferências e novos estudos. Além da ciclicidade das análises, é preciso ainda manter a recursividade no processo de obtenção das novas sequências. Uma vez que com as tecnologias de sequenciamento de alta demanda, uma maior quantidade de sequências vem sendo submetida ao GenBank a cada ano.

Com os resultados deste projeto, novas possibilidades se abrem no concerne do estudo deste agente etiológico. Uma dessas possibilidades é a identificação e mapeamento de epítopos evolutivamente estáveis no genoma. Este processo facilitará a identificação de regiões alvo para intervenção de fármacos.

Uma possível predição destas estruturas poderá resultar na identificação de uma vacina de amplo espectro, essencial para o controle da pandemia. Desta forma, o sistema proposto contribui de forma significativa para o desenvolvimento do conhecimento sobre o agente etiológico em questão. Permitindo a geração de novas ferramentas de análise, facilitando o processo de criação de novos tratamentos, vacinas e uma possível cura.

Referências Bibliográficas

[Abidi et al. 2014] Abidi, S. H., Kalish, M. L., Abbas, F., Rowland-Jones, S., e Ali, S. (2014). HIV-1 subtype A Gag variability and epitope evolution. *PloS one*, 9(6):e93415.

[Barré-Sinoussi et al. 2013] Barré-Sinoussi, F., Ross, A. L., e Delfraissy, J.-F. (2013). Past, present and future: 30 years of HIV research. *Nature reviews. Microbiology*, 11:877–83.

[Barre-Sinoussi, F., J. C. Chermann 1983] Barre-Sinoussi, F., J. C. Chermann, E. A. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):868–71.

[Brander et al. 2014] Brander, C., Haynes, B. F., e Moore, J. P. (2014). HIV Molecular Immunology 2001 Editors. *HIV Molecular Immunology*.

[Castro-Nallar et al. 2012] Castro-Nallar, E., Pérez-Losada, M., Burton, G. F., e Crandall, K. a. (2012). The evolution of HIV: inferences using phylogenetics. *Molecular phylogenetics and evolution*, 62(2):777–92.

[Chakraborty e Bandyopadhyay 2013] Chakraborty, A. e Bandyopadhyay, S. (2013). FOGSAA: Fast Optimal Global Sequence Alignment Algorithm. *Scientific reports*, 3:1746.

[Chan et al. 2014] Chan, P. a., Reitsma, M. B., Delong, A., Boucek, B., Nunn, A., Salemi, M., e Kantor, R. (2014). Phylogenetic and geospatial evaluation of HIV-1 subtype diversity at the largest HIV center in Rhode Island. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, pp. 1–9.

[Coffin e Swanstrom 2013] Coffin, J. e Swanstrom, R. (2013). HIV pathogenesis: dynamics and genetics of viral populations and infected cells. *Cold Spring Harbor perspectives in medicine*, 3(1):a012526.

[Cohen e Dolin 2013] Cohen, Y. Z. e Dolin, R. (2013). Novel HIV vaccine strategies: overview and perspective. *Therapeutic advances in vaccines*, 1(3):99–112.

[Combe e Sanjuán 2014] Combe, M. e Sanjuán, R. (2014). Variation in RNA virus mutation rates across host cells. *PLoS pathogens*, 10(1):e1003855.

[Cristina et al. 2012] Cristina, A., Vieira, D. S., Head, J. F., Maria, I., Padez, A., e Casimiro, C. (2012). A epidemia de HIV / Aids e a ação do Estado . Diferenças entre Brasil , África do Sul e Moçambique. pp. 196–206.

[Crous et al. 2012] Crous, S., Shrestha, R. K., e Travers, S. A. (2012). Appraising the performance of genotyping tools in the prediction of coreceptor tropism in HIV-1 subtype C viruses. *BMC Infectious Diseases*, 12(1):1.

[Day 2010] Day, R.-F. (2010). Examining the validity of the Needleman?Wunsch algorithm in identifying decision strategy with eye-movement data. *Decision Support Systems*, 49(4):396–403.

[de Queiróz et al. 2011] de Queiróz, a. T. L., Maracaja-Coutinho, V., Jardim, a. C. G., Rahal, P., de Carvalho-Mello, I. M. V. G., e Matioli, S. R. (2011). Relation of pretreatment sequence diversity in NS5A region of HCV genotype 1 with immune response between pegylated-INF/ribavirin therapy outcomes. *Journal of viral hepatitis*, 18(2):142–8.

[Deshmukh e Kharat 2015] Deshmukh, K. B. e Kharat, M. U. (2015). Review on Retrieving Biological Sequence Alignment using Smith-Waterman Algorithm. (1):24–26.

[Fernández-Suárez et al. 2014] Fernández-Suárez, X. M., Rigden, D. J., e Galperin, M. Y. (2014). The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic Acids Research*, 42(December 2013):1–6.

[Foley et al. 2012] Foley, B., Apetrei, C., Mizrachi, I., Rambaut, A., Korber, B., Kuiken, C., Leitner, T., Hahn, B., Mullins, J., Wolinsky, S., Abfalterer, W., Dimitrijevic, M., Funkhouser, B., Hraber, P., Krishnamoorthy, M., Macke, J., Sharma, R., Szinger, J. J., e Yoon, H. (2012). HIV Sequence Compendium 2012 Editors. pp. LA–UR–12–24653.

[Gallo et al. 1983] Gallo, R., Sarin, P., e Gelmann, E. (1983). Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science*.

[Hahn et al. 2000] Hahn, B. H., Shaw, G. M., Cock, K. M. D., e Sharp, P. M. (2000). AIDS as a Zoonosis: Scienti c and Public Health Implications. *Science*, 287(January):607–614.

[Hemelaar 2012] Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends in molecular medicine*, 18(3):182–92.

[Henn et al. 2012] Henn, M. R., Boutwell, C. L., Charlebois, P., Lennon, N. J., Power, K. a., Macalalad, A. R., Berlin, A. M., Malboeuf, C. M., Ryan, E. M., Gnerre, S., Zody, M. C., Erlich, R. L., Green, L. M., Berical, A., Wang, Y., Casali, M., Streeck, H., Bloom, A. K., Dudek, T., Tully, D., Newman, R., Axten, K. L., Gladden, A. D., Battis, L., Kemper, M., Zeng, Q., Shea, T. P., Gujja, S., Zedlack, C., Gasser, O., Brander, C., Hess, C., Günthard, H. F., Brumme, Z. L.,

Brumme, C. J., Bazner, S., Rychert, J., Tinsley, J. P., Mayer, K. H., Rosenberg, E., Pereyra, F., Levin, J. Z., Young, S. K., Jessen, H., Altfeld, M., Birren, B. W., Walker, B. D., e Allen, T. M. (2012). Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS pathogens*, 8(3):e1002529.

[Ho et al. 1995] Ho, D. D., Neumann, a. U., Perelson, a. S., Chen, W., Leonard, J. M., e Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection.

[Irahe Kasprzykowski 2013] Irahe Kasprzykowski, F. G. T. (2013). VSDBM: Modelo Computacional para gerenciamento de sequências nucleotídicas virais. Master's thesis, Centro Universitário Jorge Amado, Brasil.

[Johannessen et al. 2011] Johannessen, A., Garrido, C., Zahonero, N., Naman, E., e de Mendoza, C. (2011). HIV-1 drug resistance testing from dried blood spots collected in rural Tanzania using the ViroSeq HIV-1 Genotyping System. *The Journal of antimicrobial chemotherapy*, 66(2):260–4.

[Li et al. 2015] Li, G., Piampongsant, S., Faria, N. R., Voet, A., Pineda-Peña, A.-C., Khouri, R., Lemey, P., Vandamme, A.-M., e Theys, K. (2015). An integrated map of HIV genome-wide variation from a population perspective. *Retrovirology*, 12.

[Los Alamos National Laboratory 2015a] Los Alamos National Laboratory (2015a). Hiv circulating recombinant forms (crfs). `http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html`. Acessado em: 15-05-2015.

[Los Alamos National Laboratory 2015b] Los Alamos National Laboratory (2015b). How hiv database classifies sequences. `http://www.hiv.lanl.gov/content/sequence/HelpDocs/classification.html`. Acessado em: 15-05-2015.

[Mardis 2011] Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203.

[McGovern et al. 2010] McGovern, R. a., Thielen, A., Mo, T., Dong, W., Woods, C. K., Chapman, D., Lewis, M., James, I., Heera, J., Valdez, H., e Harrigan, P. R. (2010). Population-based V3 genotypic tropism assay: a retrospective analysis using screening samples from the A4001029 and MOTIVATE studies. *AIDS (London, England)*, 24(16):2517–25.

[MINISTÉRIO DA SAÚDE 2012] MINISTÉRIO DA SAÚDE (2012). Boletim Epidemiológico AIDS e DST.

[Needleman e Wunsch 1970] Needleman, S. B. e Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–53.

[Neher e Leitner 2010] Neher, R. A. e Leitner, T. (2010). Recombination rate and selection strength in HIV intrapatient evolution. *PLoS Computational Biology*, 6.

[Pineda-Peña et al. 2013] Pineda-Peña, A.-C., Faria, N. R., Imbrechts, S., Libin, P., Abecasis, A. B., Deforche, K., Gómez-López, A., Camacho, R. J., de Oliveira, T., e Vandamme, A.-M. (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 : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 19(100):337–48.

[Polanski e Kimmel 2007] Polanski, A. e Kimmel, M. (2007). *Bioinformatics*. Springer.

[Pruesse et al. 2012] Pruesse, E., Peplies, J., e Glöckner, F. O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, 28(14):1823–1829.

[Roberts et al. 1988] Roberts, J. D., Bebenek, K., e Kunkel, T. a. (1988). The accuracy of reverse transcriptase from HIV-1. *Science (New York, N.Y.)*, 242(4882):1171–1173.

[Roider et al. 2014] Roider, J., Meissner, T., Kraut, F., Vollbrecht, T., Stirner, R., Bogner, J. R., e Draenert, R. (2014). Comparison of experimental fine-mapping to in-silico prediction results of HIV-1 epitopes reveals ongoing need for mapping experiments. *Immunology*.

[Rouzine et al. 2014] Rouzine, I. M., Coffin, J. M., e Weinberger, L. S. (2014). Fifteen years later: hard and soft selection sweeps confirm a large population number for HIV in vivo. *PLoS genetics*, 10(2):e1004179.

[Smith, T. F.; Waterman 1981] Smith, T. F.; Waterman, M. S. (1981). Identification of Common Molecular Subsequences. pp. 195–197.

[Snoeck et al. 2011] Snoeck, J., Fellay, J., Bartha, I., Douek, D. C., e Telenti, A. (2011). Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. *Retrovirology*, 8(1):87.

[UK Colaborative Group on HIV Droug Resistance 2013] UK Colaborative Group on HIV Droug Resistance (2013). The increasing genetic diversity of HIV-1 in the UK, 2002-2010. *AIDS (London, England)*, (1):773–780.

[UNAIDS 2012] UNAIDS, W. H. O. (2012). Global Aids Response Progress Report. (1).

[UNAIDS 2013] UNAIDS, W. H. O. (2013). Report on the Global AIDS Epidemic. (1).

[van der Kuyl e Berkhout 2012] van der Kuyl, A. C. e Berkhout, B. (2012). The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. *Retrovirology*, 9:92.

[Vrancken et al. 2016] Vrancken, B., Trovão, N., Baele, G., van Wijngaerden, E., Vandamme, A.-M., van Laethem, K., e Lemey, P. (2016). Quantifying Next

Generation Sequencing Sample Pre-Processing Bias in HIV-1 Complete Genome Sequencing. *Viruses*, 8(1):12.

[Zou et al. 2015] Zou, D., Ma, L., Yu, J., e Zhang, Z. (2015). Biological databases for human research. *Genomics, Proteomics & Bioinformatics*.