DESENVOLVIMENTO DE MULTICLASSIFICADORES E DE UM SISTEMA DE IDENTIFICAÇÃO DE RESISTÊNCIA DO HIV-1 AOS ANTIRRETROVIRAIS

Letícia Martins Raposo

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Biomédica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Biomédica.

Orientador: Flavio Fonseca Nobre

Rio de Janeiro
Janeiro de 2018

DESENVOLVIMENTO DE MULTICLASSIFICADORES E DE UM SISTEMA DE IDENTIFICAÇÃO DE RESISTÊNCIA DO HIV-1 AOS ANTIRRETROVIRAIS

Letícia Martins Raposo

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA BIOMÉDICA.

Examinada por:

Profa. Rosimary Terezinha de Almeida, Ph.D.

Prof. Carlos Julio Tierra Criollo, D.Sc.

Prof. Oswaldo Gonçalves Cruz, D.Sc.

Prof. Sergio Miranda Freire, D.Sc.

Prof. Francisco Inacio Pinkusfeld Monteiro Bastos, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
JANEIRO DE 2018

Raposo, Letícia Martins

Desenvolvimento de multiclassificadores e de um sistema de identificação de resistência do HIV-1 aos antirretrovirais/Letícia Martins Raposo. – Rio de Janeiro: UFRJ/COPPE, 2018.

XXI, 126 p.: il.; 29, 7cm.

Orientador: Flavio Fonseca Nobre

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Biomédica, 2018.

Referências Bibliográficas: p. 90 – 104.

1. Multiclassificadores. 2. Sistema. 3. Resistência do HIV-1. I. Nobre, Flavio Fonseca. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Biomédica. III. Título.

Dedico esta tese aos meus pais,
Elizabeth e Gustavo, meus
maiores incentivadores.

Agradecimentos

Primeiramente, quero agradecer a Deus por tudo que Ele tem feito em minha vida. Obrigada, meu Pai, por sempre olhar por mim, me dando forças para enfrentar os desafios da vida. Muito obrigada por mais esta conquista!

Aos meus pais, meus maiores alicerces, que desde o início sempre investiram nos meus estudos, mesmo nos momentos de maior dificuldade. Obrigada por todo amor, dedicação e apoio incondicional! Sem vocês, nada disso seria possível.

Ao meu namorado Paulo, por todo carinho e amor dedicados a mim. Muito obrigada por estar sempre ao meu lado, nos momentos bons e ruins, me motivando sempre que o desânimo aparecia.

Ao meu orientador Flavio Nobre, com quem tive o prazer de conviver durante o mestrado e o doutorado. Obrigada por todos os ensinamentos, pela confiança e por estar sempre valorizando o meu trabalho. Tenho certeza de que não poderia ter um orientador melhor.

À professora Rosimary, que tem sido uma grande amiga e incentivadora ao longo dos meus anos no LESS. Obrigada pelo carinho, pelos momentos de conversa, pelos cafés pós-almoço e companhia nos retornos para casa.

À minha família, por toda contribuição valiosa ao longos desses anos. Obrigada, avós, tios e primos! Em especial, ao meu tio Werley, que desde a minha graduação, sempre me ajudou na revisão dos meus textos, procurando minhas falhas e faltas de atenção durante a escrita. Obrigada por fazer parte de mais esta conquista.

Aos meus amigos do LESS, com quem tive o prazer de conviver nesses últimos anos. Foi muito bom compartilhar com vocês todos esses momentos de doutorado. Obrigada pelas festas, almoços, ensinamentos, risadas, apoio... Agradeço em especial ao Alfonso, costarriquenho companheiro de café e apreciador de uma boa comida, ao Cesar, grande perfumista e incentivador, à Karla, companheira de orientador e de organização das festas, ao Rogério, grande conselheiro e criador do creme de milho mais gostoso que já comi, à Sulafa, amiga sudanesa, de coração gigante e cozinheira de mão cheia, e a todos os outros com quem tive a honra de trabalhar: Aline, André, Daniela, Giselle, Gleicy, Mariá e Thiago.

Aos grandes amigos que a vida me deu. Aos meus eternos companheiros Natália, Larissa e Fernando, presentes em minha vida há mais de 15 anos. Obrigada por todo

amor e amizade. Vocês também fazem parte desta conquista.

Às amigas do mestrado Aline, Bia, Carol, Kelly, Lili, Naty, Raquel e Vivi, agradeço os momentos de diversão e aventuras vividas ao longo desses anos.

Ao meu amigo Diego, que mesmo longe fisicamente, sempre esteve presente alegrando meus dias com sua criatividade e bom humor. Muito obrigada!

Aos secretários e professores do PEB pelo auxílio recebido ao longo de todos esses anos.

Às agências de financiamento Faperj, CNPq e Capes.

E a todos que, de alguma forma, contribuíram para a concretização deste trabalho, o meu mais sincero obrigada.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

DESENVOLVIMENTO DE MULTICLASSIFICADORES E DE UM SISTEMA DE IDENTIFICAÇÃO DE RESISTÊNCIA DO HIV-1 AOS ANTIRRETROVIRAIS

Letícia Martins Raposo

Janeiro/2018

Orientador: Flavio Fonseca Nobre

Programa: Engenharia Biomédica

Muitos algoritmos de interpretação genotípica têm sido elaborados com o intuito de detectar resistência do HIV aos antirretrovirais (ARV). Entretanto, esses sistemas têm apresentado discordâncias de classificação, gerando predições conflituosas da resposta terapêutica. Na prática clínica, ensaios genotípicos utilizados na detecção de resistência são realizados por meio do sequenciamento de Sanger, uma técnica com sensibilidade limitada, detectando apenas as variantes do HIV presentes em mais de 15-20% da população viral. Novas técnicas de sequenciamento de DNA, como o sequenciamento de nova geração (NGS), têm sido exploradas nos testes genotípicos de resistência do HIV. Essas técnicas são capazes de detectar mutações de resistência presentes em baixas frequências não detectáveis pela genotipagem atual. Os objetivos deste estudo foram desenvolver multiclassificadores de resistência a partir dos algoritmos de interpretação genotípica e implementar um ambiente integrado capaz de identificar as mutações de resistência do HIV-1 e os níveis de suscetibilidade aos ARVs a partir de dados brutos de NGS. Três estratégias diferentes foram utilizadas no desenvolvimento dos multiclassificadores: voto majoritário (VM), escolha do melhor algoritmo de interpretação genotípica (MS) e técnica *stacking*, com metaclassificadores *naïve Bayes* (NB) e k-NN. No geral, as abordagens NB e MS obtiveram os melhores resultados, com o NB sendo estatisticamente superior a pelo menos uma das outras três estratégias para quatro fármacos. O ambiente integrado recebeu o nome de SIRA-HIV e foi implementado na linguagem R. O sistema realiza uma avaliação abrangente dos dados de NGS, fornecendo ao usuário uma lista dos aminoácidos (e suas frequências) encontrados nas regiões analisadas, além da classificação de resistência do HIV-1 aos ARVs segundo dois pontos de corte.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

DEVELOPMENT OF ENSEMBLE CLASSIFIERS AND A SYSTEM OF IDENTIFICATION OF HIV-1 RESISTANCE TO ANTIRETROVIRALS

Letícia Martins Raposo

January/2018

Advisor: Flavio Fonseca Nobre

Department: Biomedical Engineering

Many genotypic interpretation algorithms have been elaborated to detect HIV resistance to antiretrovirals (ARV). However, these systems have shown discordances in classification, generating different predictions of the therapeutic response. In clinical practice, genotypic assays are performed by Sanger sequencing, a technique with limited sensitivity, detecting only HIV variants present in more than 15-20% of the viral population. New DNA sequencing techniques, such as new generation sequencing (NGS), have been used in HIV genotypic resistance assays. These techniques can identify HIV-1 drug resistance mutations present at low frequencies not detectable by current HIV-1 genotyping. This study aimed to develop ensemble classifiers from interpretation algorithms and to implement an integrated environment capable of identifying the HIV-1 resistance mutations and the levels of susceptibility to ARVs from raw NGS data. Three different strategies were used to develop the ensemble classifiers: majority voting (MV), choice of the best genotypic interpretation system (MS) and stacking technique, with naïve Bayes (NB) and k-NN as meta-classifiers. In general, NB and MS obtained the best results, with NB showing a statistically superior performance to at least one of the other three strategies for four drugs. The integrated environment was called SIRA-HIV, and it was implemented in the R language. The system performs a complete evaluation of the NGS data, providing to the user a list of amino acids and their frequencies found in the regions analyzed, and the HIV-1 resistance classification to ARVs according to two cut-offs.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

3TC	Lamivudina
A	Adenina
ABC	Abacavir
ANRS	Agence Nationale de Recherches sur le Sida
ART	Terapia Antirretroviral (*Antiretroviral Therapy*)
ARV	Antirretroviral
ATV	Atazanavir
AZT	Zidovudina
C	Citosina
CCR5	*Chemokine (C-C Motif) Receptor 5*
CCS	Centro de Ciências da Saúde
CHL	Centre Hospitalier de Luxembourg
CRF	Forma Circulante Recombinante (*Circulating Recombinant Form*)
CXCR4	*Chemokine (C-X-C Motif) Receptor 4*
d4T	Estavudina
ddI	Didanosina
ddNTP	Didesoxinucleotídeo Modificado (*Dideoxynucleoside Triphosphates*)
DLV	Delavirdina
DMC	Detroit Medical Centre
DNA	Ácido Desoxirribonucleico (*Deoxyribonucleic Acid*)

dNTP	Didesoxinucleotídeo Normai (*Deoxynucleoside Triphosphates*)
DRV	Darunavir
DTG	Dolutegravir
EDK	Estimativa de Densidade de *Kernel*
EFV	Efavirenz
EI	Inibidores de Entrada (*Entry Inhibitors*)
ETR	Etravirina
EVG	Elvitegravir
FDA	U. S. Food and Drug Administration
FI	Inibidores de Fusão (*Fusion Inhibitors*)
FN	Falso Negativo
FP	Falso Positivo
FPV	Fosamprenavir
FTC	Emtricitabina
G	Guanina
GAV	Grupo de Aconselhamento Virológico
HIV	Vírus da Imunodeficiência Humana (*Human Immunodeficiency Virus*)
HIVdb	Stanford HIV database
I	Intermediário
IDV	Indinavir
IN	Integrase
INSTI	Inibidores da Transferência da Cadeia pela Integrase (*Integrase Strand Transfer Inhibitors*)
k-NN	*k-Nearest Neighbors*
LPV	Lopinavir

MI	Inibidores de Maturação (*Maturation Inhibitors*)
MMR	Mutação Minoritária Resistente
MS	Melhor Sistema
MVC	Maraviroque
NB	*Naïve Bayes*
NFV	Nelfinavir
NGS	Sequenciamento de Nova Geração (*Next Generation Sequencing*)
NNRTI	Inibidores Não Análogos de Nucleosídeos da Transcriptase Reversa (*Non-Nucleoside Reverse Transcriptase Inhibitors*)
NRTI	Inibidores Análogos de Nucleosídeos da Transcriptase Reversa (*Nucleoside-analogue Reverse Transcriptase Inhibitors*)
NVP	Nevirapina
PCR	Reação em Cadeia da Polimerase (*Polymerase Chain Reaction*)
PI	Inibidores da Protease (*Protease Inhibitors*)
PR	Protease
qPCR	Reação em Cadeia da Polimerase em Tempo Real (*quantitative real-time Polymerase Chain Reaction*)
R	Resistente
RAL	Raltegravir
Rega	Rega Institute
RENAGENO	Rede Nacional de Genotipagem
RNA	Ácido Ribonucleico (*Ribonucleic Acid*)
RPV	Rilpivirina
RT	Transcriptase Reversa (*Reverse Transcriptase*)
RTV	Ritonavir
S	Suscetível
sdNVP	*single-dose Nevirapine*

Sinan	Sistema de Informação de Agravos de Notificação
SMOTE	*Synthetic Minority Over-sampling Technique*
SNP	Polimorfismo de Nucleotídeo Simples (*Single Nucleotide Polymorphism*)
SQV	Saquinavir
SUS	Sistema Único de Saúde
T	Timina
T-20	Enfuvirtida
TDF	Tenofovir
TPV	Tipranavir
VM	Voto Majoritário
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
VPN	Valor Preditivo Negativo
VPP	Valor Preditivo Positivo

Capítulo 1

Introdução

O vírus da imunodeficiência humana (*Human Immunodeficiency Virus*, HIV) é o agente etiológico causador da aids, uma das principais causas de morte em adultos jovens. No final de 2016, 36,7 milhões de pessoas viviam com o HIV no mundo, sendo que 1,8 milhões representavam novas infecções. No mesmo ano, cerca de 1,0 milhão de pessoas morreram de causas relacionadas à aids [1]. No Brasil, de 1980 a junho de 2017, foram registrados 882.810 casos de aids e, até dezembro de 2016, foram identificados 316.088 óbitos tendo como causa básica a HIV/aids. No Sistema de Informação de Agravos de Notificação (Sinan), foram notificados, no período de 2007 até junho de 2017, 194.217 casos de infecção pelo HIV. [2].

Os primeiros casos reconhecidos de aids ocorreram nos EUA, no início de 1980, em homens homossexuais [3]. A partir de 1983, quando o HIV foi identificado como agente responsável pela infecção e seu ciclo de vida foi caracterizado, a comunidade científica e médica iniciou uma busca por medicamentos. O primeiro antirretroviral (ARV) introduzido no mercado, em 1987, recebeu o nome de AZT (zidovudina) e permitiu um decréscimo da mortalidade e das infecções oportunistas em indivíduos com aids [4].

Durante as últimas três décadas, inúmeras combinações de diferentes ARVs foram desenvolvidos e os progressos obtidos com as opções terapêuticas foram significativos. Apesar dos benefícios clínicos alcançados com essas terapias, os casos de supressão viral incompleta e resistência do HIV aos ARVs são relativamente frequentes [5], o que limita o número de tratamentos disponíveis e contribui na transmissão de variantes genéticas resistentes [6].

A resistência aos ARVs é considerada um dos maiores obstáculos para o sucesso da terapia contra a infecção pelo HIV, sendo a grande responsável pela redução progressiva da potência dos componentes do esquema terapêutico. O uso de testes laboratoriais na detecção desse problema tem desempenhado um importante papel nas tomadas de decisão em relação às terapias a serem adotadas pelos indivíduos HIV+.

As abordagens mais utilizadas na detecção de resistência aos ARVs são a fenotipagem e a genotipagem. A fenotipagem é o método mais direto e mede a suscetibilidade do HIV aos medicamentos por meio de cultura, na presença de concentrações crescentes do fármaco de interesse. A genotipagem, por sua vez, determina a sequência do gene alvo do fármaco e usa essa informação para inferir a suscetibilidade do ARV [7]. Esse método é mais comumente utilizado na prática clínica, uma vez que ele é, geralmente, menos custoso, com processamento mais rápido e é menos laborioso do que a fenotipagem [8–10].

Muitos sistemas de interpretação de dados genotípicos têm sido desenvolvidos, ao longo desses anos, com o intuito de detectar os níveis de suscetibilidade do HIV aos ARVs. Os sistemas baseados em regras mais utilizados em todo o mundo são o ANRS (Agence Nationale de Recherches sur le Sida) [11], o Rega (Rega Institute) [12] e o HIVdb (Stanford University, HIV database) [13]. Esses algoritmos baseiam-se nas mutações identificadas pela genotipagem, fornecendo pontuações específicas, traduzidas em diferentes níveis de suscetibilidade para a grande maioria dos ARVs aprovados pelo FDA (U. S. Food and Drug Administration).

Esses algoritmos foram desenvolvidos utilizando diferentes conjuntos de dados, constituídos de variados subtipos virais, provenientes tanto de indivíduos HIV+ que fizeram uso de ARVs quanto daqueles que nunca experimentaram algum regime terapêutico [14]. Essas diferenças podem ter contribuído para as discordâncias de classificação entre os sistemas que têm sido observadas em diversos estudos [15–19], variando desde resultados distintos mas próximos até completamente opostos, o que gera predições conflituosas da resposta terapêutica.

Atualmente, os dados utilizados pelos algoritmos de interpretação genotípica são provenientes da genotipagem, executada por meio do sequenciamento de Sanger [20]. Esse tipo de sequenciamento possui uma menor sensibilidade que outros métodos mais modernos, detectando apenas as variantes do HIV presentes em mais de 15-25% da população viral [21–23].

Novas técnicas de sequenciamento de DNA (*Deoxyribonucleic Acid*), como os sequenciamentos de nova geração (*Next Generation Sequencing*, NGS), têm sido exploradas nos testes genotípicos de resistência do HIV. Essas estratégias apresentam uma grande velocidade de processamento, o que permite o sequenciamento de um enorme volume de dados. Além do mais, ensaios genotípicos baseados em NGS são capazes de detectar variantes minoritárias presentes na população viral com frequências de até 0,1-1% [20, 24], gerando informações adicionais que podem auxiliar no desenvolvimento de protocolos clínicos mais refinados e, assim, fornecer benefícios clínicos adicionais às pessoas vivendo com o HIV.

Várias plataformas de NGS têm sido desenvolvidas nos últimos 10 anos e se expandido na prática laboratorial. Além disso, numerosos estudos têm utilizado

a técnica de NGS para detectar variantes minoritárias de cepas virais resistentes aos medicamentos [25–31]. Alguns trabalhos identificaram variantes minoritárias do HIV contendo mutações, múltiplas ou únicas, resistentes a uma ou diversas classes de ARVs, tanto em indivíduos HIV+ que já fizeram uso prévio de medicamentos quanto naqueles que ainda não experimentaram nenhuma terapia [25–28, 30, 32].

A análise dos dados de NGS é uma etapa particularmente importante na identificação das variantes virais, demandando conhecimentos computacionais que, com certa frequência, são um campo desconhecido para muitos da área da saúde. A avaliação desses dados requer alguns conhecimentos de bioinformática, que englobam desde o uso de sistemas operacionais diferentes do padrão até conhecimentos de programação em linhas de comando, o que dificulta a interpretação das informações e limita a amplitude de estudos nessa área [33]. É importante que o acesso às análises dos dados de NGS se torne tão acessível quanto sua geração.

Diante do cenário apresentado, observa-se a necessidade de se obter um único perfil de resistência a partir das classificações geradas pelos principais algoritmos de interpretação, a fim de auxiliar na escolha do regime terapêutico a ser administrado a cada indivíduo HIV+. Além disso, o desenvolvimento de um sistema com interface simples e amigável capaz de analisar sequências obtidas do NGS seria importante para fornecer ao usuário informações acerca da população viral presente no organismo do paciente, assim como classificações do nível de resistência do HIV-1 aos ARVs disponíveis.

1.1 Objetivos

Os objetivos deste trabalho foram desenvolver multiclassificadores de resistência a partir dos algoritmos de interpretação genotípica e implementar um ambiente integrado capaz de identificar as mutações de resistência do HIV-1 e os níveis de suscetibilidade aos ARVs a partir de dados brutos de NGS provenientes de amostras de indivíduos HIV+.

Como objetivos específicos, o presente trabalho buscou:

- Desenvolver multiclassificadores para cada ARV a partir dos algoritmos ANRS, HIVdb e Rega por meio de três estratégias:
  - Voto majoritário;
  - Seleção do melhor algoritmo de interpretação genotípica;
  - Adaptação da técnica *stacking*.
- Desenvolver um sistema que:

- Identifique as variantes do HIV-1, presentes nas amostras dos indivíduos HIV+, e suas respectivas frequências na população viral;
- Forneça o nível de suscetibilidade do HIV-1 aos ARVs segundo os algoritmos ANRS, HIVdb, Rega e Algoritmo Brasileiro;
- Forneça o nível de suscetibilidade do HIV-1 aos ARVs segundo dois pontos de corte das frequências das variantes: $\geq 20\%$ e $\geq 1\%$.

1.2 Organização do Trabalho

A tese está organizada em sete capítulos. O segundo capítulo contempla a fundamentação teórica sobre o HIV, as terapias antirretrovirais e o problema da resistência, as estratégias de sequenciamento de DNA e as técnicas de combinação de classificadores. No capítulo 3, a revisão de literatura aborda os trabalhos publicados na área, assim como os achados importantes que motivaram o desenvolvimento do presente trabalho. O quarto capítulo detalha os métodos e as estratégias aplicados na obtenção dos classificadores a partir dos algoritmos de interpretação genotípica, assim como a metodologia empregada no desenvolvimento do sistema. O capítulo 5 apresenta os resultados encontrados na modelagem dos classificadores e exibe o sistema desenvolvido, seus módulos e suas funções. O capítulo 6 discute os achados encontrados e as limitações do estudo. Por fim, o capítulo 7 fornece uma conclusão da tese desenvolvida e sugere trabalhos futuros na área.

Capítulo 2

Fundamentação Teórica

2.1 HIV

2.1.1 Classificação

O HIV é um vírus RNA do gênero Lentivírus pertencente à família *Retroviridae*, sendo altamente mutável, o que gera inúmeras cepas (variantes genéticas) diferentes de HIV [34]. Com base em semelhanças genéticas, as cepas de vírus podem ser classificadas em tipos, grupos e subtipos.

O HIV pode ser classificado em dois tipos, HIV-1 e HIV-2, ambos transmitidos por contato sexual, pelo sangue e de mãe para filho. O HIV-1 é responsável pela grande maioria das infecções por HIV no mundo, sendo o tipo dois menos transmissível e mais restrito à África Ocidental [35]. As cepas do HIV-1 podem ser classificadas em quatro grupos: grupo M (*major*), grupo O (*outlier*) e dois novos grupos, N (*non-M*, *non-O*) e P. O grupo O parece estar restrito à África Centro Ocidental e o grupo N, uma cepa descoberta em 1998 em Camarões, é extremamente rara. Em 2009, uma nova cepa próxima a do HIV do gorila foi descoberta em uma mulher camaronesa, sendo designada como HIV-1 do grupo P [36, 37].

Dentro do grupo M, responsável por mais de 90% das infecções pelo HIV-1, são conhecidos, pelo menos, nove subtipos distintos geneticamente: A, B, C, D, F, G, H, J e K [38, 39]. Ocasionalmente, dois vírus de diferentes subtipos podem se reunir na célula de uma pessoa infectada e misturar o seu material genético, criando um novo vírus híbrido [40]. A maioria das recombinações não sobrevive durante muito tempo, mas aquelas que infectam mais de uma pessoa são conhecidas como formas circulantes recombinantes (*Circulating Recombinant Form*, CRF). Um exemplo de mistura é a dos subtipos A e B, representada por CRF A/B. Na Figura 2.1 está representado esquematicamente os grupos e subtipos do HIV-1.

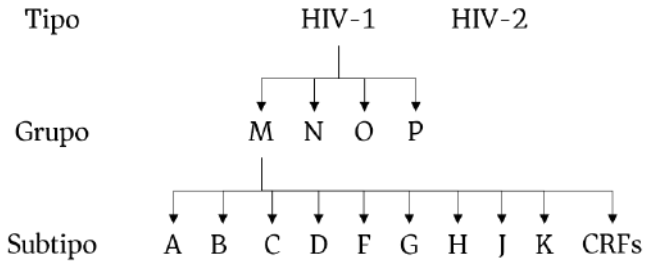


Figura 2.1: Cepas do HIV-1. Esse tipo é o mais comum da espécie e é classificado em quatro grupos e nove subtipos, além das formas circulantes recombinates.

As formas genéticas mais frequentes do HIV-1 são os subtipos A, B e C, sendo que cerca de 50% das infecções por HIV-1 em todo o mundo são decorrentes do subtipo C. O subtipo A é predominante na ilha de Madagascar e em países do Leste Europeu que constituíam a União Soviética [39]. O subtipo B é a principal forma genética na Europa Ocidental e nas Américas, incluindo o Brasil [41], sendo comum também em vários países do sudeste Asiático, norte da África e no Oriente Médio. O subtipo C é predominante na África do Sul e Índia [38, 39].

2.1.2 Estrutura e Genoma

As partículas virais do HIV-1 (vírion) são unidades esféricas com um diâmetro de cerca de 100-120 nm. Cada partícula é composta por duas cadeias simples de RNA (*Ribonucleic Acid*), encapsuladas por uma camada proteica (ou nucleocapsídeo), um capsídeo e um envelope externo composto por duas camadas fosfolipídicas [42]. Adicionalmente, os vírions são formados por 15 tipos de proteínas virais e algumas proteínas originárias da última célula hospedeira infectada [43]. A nomenclatura dessas proteínas adota a abreviação "gp" para glicoproteína e "p" para proteína, seguida de um número indicador do peso molecular em kilodáltons (kd). A Figura 2.2 mostra uma representação da estrutura do HIV, com alguns de seus componentes.

As proteínas do HIV-1 são sintetizadas a partir de nove genes: três genes que codificam os precursores poliproteicos (*gag, pol* e *env*) e seis genes acessórios com funções regulatórias e auxiliares: *vif, vpr, vpu, tat, rev* e *nef* [43]. Na Figura 2.3 está representado o genoma do HIV-1.

Os principais genes (*gag, pol* e *env*) codificam poliproteínas que são posteriormente proteolisadas em proteínas individuais [43]. As quatro proteínas estruturais provenientes da p55, codificada pelo gene *gag* (*group specific antigen*), são representadas pela p6, p7, p17 e p24. Cada fita simples de RNA está fortemente ligada às proteínas p6 e p7 do nucleocapsídeo, formado pela proteína p24. O capsídeo é envolto por uma matriz esférica composta pelas proteínas p17 [43, 44].

As duas proteínas codificadas pelo gene *env* (*envelope*), representadas pela gp120

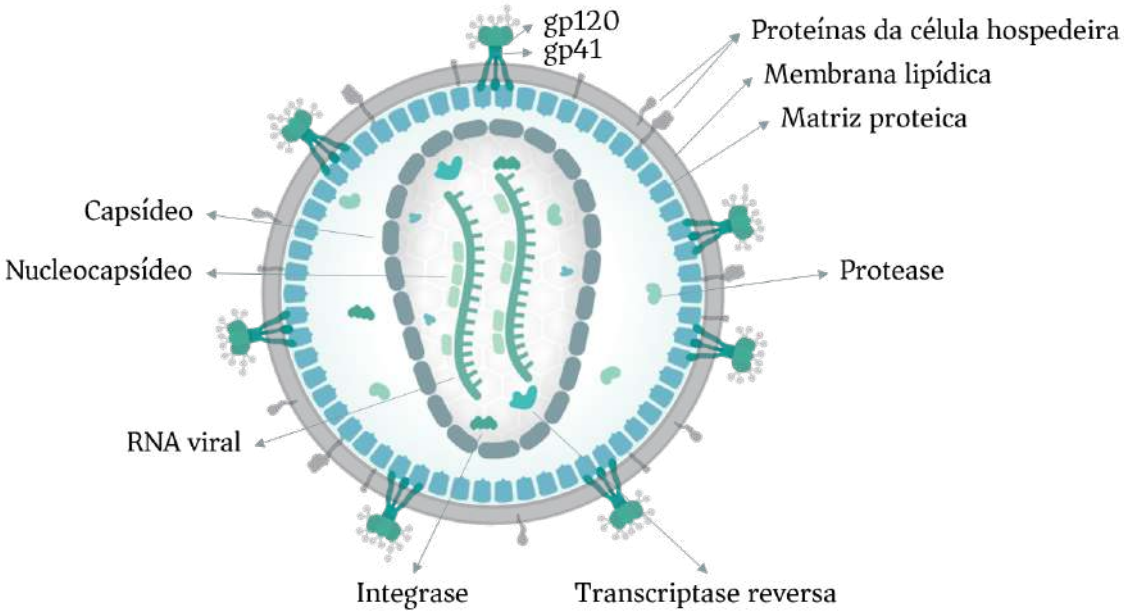


Figura 2.2: Diagrama da partícula viral do HIV-1. Adaptado de Thomas Splettstoesser, 2014.

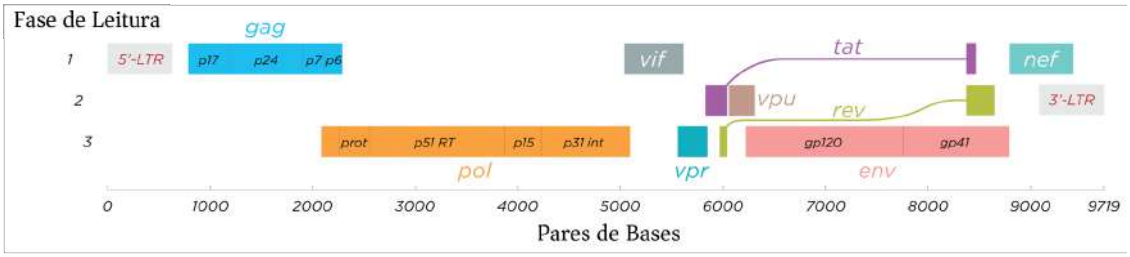


Figura 2.3: Representação do genoma do HIV-1 com as localizações relativas dos genes e as proteínas codificadas por cada um. Adaptado de Thomas Splettstoesser, 2014.

e gp41, são proteínas estruturais anexadas ao envelope viral e envolvidas na ligação aos receptores do HIV nas células do hospedeiro e na fusão do envelope viral com a membrana celular [43, 44].

As três proteínas codificadas pelo gene *pol* (*polymerase*) são [43, 44]:

- a protease (PR ou p10), que realiza a clivagem de precursores proteicos em unidades ativas menores após a liberação da partícula viral da célula do hospedeiro;
- a transcriptase reversa (*Reverse Transcriptase* (RT) ou p51/p66), um heterodímero constituído por duas subunidades estruturalmente distintas, necessária à replicação do HIV por transcrever RNA em DNA e;
- a integrase (IN ou p31), que tem como função principal promover a integração do DNA do HIV ao genoma do hospedeiro.

O HIV-1 codifica seis proteínas adicionais com função reguladora ou acessória [43, 44]:

- o gene *tat* (*HIV trans-activator*) codifica a p14, uma proteína reguladora que ativa a transcrição de genes provirais do HIV;
- o gene *rev* (*regulator of expression of virion proteins*) codifica a p19, uma proteína que transporta o RNA viral para a tradução no citoplasma;
- o gene *nef* (*negative factor*) codifica a p27, que modifica a célula hospedeira para aumentar a replicação viral e torná-la menos suscetível a ser destruída pelo sistema imune do hospedeiro;
- o gene *vpu* (*virus protein U*) codifica a p16 que auxilia na montagem eficiente dos vírions, brotamento dos mesmos para fora da célula hospedeira e promoção da morte celular;
- o gene *vpr* (*virus protein R*) codifica a p15, que auxilia na integração do DNA do HIV no núcleo da célula hospedeira e;
- o gene *vif* (*viral infectivity factor*) codifica a p23, estabilizando o recém-sintetizado DNA do HIV e facilitando o seu transporte para o núcleo.

2.1.3 Tropismo Viral

O tropismo viral está relacionado à capacidade dos vírus em entrar e infectar células hospedeiras específicas por meio da interação com receptores presentes nas membranas dessas células. *In vivo*, o HIV-1 infecta predominantemente as células do sistema imune, como células T, macrófagos e, em alguns casos, células dendríticas, por apresentarem receptores de superfície celular necessários ao desencadeamento da fusão das membranas virais e celulares e à entrada do vírus nas células [45].

A entrada do HIV-1 em uma célula-alvo depende, primariamente, da ligação da glicoproteína do envelope a um receptor CD4 e, também, a um correceptor, podendo este ser CXCR4 (*Chemokine (C-X-C Motif) Receptor 4*) ou CCR5 (*Chemokine (C-C Motif) Receptor 5*). O correceptor mais utilizado *in vivo* é o receptor de quimiocina CCR5, entretanto, durante a evolução da infecção, algumas variantes utilizam o correceptor CXCR4 [45].

As designações fenotípicas para os vírus são: (i) R5-trópicos (ou simplesmente R5) para as variantes que utilizam o correceptor CCR5, (ii) X4-trópicos (ou simplesmente X4) quando o correceptor CXCR4 é utilizado, e (iii) duplo-trópicos ou R5/X4 quando o vírus é capaz de utilizar ambos os correceptores [46]. Na Figura 2.4 está representado esquematicamente os tipos de HIV-1, segundo o tropismo, e os correceptores.

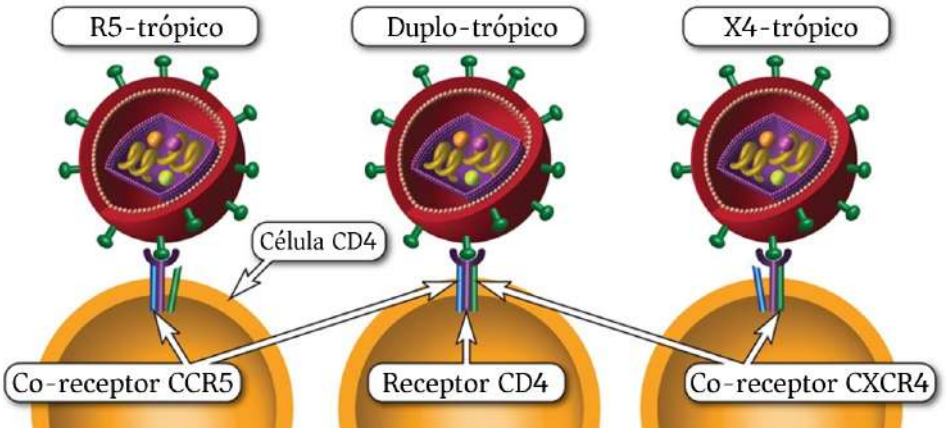


Figura 2.4: Tropismo do HIV-1. O vírus pode se ligar a uma célula CD4 hospedeira por meio dos correceptores CCR5, CXCR4 ou utilizar ambos. Adaptado de AIDSinfo, 2017.

2.1.4 Ciclo de Replicação

O ciclo de replicação do HIV-1 envolve diversas etapas. Ele se inicia com a entrada do vírus na célula hospedeira por meio das proteínas do envelope, que se ligam aos receptores CD4 presentes na superfície das células-alvo [42]. A ligação da proteína gp120 a esses receptores provoca alterações conformacionais na proteína, facilitando a exposição de um segundo local de ligação para o correceptor CCR5 ou CXCR4 [47]. Essa interação desencadeia um processo de fusão das membranas externas viral e celular, permitindo a entrada do conteúdo do vírus na célula.

Na etapa seguinte, a RT inicia a formação de uma molécula de cadeia dupla de DNA viral a partir das cadeias simples de RNA contidas na partícula viral [48]. Nessa fase, ocorre uma alta frequência de recombinação genética, que, devido às altas taxas de erro da RT, gera uma população viral altamente heterogênea [49].

No núcleo, a IN catalisa a inserção do DNA no cromossomo da célula hospedeira, permitindo que o mesmo seja expresso pela célula como se fosse um gene próprio [50]. O DNA integrado é transcrito pelo próprio maquinário da célula hospedeira, dando origem ao novo RNA viral. Esse irá compor o material genético de novas partículas virais ou migrar para o citoplasma da célula infectada, levando o código para a síntese de proteínas e enzimas virais. O código é traduzido em cadeias polipeptídicas, que se dobram para formar os componentes das proteínas e enzimas das novas partículas virais [42].

Durante e depois da montagem e liberação das partículas, a PR é responsável pelo processo de maturação. Ela cliva as cadeias polipeptídicas do HIV-1 em várias posições, tornando os vírions partículas infecciosas [42]. Uma única célula infectada pode liberar muitas novas partículas de HIV-1, as quais se movem para infectar

outras células em várias partes do corpo, onde novos ciclos de replicação serão iniciados. O ciclo de vida do HIV-1 está esquematizado na Figura 2.5.

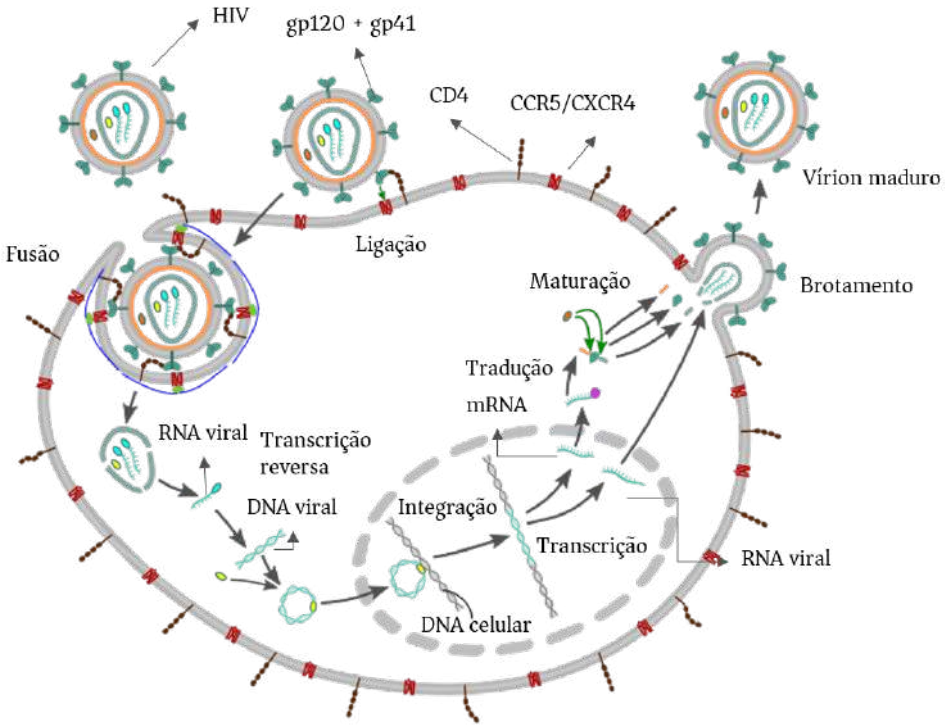


Figura 2.5: Esquematização do ciclo de replicação do HIV-1. Adaptado de Jordi March Nogué, 2017.

2.2 Terapia Antirretroviral

As terapias antirretrovirais (*Antiretroviral Therapy*, ART) padrão consistem na combinação de ARVs que visam reduzir ao máximo a replicação do HIV e impedir a progressão da infecção pelo vírus. As ARTs também são capazes de prevenir a transmissão do HIV. Reduções significantes das taxas de mortalidade e infecção pelo HIV têm sido observadas quando é feito o uso de um esquema antirretroviral potente, principalmente nos estágios iniciais da infecção [51].

A Organização Mundial da Saúde recomenda que todas as pessoas infectadas pelo HIV iniciem a terapia imediatamente após o diagnóstico, independente da contagem de células CD4 [52]. Segundo dados da UNAIDS, em junho de 2017, 20,9 milhões de pessoas, no mundo, tinham acesso aos ARVs, representando um aumento de aproximadamente 14,8% em relação ao meio do ano de 2016 [1]. No Brasil, de 2009 a 2015, o número de pessoas em tratamento no Sistema Único de Saúde (SUS) era de 455 mil pessoas [53].

O Brasil foi um dos primeiros países em desenvolvimento a garantir o acesso universal e gratuito aos ARVs no SUS, a partir de 1996, promovendo um aumento

da sobrevida de pessoas vivendo com HIV [54]. Antes de 1996, as opções de ARTs para a infecção por HIV-1 eram escassas no mercado, tendo como primeiro e único medicamento a AZT, em 1987 [4]. A revolução do tratamento só ocorreu em meados de 1990 com o desenvolvimento de inibidores da RT e da PR e a introdução de terapia antirretroviral altamente ativa (*Highly Active Antiretroviral Therapy*, HAART) que combinava essas duas classes de ARVs a fim de aumentar a eficácia e durabilidade dos tratamentos [55].

2.2.1 Antirretrovirais

Atualmente, existem seis classes principais de ARVs atuando em etapas e enzimas diferentes do ciclo de replicação do HIV-1. A primeira classe de fármacos aprovada pelo FDA foi a dos inibidores análogos de nucleosídeos/nucleotídeos da transcriptase reversa (*Nucleoside-analogue Reverse Transcriptase Inhibito*r, NRTI) [56]. Eles apresentam uma estrutura muito semelhante à dos nucleosídeos/nucleotídeos verdadeiros, competindo com esses na fase de replicação, impedindo que a síntese da cadeia de DNA tenha continuidade [57]. Os NRTIs aprovados atualmente pelo FDA são: abacavir (ABC), didanosina (ddI), emtricitabina (FTC), lamivudina (3TC), estavudina (d4T), tenofovir (TDF) e zidovudina (AZT) [58].

Outra classe que atua no desempenho da RT é a dos inibidores não análogos de nucleosídeos da transcriptase reversa (*Non-Nucleoside Reverse Transcriptase Inhibitor*, NNRTI). Essas pequenas moléculas possuem a capacidade de se ligar a uma região específica da RT, flexibilizando-a e bloqueando a síntese de DNA [59]. Atualmente, existem 5 NNRTI aprovados pelo FDA: delavirdina (DLV), efavirenz (EFV), etravirina (ETR), nevirapina (NVP) e rilpivirina (RPV) [58].

Os inibidores da protease (*Protease Inhibitor*, PI) atuam sobre a PR, responsável por promover a maturação das novas partículas virais produzidas no ciclo de replicação. Nove representantes estão aprovados atualmente pelo FDA. São eles: atazanavir (ATV), darunavir (DRV), fosamprenavir (FPV), indinavir (IDV), lopinavir (LPV), nelfinavir (NFV), ritonavir (RTV), saquinavir (SQV) e tipranavir (TPV) [58].

A classe dos inibidores da integrase, também denominados inibidores da transferência da cadeia pela integrase (*Integrase Strand Transfer Inhibitor*, INSTI), são responsáveis por bloquear a enzima IN, responsável por integrar o material genético do vírus ao DNA da célula infectada [58]. Atualmente, três representantes estão disponíveis: dolutegravir (DTG), elvitegravir (EVG) e raltegravir (RAL).

Duas classes de ARVs apresentam apenas um medicamento aprovado pelo FDA: a classe dos inibidores de entrada (*Entry Inhibitor*, EI), representado pelo maraviroque (MVC), que interferem na capacidade do vírus de se ligar aos receptores da

superfície externa das células CD4, e os inibidores de fusão (*Fusion Inhibitor*, FI), representado pela enfuvirtida (T-20), que impedem a fusão das membranas viral e celular, evitando a entrada do HIV-1 na célula [58].

Há ainda um classe experimental de ARV representada pelos inibidores de maturação (*Maturation Inhibitor*, MI). Esse fármacos, ao atuarem sobre a PR, interferem na clivagem entre a p24 e um pequeno peptídeo na poliproteína *gag*. Esse bloqueio ocasiona a formação de partículas virais imaturas incapazes de completar seu ciclo de vida e que, consequentemente, não são infectantes [60]. Seu representante atual é o Berivimat. Na Figura 2.6 encontra-se um esquema representativo dos locais de ação dos ARVs.

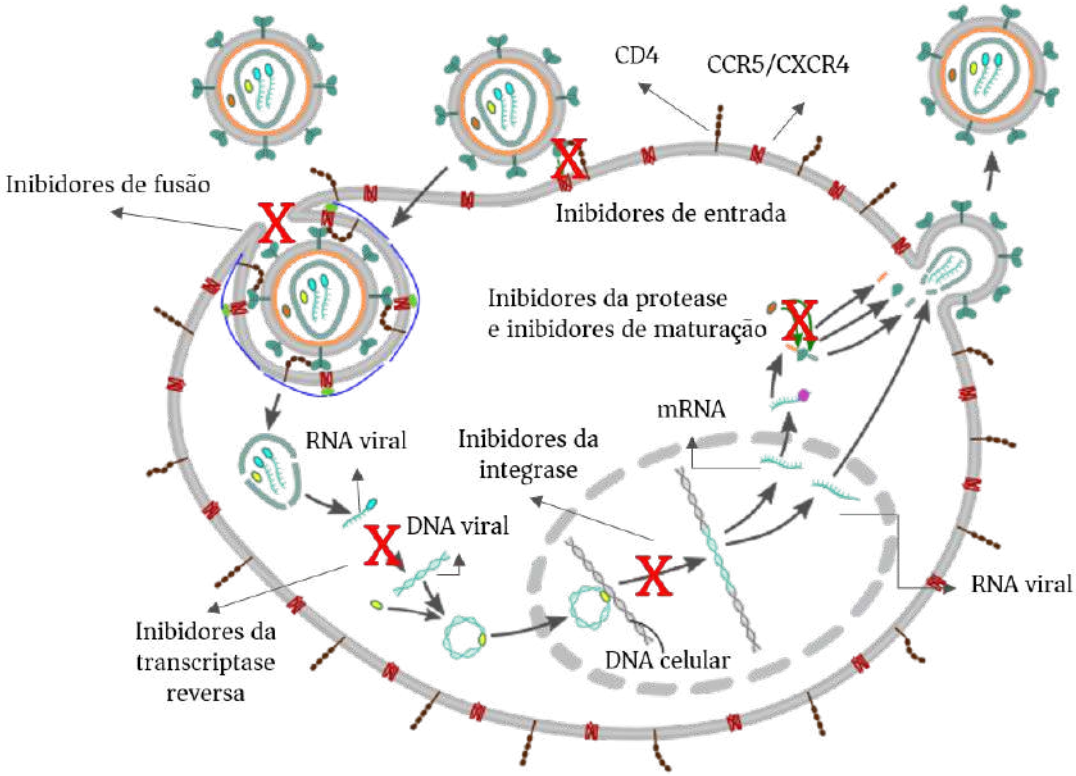


Figura 2.6: Descrição esquemática do mecanismo das seis classes de antirretrovirais atualmente disponíveis contra o HIV-1: inibidores de entrada, inibidores de fusão, inibidores da transcriptase reversa, inibidores da integrase, inibidores da protease e inibidores de maturação. Adaptado de Thomas Splettstoesser, 2013.

Alguns fatores podem prejudicar o efeito das ARTs e levar à falha virológica (não-obtenção ou não-manutenção de carga viral indetectável), quando não há supressão viral (RNA viral confirmado no plasma acima de 200 cópias/mL) [61]. Destacam-se a falta de aderência e a baixa tolerância aos medicamentos pelo indivíduo HIV+, ARVs pouco absorvidos pelo organismo, interações medicamentosas entre os ARVs e outros fármacos e, principalmente, a resistência aos ARVs [61, 62], um dos principais

problemas enfrentados pelos indivíduos vivendo com o HIV.

2.2.2 Resistência aos Antirretrovirais

A resistência do HIV aos ARVs consiste na capacidade do vírus de se replicar mesmo na presença do fármaco, decorrente de mutações nas proteínas virais alvo dos medicamentos [63]. As principais consequências da resistência são: falha do tratamento, necessidade de uso de medicamentos de segunda e terceira linha mais caros, e muitas vezes não disponíveis em diversos contextos, aumento dos custos de saúde associados ao uso e monitoração desses fármacos, propagação do vírus resistente e a necessidade de desenvolvimento de novos ARVs [64].

As mutações de resistência do HIV são classificadas em mutações primárias (ou principal) e secundárias (ou acessórias/compensatórias). As primeiras, quando presentes, reduzem significativamente a suscetibilidade do vírus a ARVs específicos por gerarem alterações em regiões de ligação do fármaco. As segundas, por outro lado, não possuem efeito significativo no fenótipo quando presentes sozinhas, mas podem contribuir para o aumento da resistência quando combinadas com as mutações primárias [65, 66].

Os testes de resistência mais utilizados são os ensaios fenotípicos, que medem a suscetibilidade do HIV aos ARVs em diferentes concentrações do fármaco, e os ensaios genotípicos, que identificam mutações de resistência no genoma viral.

Na fenotipagem, o nível de suscetibilidade do HIV em relação ao ARV é dado pela concentração do fármaco necessária para inibir a replicação do vírus em 50% (IC_{50}). Ela é normalmente expressa como *fold-change*, uma medida da expressão gênica que descreve o quanto uma quantidade muda de um valor inicial para um valor final. Nesse caso, a *fold-change* é dada pelo quociente entre a IC_{50} para o vírus do paciente e a IC_{50} para o vírus de referência.

A interpretação dos resultados da fenotipagem é padronizada a partir da definição de pontos de corte, que se destinam a definir um limiar acima do qual a suscetibilidade do vírus em relação a um dado fármaco começa a declinar. Pontos de corte clínicos, baseados em dados de resposta virológica nos ensaios clínicos, fornecem o limiar de maior relevância, mas também são os mais difíceis de se estabelecer. Pontos de corte biológicos são definidos por um limite superior da distribuição de suscetibilidade exibida pelo vírus do tipo selvagem [67].

Os testes de genotipagem se baseiam na detecção de mutações no genoma viral que conferem resistência fenotípica. Os fragmentos de RNA viral são obtidos a partir do plasma do paciente, convertidos em DNA e amplificados pela técnica de reação em cadeia da polimerase (*Polymerase Chain Reaction*, PCR). A análise dos padrões genotípicos de resistência baseia-se na avaliação das mutações genotípicas

identificadas juntamente com dados de resistência fenotípica [8].

Embora a fenotipagem seja considerada uma melhor ferramenta para avaliar a resistência, ela consiste em um teste demorado, trabalhoso e caro [7]. Por outro lado, a genotipagem é menos dispendiosa, tem custos mais baixos e é de execução mais fácil, permitindo que essa técnica seja amplamente utilizada [8–10]. Diretrizes de tratamento recomendam testes de resistência genotípica antes de iniciar a ART e no caso de falha virológica [68].

O teste genotípico padrão disponível comercialmente é baseado no sequenciamento de Sanger. Atualmente, apenas o teste ViroSeq HIV-1, da Abott (Chicago, Illinois, EUA), está comercialmente disponível [69]. Embora o sequenciamento de Sanger esteja disponível em sistemas automáticos, a interpretação dos seus resultados normalmente é conduzida manualmente, caracterizando-a como dependente do operador. Isso pode gerar uma limitação da reprodutibilidade e uma redução da sensibilidade, caso haja uma baixa concordância interoperador, que é relativamente frequente [70]. Ademais, os ensaios de genotipagem padrão possuem a capacidade de detectar apenas as mutações presentes em pelo menos 15-25% da população viral [21–23].

Nos últimos anos, diversas tecnologias modernas, como a reação em cadeia da polimerase em tempo real (*quantitative real-time Polymerase Chain Reaction*, qPCR) e o NGS, surgiram com a proposta de sequenciar fragmentos únicos de DNA de complexas misturas com um custo relativamente baixo, alto rendimento [71] e sensibilidade maior que a do método padrão, permitindo a detecção de mutações com frequência de até 0,1-1% [20, 24].

2.3 Sequenciamento de DNA

Os métodos de sequenciamento de DNA têm como objetivo determinar a ordem das bases nucleotídicas adenina (A), guanina (G), citosina (C) e timina (T) em uma molécula de DNA. O primeiro método popular foi o sequenciamento de Sanger [72], desenvolvido em 1977 pelo bioquímico britânico Dr. Frederick Sanger.

Mais recentemente, o método de Sanger foi suplantado por técnicas de NGS, especialmente em análises de grande escala. Entretanto, o sequenciamento tradicional permanece em largo uso em projetos de menor escala e na validação dos resultados de novas técnicas, constituídas pelas etapas de preparação do modelo de DNA (*template*), sequenciamento e imagem, alinhamento do genoma e métodos de montagem [71].

2.3.1 Sequenciamento de Sanger

O método clássico de terminação da cadeia, ou método de Sanger [72], é caracterizado por utilizar didesoxinucleotídeos modificados (*Dideoxynucleoside Triphosphates*, ddNTP) como terminadores da cadeia de DNA. Esse sequenciamento requer um molde de cadeia simples de DNA, um *primer* de DNA, uma DNA polimerase, desoxinucleotídeos padrão (*Deoxynucleoside Triphosphates*, dNTP), e ddNTPs. Esses últimos não possuem um grupo 3'-OH necessário para a formação de uma ligação fosfodiéster entre dois nucleotídeos, fazendo com que a DNA polimerase finalize a extensão do DNA quando esses nucleotídeos modificados são incorporados [72, 73]. Os ddNTPs podem ser marcados radioativamente ou com produtos fluorescentes para a detecção em máquinas de sequenciamento automático [74]. A Figura 2.7 esquematiza a ideia básica do sequenciamento de Sanger.

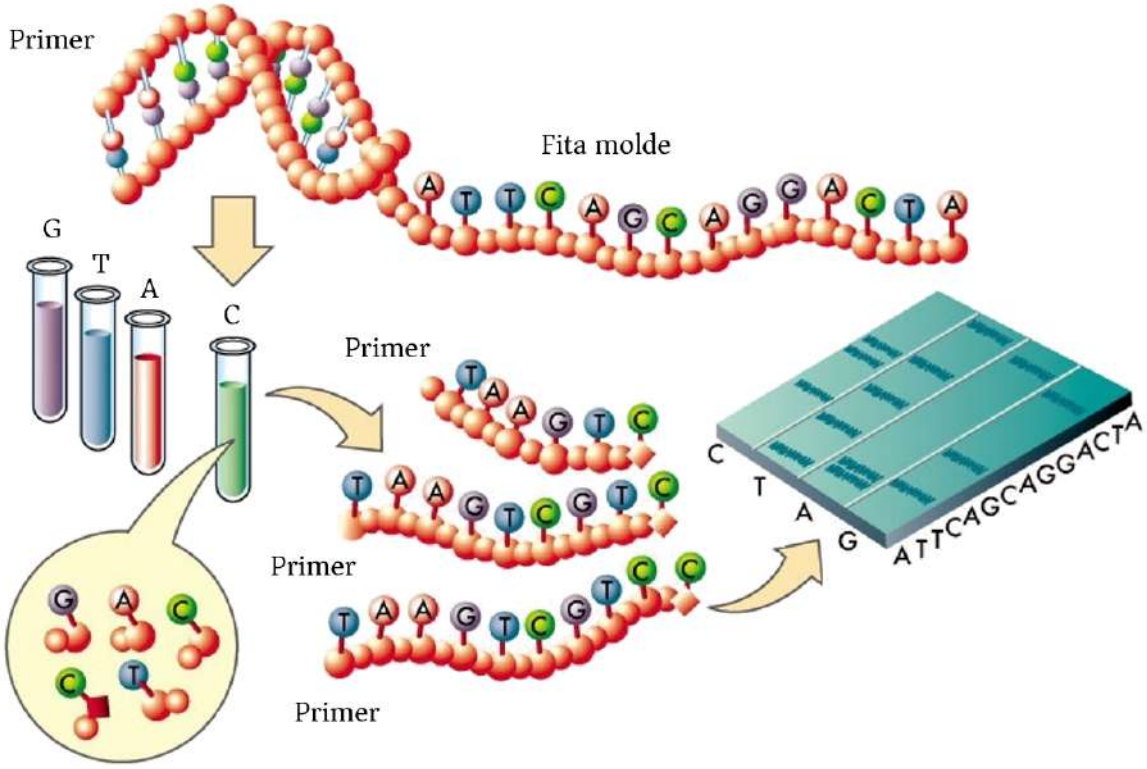


Figura 2.7: Esquematização do sequenciamento de Sanger. Cada tubo de reação é alimentado com DNA polimerase, *primer*, fita molde, dNTPs e ddNTP específico (dATP, dTTP, dCTP ou dGTP). A reação de polimerização ocorre até um ddNTP ser incorporado. Na eletroforese é possível determinar a sequência complementar da fita molde. Adaptado de Ned Shaw.

Inicialmente, milhões de cópias da sequência a ser determinada são purificadas ou amplificadas, dependendo da fonte da sequência. Em seguida, a amostra a ser sequenciada é destinada a quatro regiões (tubos, por exemplo) de sequenciamento separadas, contendo, em cada uma, todos os dNTPs e a polimerase [75]. Em cada

região é adicionado apenas um dos quatro didesoxinucleotídeos modificados (ddATP, ddGTP, ddCTP ou ddTTP), em concentrações menores que as dos dNTPs [76].

Quando um ddNTP é incorporado no lugar de um dNTP, a síntese da cadeia é prematuramente finalizada, gerando, em cada uma das regiões, cadeias de DNA de tamanhos diferentes. Cada mistura é então separada em um gel de sequenciamento por eletroforese para detectar os nucleotídeos presentes na sequência de DNA. As sequências obtidas irão diferir entre si apenas por um nucleotídeo [76]. Cada banda indica um fragmento de DNA, resultado da terminação da cadeia após a incorporação de um ddNTP. As posições relativas das diferentes bandas entre as quatro pistas da eletroforese, de baixo para cima, fornecem a sequência da cadeia de DNA complementar à cadeia usada como molde [76]. O método de Sanger é capaz de sequenciar fragmentos de DNA (também denominados leituras ou *reads*) com tamanhos entre 800 a 900 pares de base [77].

Os testes genotípicos empregados atualmente na prática clínica para a detecção de resistência do HIV aos ARVs são realizados por meio do método de Sanger. Nesse tipo de sequenciamento, nem todas as variantes que compõem a diversidade viral presente no indivíduo infectado são detectadas, uma vez que esse método é sensível apenas às mutações prevalentes em pelo menos 15-25% da população viral [21–23].

2.3.2 Sequenciamentos de Nova Geração

Os métodos de NGS, também conhecidos como sequenciamentos de alto rendimento, têm como característica permitir milhares de reações de sequenciamento em paralelo, produzindo um enorme volume de dados em alta velocidade [78, 79].

Essas metodologias, ao proporcionarem um aumento significativo na quantidade de bases sequenciadas, geram uma relação custo/benefício superior em relação ao sequenciamento de Sanger, pois oferecem um preço final por base sequenciada mais atraente e um processo realizado em menor tempo [71, 78]. Entretanto, as plataformas de NGS apresentam algumas limitações, como a geração de erros de sequenciamento e a produção de uma enorme quantidade de dados na forma de sequências curtas, representando um desafio para o desenvolvimento de softwares e algoritmos computacionais mais eficientes [78].

Atualmente, alguns dos sistemas de NGS disponíveis são: SOLiD/Ion Proton/Ion 5S/Ion Torrent Personal Genome Machine (PGM) (Thermo Fisher Scientific, Carlsbad, CA, USA), Genome Analyzer/HiSeq 2500/HiSeq 3000/4000/ HiSeq X/MiSeq/MiniSeq/NextSeq 500 (Illumina, San Diego, CA, USA), GS FLX Titanium/GS Junior (454 Life Sciences, Roche Diagnostics, Branford, CT, USA), GeneReader (QiAGEN, Venlo, Netherlands), Pacific BioSciences RS II/Sequel (PACIFIC BIOSCIENCES, Menlo Park, CA, USA), BGISEQ-500 (BGI Platform, Yantian Dis-

trict, Shenzhen, China) e MiniION/PromethION (Oxford, OX4 4GA, UK) [80].

Embora cada tecnologia de NGS seja única em seu funcionamento, as principais plataformas atuais possuem metodologias semelhantes que incluem a preparação dos modelos de DNA, o sequenciamento e a análise de dados [71].

A preparação do modelo consiste em construir e amplificar uma biblioteca de DNA. Essas bibliotecas são construídas por fragmentos da amostra de DNA, possuindo, basicamente em suas extremidades, adaptadores ligados covalentemente. Esses adaptadores são sequências universais de DNA sintético, específicas de cada plataforma [81]. Uma vez construídas, as bibliotecas são clonalmente amplificadas para a etapa de sequenciamento [82].

Predominantemente, as tecnologias de NGS baseiam-se em um princípio de sequenciamento por síntese. Nesse tipo de sequenciamento, os fragmentos de DNA atuam como moldes para a síntese de novos fragmentos. À medida que os nucleotídeos se incorporam na cadeia de DNA crescente, há a emissão de um sinal único, podendo esse ser uma molécula fluorescente, o que ocorre na maioria dos sistemas de NGS, ou um sinal na forma de uma alteração de pH, como é o caso do Ion Torrent [82, 83]. Na Figura 2.8 estão representadas as etapas básicas do sequenciamento por síntese.

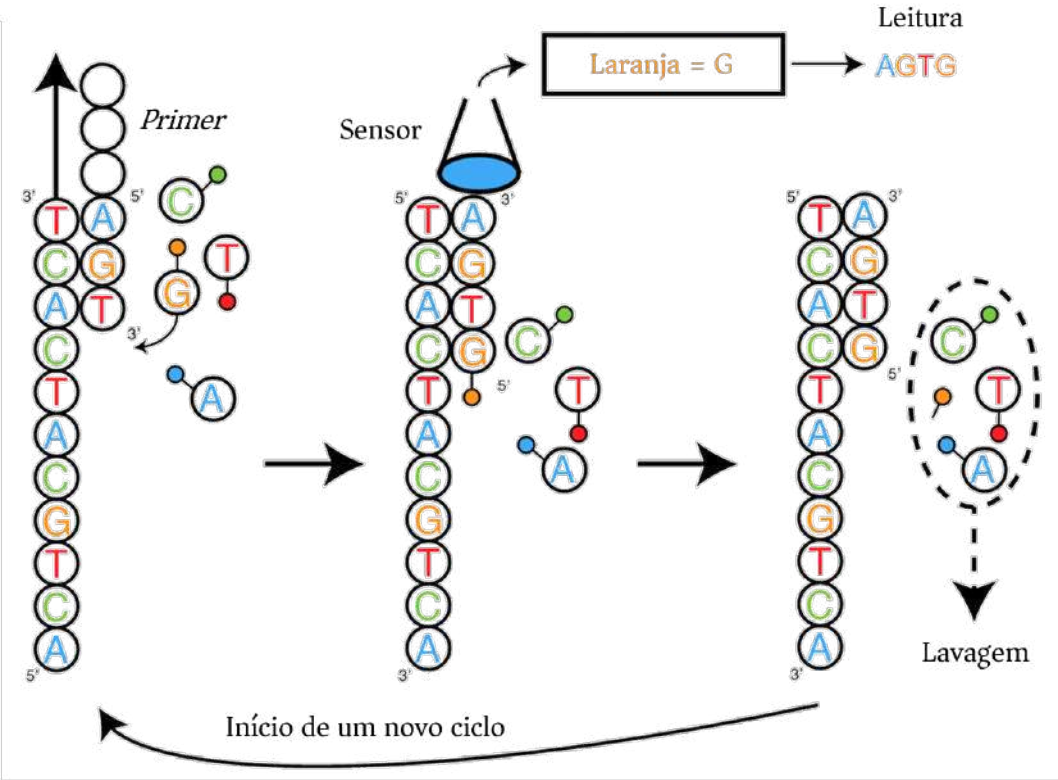


Figura 2.8: Sequenciamento por síntese. Os fragmentos de DNA atuam como moldes para a síntese de novos fragmentos. À medida que os nucleotídeos se incorporam na cadeia de DNA, há a emissão de um sinal único a ser registrado por um sensor.

Após o sequenciamento, os dados brutos são submetidos a várias etapas de análise, como o pré-processamento das sequências para remoção dos adaptadores e das leituras de baixa qualidade, e o mapeamento dos dados em relação a um genoma de referência. Dependendo do tipo de objetivo do sequenciamento aplicado, a análise de dados de NGS pode incluir a detecção de polimorfismos de nucleotídeo simples (*Single Nucleotide Polymorphism*, SNP) ou *indels* (inserção ou deleção de bases), a identificação de novos genes ou elementos reguladores, e a avaliação dos níveis de expressão de transcritos. Muitas ferramentas e softwares estão disponíveis para realizar a análise desses dados [84].

O NGS vem sendo aplicado em diversas áreas da ciência, como no sequenciamento e investigação da diversidade do genoma, na metagenômica, na epigenética, na descoberta de RNAs não-codificantes e de locais de ligação às proteínas [85, 86] e na detecção de alelos ligados ao câncer [78]. Na microbiologia e na virologia, os métodos de NGS abrangem a descoberta de novos microrganismos e vírus usando abordagens de metagenômica [87], a análise de variabilidade do genoma viral dentro do hospedeiro [88] e a detecção de mutações de resistência a medicamentos antivirais, presentes em menor quantidade, em pacientes infectados pelo HIV [86].

As técnicas de NGS permitem detectar e quantificar variantes resistentes presentes em reduzidas frequências (0,1-1% da população viral) [20, 24], apresentando maior sensibilidade que a técnica tradicional de sequenciamento de Sanger. Apesar de ser um tópico objeto de debate, já foi observado que HIVs resistentes que correspondem a pequenas frações da população viral podem ser clinicamente importantes, uma vez que essas variantes podem crescer rapidamente sob a pressão seletiva exercida pelos medicamentos, podendo levar futuramente à falha virológica [25].

2.4 Algoritmos de Interpretação Genotípica de Resistência

Os algoritmos de interpretação genotípica são sistemas de classificação de resistência desenvolvidos com base em dados da literatura científica e conhecimento de especialistas. Esses sistemas são constituídos por regras que descrevem interações e/ou combinações entre determinadas mutações relacionadas ao problema da resistência [14].

O início do desenvolvimento desses algoritmos ocorreu em 1997 em centros de pesquisa em Paris (ANRS), Califórnia (Stanford HIVdb) e Leuven (Rega), sendo esses sistemas os mais amplamente utilizados. As versões iniciais foram desenvolvidas com base no grupo M, subtipo B do HIV-1, e, com o passar dos anos, estendido aos

outros subtipos, como grupos A e B do HIV-2 e grupo O do HIV-1 [89].

Os algoritmos ANRS, HIVdb e Rega foram desenvolvidos a partir de diferentes conjuntos de dados, subtipos, tipo de pacientes (fizeram ou não uso de ARVs), entre outras diferenças, podendo gerar resultados discordantes de classificação entre eles. Os três sistemas podem ser acessados diretamente na página do Stanford Drug Resistance Database (http://hivdb.stanford.edu/) em suas versões atuais.

No Brasil, um algoritmo de interpretação genotípica também foi desenvolvido com a finalidade de abranger, em suas regras, características específicas do HIV no país. O sistema recebeu o nome de Algoritmo Brasileiro e foi implementado em 2003.

2.4.1 ANRS

O algoritmo ANRS foi desenvolvido para os tipos HIV-1 e HIV-2 a fim de orientar os médicos na escolha da ART. O sistema baseia-se estatisticamente na correlação entre as mutações de resistência presentes no genoma do vírus e os desfechos virológicos de pacientes com falha terapêutica [11].

O ANRS foi projetado para indicar o nível de suscetibilidade do HIV aos PIs, NRTIs, NNRTIs, INSTIs e FI, classificando as amostras em três níveis: suscetível, possível resistência e resistente. Uma classificação suscetível indica que o ARV, em particular, será eficaz no tratamento da infecção pelo HIV. A possível resistência, por sua vez, indica que o fármaco é parcialmente eficaz. No caso do nível resistente, o ARV perde sua eficácia [9, 14].

O sistema fornece tabelas de regras contendo as mutações que conferem resistência ou possível resistência genotípica aos fármacos. As regras são atualizadas regularmente, estando disponível atualmente a versão 27, de setembro de 2017 (http://www.hivfrenchresistance.org/)

2.4.2 HIVdb

O algoritmo HIVdb [13], provavelmente o mais reconhecido dentre os sistemas de interpretação genotípica, gera um perfil de resistência usando regras booleanas e a imposição de penalidades. O sistema classifica sequências da PR, RT e IN e aceita como entrada dados no formato FASTA ou uma lista de mutações, que são posteriormente comparadas com a sequência consenso do subtipo B.

O HIVdb atribui uma pontuação de penalidade para cada mutação de resistência aos ARVs, classificando o nível de suscetibilidade em cinco categorias: suscetível (0 a 9 pontos), potencial resistência de baixo nível (10 a 14 pontos), resistência de baixo nível (15 a 29 pontos), resistência intermediária (30 a 59 pontos) e resistência

de alto nível (acima de 60 pontos). A saída do sistema contém a pontuação total e os comentários sobre as mutações presentes na amostra. As pontuações são atualizadas com base em novos dados publicados e no parecer dos usuários e especialistas da área [9, 14]. A versão atual do HIVdb é a 8.4, de junho de 2017 (http://sierra2.stanford.edu/sierra/servlet/JSierra).

2.4.3 Rega

O algoritmo Rega [12], assim como o ANRS, classifica os níveis de suscetibilidade do HIV aos ARVs em três categorias: suscetível, intermediário e resistente. Esse sistema foi estabelecido para uso diário na prática clínica [9], com regras de classificação voltadas para os PIs, NRTIs, NNRTIs, INSTIs e EI, para ambos HIV-1 e HIV-2. Atualmente, o Rega está disponível na versão 9.1.0, lançada em novembro de 2013, para o HIV-1, e na versão 8.0.2 para o HIV-2, de junho de 2009 (https://rega.kuleuven.be/cev/avd/software/rega-algorithm).

2.4.4 Algoritmo Brasileiro

O Algoritmo Brasileiro é um sistema de interpretação genotípica desenvolvido com base nas regras de resistência aos ARVs criadas pelo comitê da Rede Nacional de Genotipagem (RENAGENO). O algoritmo considera três faixas de interpretação (suscetível, resistência intermediária e resistente) e classifica a suscetibilidade do HIV-1 apenas aos PIs, NRTIs e NNRTIs. A entrada dos dados ocorre por meio de um arquivo no formato FASTA ou via digitação das mutações.

As regras de interpretação de resistência do HIV-1 são apresentadas na forma de uma tabela e estão disponíveis no site do Departamento de DST/AIDS e Hepatites Virais do Ministério da Saúde (http://50.116.24.135:8080/HIV/resistencia.jsp). O algoritmo encontra-se na versão 13, de dezembro de 2015.

2.5 Multiclassificadores

Um multiclassificador é um sistema composto por um conjunto de classificadores treinados individualmente (tais como redes neurais ou árvores de decisão), cujas predições individuais são combinadas para classificar novas instâncias. Muitos estudos têm investigado a técnica de combinar as predições de múltiplos classificadores a fim de produzir um único sistema de classificação com melhor desempenho global [90–92].

Um multiclassificador é concebido em duas etapas:

1. Construção dos classificadores-base, que podem ser homogêneos e, portanto,

são gerados a partir de um mesmo algoritmo de aprendizagem de máquina, mas com subconjuntos de treinamento diferentes [93]; ou heterogêneos, gerados a partir de algoritmos diferentes, mas com dados de treinamento iguais ao da amostra original [94];

2. Combinação dos classificadores-base, por meio de um esquema de votação das saídas dos modelos ou mediante o uso de um algoritmo de aprendizagem de máquina que tem como entrada as saídas de cada classificador-base.

Três métodos mais populares para a criação de multiclassificadores são: *bagging* [95] e *boosting* [96–98], ambos representando a classe dos multiclassificadores homogêneos, e *stacking* [91, 92], um sistema multiclassificador heterogêneo.

2.5.1 *Bagging*

Esta estratégia, cujo nome vem da composição de *bootstrap* e *aggregating*, foi proposta por Breiman, em 1996 [95], e baseia-se na técnica de amostragem por *bootstrap* desenvolvida por Efron e Tibshirani (1994) [99].

Nessa metodologia, vários classificadores do mesmo tipo (por exemplo, redes neurais artificiais) são gerados a partir de subconjuntos diferentes obtidos com reposição dos dados a partir do conjunto original [100]. Os subconjuntos apresentam o mesmo tamanho da amostra original, com algumas instâncias aparecendo mais de uma vez nos subconjuntos e outras não sendo selecionadas na amostragem.

Dessa forma, dado um conjunto de treinamento $\mathcal{T} = \{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$, em que y representa a classe ou a resposta numérica, B amostras *bootstrap* $\left\{\mathcal{T}^{(B)}\right\}$ são aleatoriamente obtidas de $\mathcal{T}$, formando, a partir de cada uma delas, B preditores $\left\{\varphi\left(\mathbf{x}, \mathcal{T}^{(B)}\right)\right\}$. $\left\{\mathcal{T}^{(B)}\right\}$ consistem em amostras formadas por n instâncias obtidas por amostragem com reposição do conjunto de dados original $\mathcal{T}$.

Se y é numérico, a predição final φ_B será dada pela média simples dos B desfechos gerados por cada preditor

$$\varphi_B(\mathbf{x}) = média_B \varphi\left(\boldsymbol{x}, \mathcal{T}^{(B)}\right). \tag{2.1}$$

Se y for uma classe, o desfecho final $\varphi_B(\mathbf{x})$ será obtido pelo número de votos de cada preditor [95]. Na Figura 2.9 encontra-se um desenho esquemático do algoritmo *bagging*.

Esse algoritmo é comumente recomendado quando os preditores possuem um comportamento instável, isto é, quando pequenas mudanças no conjunto de treinamento ocasionam grandes mudanças no preditor. Assim, ao combinar múltiplos classificadores instáveis, é mais provável que as respostas finais possuam maior chance de acerto.

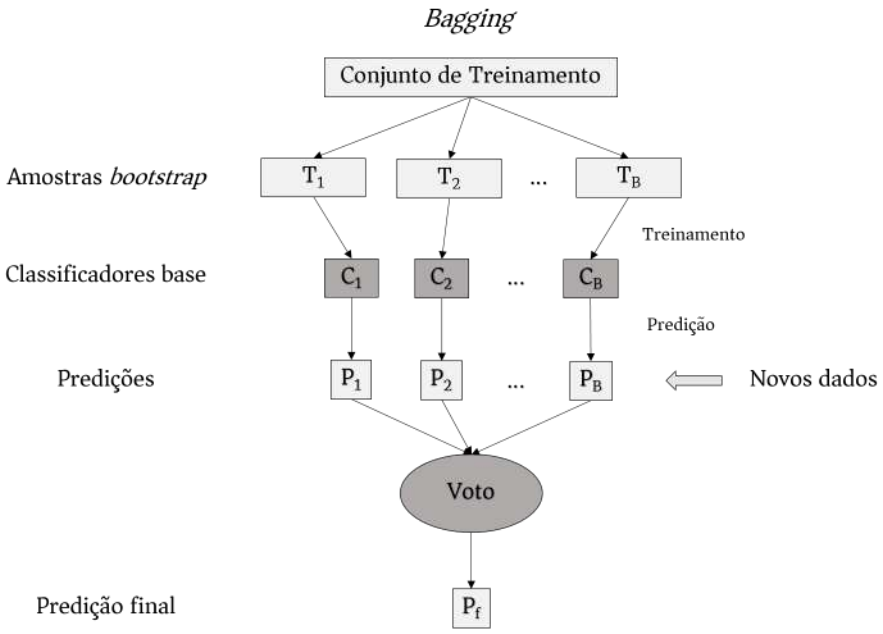


Figura 2.9: Algoritmo *Bagging*. A partir das amostras obtidas por *bootstrap*, classificadores-base são desenvolvidos e seus desfechos são combinados a fim de fornecer uma única predição para cada instância.

2.5.2 *Boosting*

A técnica *boosting* foi proposta por Schapire e colaboradores, em 1990 [96], aperfeiçoada por Freund, em 1995 [97], e tem como objetivo combinar classificadores fracos de modo a alcançar um classificador final mais acurado.

Nessa abordagem, a seleção de cada instância depende de uma probabilidade associada a ela, que varia conforme a sua contribuição para o erro dos classificadores já treinados [100]. Assim, caso uma instância não tenha sido classificada corretamente pelo presente classificador, sua probabilidade de escolha para o classificador seguinte aumenta. Nessa técnica, é necessário que os classificadores-base sejam treinados sequencialmente.

Um dos algoritmos de *boosting* mais utilizadas é o *AdaBoost* (*Adaptative Boosting*), em que os subconjuntos de treino são amostrados de forma adaptativa, de acordo com a contribuição das instâncias para o erro dos classificadores-base [101]. A Figura 2.10 apresenta o funcionamento geral do algoritmo *AdaBoost*.

No *AdaBoost* discreto, dado um conjunto de treinamento $\mathcal{T} = \{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$, em que $y_i \in \{-1, +1\}$ representa a classe de cada instância, em cada iteração $t = 1, \ldots, T$, pesos $w_{i,t}$ são computados para cada uma das n instâncias de treinamento.

Na primeira rodada, todos os pesos são iguais, com cada instância apresentando probabilidade de ser selecionada para o subconjunto de treinamento igual a $\frac{1}{n}$. Dessa forma, n instâncias são selecionadas e um classificador-base h_t é treinado e testado

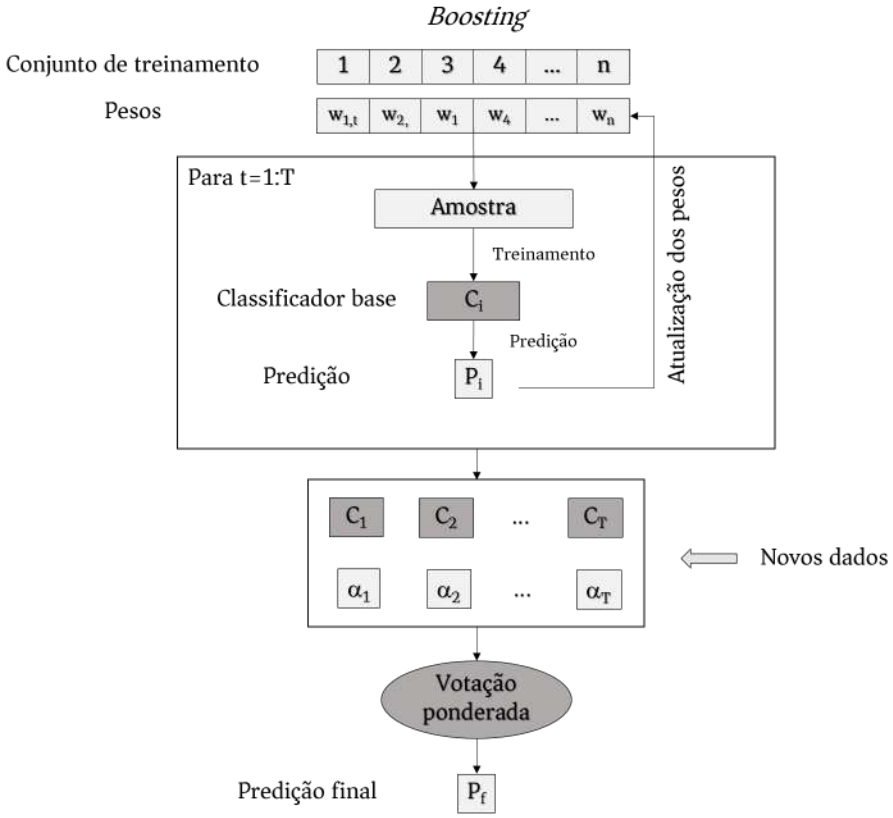


Figura 2.10: Algoritmo *AdaBoost*. Em cada iteração, a distribuição de pesos de cada instância é atualizada a fim de que os exemplos classificados erroneamente recebam maiores pesos e, assim, maior probabilidade de serem selecionados para compor o classificador seguinte.

com os exemplos de treinamento. A soma dos erros ponderados para as instâncias classificadas erroneamente é computada como

$$\varepsilon_t = \sum_{\substack{i=1 \\ h_t(\boldsymbol{x}_i) \neq y_i}}^{n} w_{i,t} \tag{2.2}$$

e para a rodada seguinte, os pesos são atualizados pela fórmula

$$w_{i,t+1} = \frac{w_{i,t} \exp(-\alpha_t y_i h_t(\mathbf{x}_i))}{Z_t} \tag{2.3}$$

em que $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$, em seu valor ótimo, minimiza o erro de treinamento, e $Z_t = \sum_{i=1}^{n} w_{i,t} \exp(-\alpha_t y_i h_t(\mathbf{x}_i))$ é um fator de normalização que transforma $w_{i,t+1}$ em uma distribuição de probabilidade.

As instâncias que mais contribuem para o erro apresentarão maior probabilidade de serem selecionadas para compor o próximo subconjunto de treinamento a ser utilizado no classificador-base seguinte. Por fim, as saídas dos classificadores-base são combinadas por um esquema de votação para compor o desfecho do classificador

final, sendo determinado por

$$H(\mathbf{x}) = sinal\left(\sum_{t=1}^{t} \alpha_t h_t(\mathbf{x}_i)\right) \tag{2.4}$$

em que $sinal\,(\boldsymbol{x})$ é a função

$$sinal\,(\boldsymbol{x}) = \begin{cases} -1, & \text{se x menor que 0} \\ +1, & \text{caso contrário} \end{cases}. \tag{2.5}$$

2.5.3 *Stacking*

A técnica *stacked generalization* (ou *stacking*) foi desenvolvida, em 1992, por Wolpert [91]. A ideia dessa estratégia é aprender quais classificadores-base, gerados por diferentes algoritmos de aprendizagem de máquina, apresentam melhor desempenho, e tentar identificar uma forma de combinar suas saídas, obtendo um classificador único com melhor acurácia.

O algoritmo pode ser dividido em duas etapas. Na primeira, denominada nível 0, o *stacking* constrói um metaconjunto de dados combinando a saída dos n classificadores-base (classificadores de nível 0) diferentes $\{C_1, \ldots, C_n\}$. Dessa forma, o número de atributos do metaconjunto será igual ao número de classificadores-base.

A partir de um conjunto de treinamento $\mathcal{T}$, os dados são particionados em k subconjuntos disjuntos, em que, por vez, $k-1$ conjuntos são utilizados no treinamento dos classificadores-base e o subconjunto restante é utilizado para obter as predições dos classificadores que serão utilizadas como entrada no metaclassificador. Este processo é realizado k vezes alternando de forma circular o subconjunto utilizado para gerar as predições.

Na etapa seguinte, nível 1, o conjunto de dados formado pelas classes preditas $s = (c_1, c_2, \ldots, c_n)$ é usado no treinamento de um metaclassificador (classificador de nível 1) M_c, responsável pela predição final do sistema, e do qual se espera um desempenho global superior ao dos classificadores-base.

A fim de prever a classe de um novo exemplo utilizando o meta classificador obtido, a instância é categorizada pelos classificadores de nível 0 e suas saídas $(c_1, c_2, \ldots, c_n)$ geram uma instância de nível 1. O classificador de nível 1, portanto, combina essas saídas e fornece uma predição final (Figura 2.11).

Nessa técnica, a decisão final não é obtida por meio de votos, o que poderia ser um problema caso um dos classificadores-base gerasse predições grosseiramente incorretas. Normalmente, como os classificadores de nível 0 são os responsáveis pela tarefa mais importante de classificação, o classificador de nível 1 pode ser um algoritmo mais simples [102].

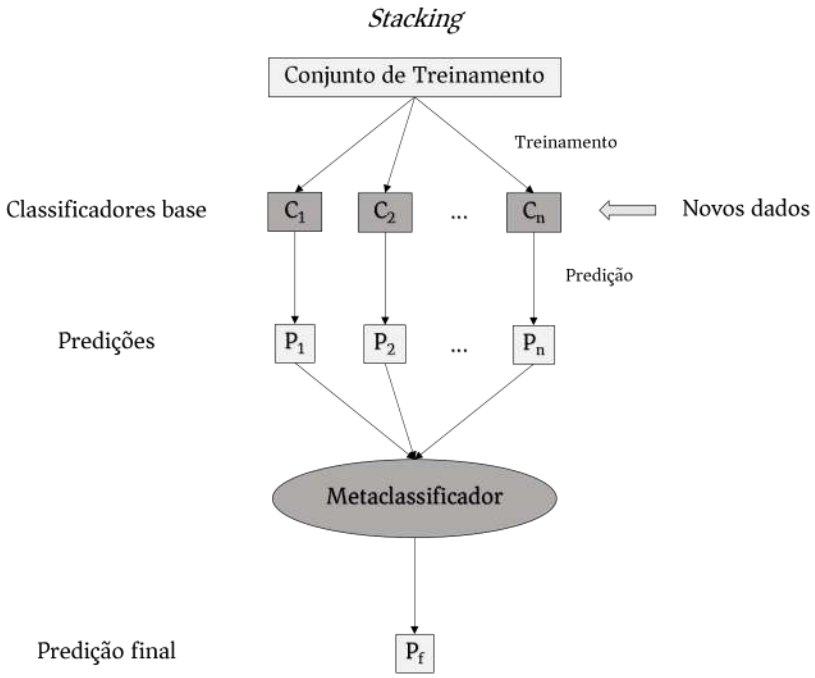


Figura 2.11: Algoritmo *stacking*. As saídas dos classificadores-base são utilizadas como entradas de um metaclassificador que fornece uma predição final.

O algoritmo *Stacking* também pode ser aplicado na predição numérica. Nesse caso, os modelos de nível 0 e o modelo de nível 1 predizem valores numéricos. O mecanismo básico continua o mesmo. A única diferença reside na natureza dos dados de nível 1. Cada atributo do meta conjunto de dados representa a previsão numérica feita por um dos modelos de nível 0 e, em vez de um valor de classe, a saída é um valor numérico relacionado aos exemplos de treinamento do nível 1 [102].

Capítulo 3

Revisão de Literatura

Ensaios genotípicos tradicionais utilizados na determinação de resistência do HIV-1 aos ARVs têm uma capacidade de detecção de mutações presentes apenas nas populações virais com frequências acima de 15-25% [21–23], não conseguindo identificar variantes virais minoritárias.

Uma série de ensaios ultrassensíveis, como a qPCR e o NGS, é capaz de detectar mutações em uma frequência muito menor do que os métodos padrão, alcançando níveis de 0,1-1% [20, 24]. Muitos trabalhos têm estudado a relação entre a presença das mutações minoritárias resistentes (MMR) e a resposta aos tratamentos.

No trabalho de Johnson e colaboradores (2008) [25], foram desenvolvidos dois estudos baseados na utilização da qPCR: um transversal e um caso-controle. O estudo transversal contou com a participação de indivíduos HIV+ recém-diagnosticados, sem uso prévio de ARVs e sem mutações detectáveis (vírus selvagem), e indivíduos HIV+ com uma ou mais mutações resistentes, segundo o sequenciamento de Sanger. Oito ensaios de qPCR foram usados para identificar algumas MMRs com frequências entre 0,4 e 19%. O ensaio sensível de qPCR identificou de uma a três MMRs em 17% dos indivíduos que possuíam o vírus selvagem, com 2% desses contendo mutações resistentes a duas classes de ARVs. Nos indivíduos que já possuíam mutações resistentes, 10% deles apresentaram pelo menos uma MMR diferente das mutações previamente presentes. No estudo de caso-controle, o grupo avaliou o impacto de três mutações em indivíduos que iniciaram um tratamento com EFV, mas sem registro prévio de ARVs e sem evidência de resistência segundo o sequenciamento de Sanger. O ensaio sensível identificou 3% dos indivíduos com uma ou duas MMRs. Em relação aos indivíduos que apresentaram falha virológica, 7% deles apresentavam MMRs no início do acompanhamento, enquanto que, naqueles que obtiveram sucesso no tratamento, apenas 0,9% apresentavam MMRs, indicando, dessa forma, que as variantes minoritárias poderiam ter consequências clínicas para o paciente.

Simen e colaboradores, em 2009 [26], desenvolveram um estudo com o intuito de determinar o impacto das MMRs nos desfechos clínicos, utilizando como métodos

de genotipagem o sequenciamento de Sanger e a plataforma 454 Life Sciences. O grupo observou que a técnica de NGS, por ter maior sensibilidade, encontrou mais mutações que a genotipagem padrão (28% vs. 14%). Entre os indivíduos que já haviam feito uso de NNRTIs, todos que apresentavam uma mutação de resistência a esse regime identificada pelo NGS tiveram falha virológica, com um risco maior do que aqueles que não apresentavam mutações resistentes aos NNRTIs.

No mesmo ano, Metzner e equipe [27] também avaliaram o papel das MMRs em pacientes com falha virológica. Nesse trabalho, foram acompanhados 4 indivíduos HIV+ com falha precoce a um regime de NNRTI e 18 indivíduos, no grupo controle, sem falha associada ao mesmo regime de tratamento, com subtipos e cargas virais comparáveis às dos 4 indivíduos com falha virológica. O estudo de caso-controle foi realizado por meio de ensaios de qPCR alelo-específica. Os resultados foram comparados com os da análise de fenotipagem virtual, utilizando o sequenciamento de Sanger. Eles observaram que antes do início do regime terapêutico, nenhum dos indivíduos apresentava resistência, segundo a genotipagem tradicional. Entretanto, quando utilizado o ensaio sensível, todos os indivíduos, que posteriormente apresentaram falha virológica, estavam infectados por vírus resistentes em baixas frequências (0,07-2,0%) antes do tratamento, com 1 a 4 mutações por paciente. No grupo controle, apenas 3 dos 18 indivíduos foram detectados com variantes minoritárias. Ao longo das semanas de acompanhamento, observou-se que a maioria das quasispécies minoritárias nos 4 indivíduos com falha foi rapidamente selecionada, tornando-se a principal população viral dentro de algumas semanas após os pacientes começarem a ART. O estudo observou um rápido crescimento das quasispécies minoritárias resistentes aos medicamentos, não detectadas no início do estudo por genotipagem convencional. Sob a pressão evolutiva artificial exercida pelos medicamentos, essas variantes com menos *fitness* em situação normal se tornam comparativamente mais eficientes que as majoritárias. Dessa forma, as variantes minoritárias podem se tornar a população viral majoritária e, posteriormente, levar à falha virológica precoce apesar da excelente aderência e da utilização de um potente regime terapêutico.

Em 2009, Le e colaboradores [28] examinaram a prevalência e os padrões de MMRs em 22 indivíduos HIV+ que experimentaram falha virológica, assim como o efeito dessas mutações nas classificações de suscetibilidade segundo o algoritmo HIVdb. Nesse trabalho, os resultados do sequenciamento de Sanger foram comparados aos do sequenciador Genome Sequencer FLX. O NGS foi capaz de detectar MMRs em todos os 22 indivíduos, enquanto que o método de Sanger não foi capaz de detectar MMRs em 19 indivíduos. No total, as MMRs foram responsáveis por 90 das 247 mutações (36%) detectadas por NGS, sendo que a maioria destas (95%) não foi detectada pela genotipagem padrão. As diferenças nos números de mutações detectadas pelos dois métodos se mostraram estatisticamente significantes para as

três classes de inibidores NRTI, NNRTI e PI. As MMRs com frequências menores que 13%, detectadas pelo NGS, não foram detectadas pelo sequenciamento de Sanger e apenas 57% das mutações com frequência entre 20 e 30,5% foram identificadas pela genotipagem padrão. Ao utilizar o HIVdb, 77% dos indivíduos tiveram seu nível de resistência aumentado com relação a, pelo menos, um ARV devido à presença das MMRs. Em metade dos indivíduos foi evidenciada resistência até então desconhecida ao menos a um ARV pelo mesmo motivo. O estudo mostrou que as MMRs são comumente identificadas em indivíduos que já fizeram uso de ARVs no momento da falha virológica e que o conhecimento sobre a eventual presença dessas MMRs pode auxiliar na tomada de decisão em relação aos regimes terapêuticos a serem adotados.

Boltz e colaboradores [29], em 2011, observaram o papel das variantes minoritárias do HIV-1 na falha da ART contendo o fármaco NVP em mulheres previamente expostas a uma dose única do mesmo medicamento (*single-dose nevirapine*, sdNVP). As mulheres que não apresentavam resistência a NNRTI por genotipagem padrão, mas apresentavam mutações K103N ou Y181C detectadas por um ensaio de PCR alelo-específico, apresentaram um risco de falha virológica ou morte três vezes maior que as mulheres sem essas mutações.

No estudo de Cozzi-Lepri e equipe (2014) [103], as MMRs também foram avaliadas com o intuito de verificar o papel dessas mutações nos desfechos virológicos de indivíduos HIV+ que faziam uso de NNRTIs de primeira linha. O estudo contou com 76 indivíduos com falha virológica e outros 184 com sucesso no tratamento até o momento do estudo. As MMRs foram consideradas aquelas com frequência entre 1-25% e detectadas pelo sequenciador Genome Sequencer FLX. O grupo observou que 31,6% dos casos apresentavam MMRs antes da falha virológica e 16,8% dos controles apresentavam MMRs. Além disso, a detecção de pelo menos uma MMR comparada a nenhuma se mostrou associada a um aumento do risco de falha virológica.

Vandenhende e colaboradores (2014) [30] também desenvolveram um estudo com o objetivo de determinar a prevalência das MMRs bem como o seu efeito sobre a resistência aos ARVs antes do tratamento e na falha virológica. O estudo contou com a participação de 29 indivíduos HIV+ sem tratamento prévio, tendo iniciado a ART entre 2000 e 2009 e experimentado falência virológica. As mutações de resistência da RT e da PR foram identificadas utilizando o sequenciamento de Sanger e o NGS, com a plataforma Genome Sequencer Junior, antes do início da ART e na vigência de falha virológica. O estudo observou que o NGS revelou, em média, 1,35 vezes mais mutações resistentes do que o sequenciamento de Sanger, devido à detecção das MMRs, que foram encontradas em 21 pacientes no início do estudo e em 18 pacientes com falha virológica. O estudo observou que as MMRs detectadas antes do início do tratamento e na falha virológica podem aumentar a carga geral de resistência aos

ARVs. Entretanto, foi observado, por meio do algoritmo ANRS v22, que apenas em poucos casos essas mutações alteravam a predição de suscetibilidade ao tratamento prescrito.

Em 2015, Fonager e colaboradores [31] compararam as abordagens de NGS, por meio do sequenciador Illumina MiSeq, e do sequenciamento de Sanger na detecção de MMRs a três INSTIs: RAL, EVG e DTG. As duas técnicas de sequenciamento foram utilizadas para analisar as sequências contemplando a região codificante da IN de seis pacientes, além de amostras seriadas de dois pacientes. O sequenciamento de Sanger detectou resistência aos INSTIs e mutações acessórias em três dos pacientes (denominados INSTI Res+), enquanto não foram observadas resistência ou mutações acessórias nos três pacientes restantes (INSTI Res-). O NGS identificou 142 MMRs, das quais 46 foram identificadas pelo sequenciamento de Sanger. Essas mutações de resistência aos INSTIs e/ou mutações acessórias adicionais foram identificadas nos três pacientes INSTI Res+ e em um paciente INSTI Res-. O NGS demonstrou uma sensibilidade maior do que o sequenciamento de Sanger na identificação de MMRs aos INSTIs, mostrando ser uma ferramenta potencialmente útil no monitoramento das MMRs em pacientes que venham a fazer uso de INSTIs.

Como observado anteriormente, muitos dos estudos recentes que buscam avaliar a importância das variantes minoritárias no desempenho dos esquemas terapêuticos fazem uso das técnicas de NGS. Essas estratégias de sequenciamento em paralelo produzem um grande número de sequências curtas que podem apresentar erros intrínsecos às plataformas, resultando em problemas computacionais complexos [104]. É importante realizar uma análise cuidadosa desses dados, particularmente na detecção dessas variantes, que podem ser confundidas com erros de sequenciamento, uma vez que ambos apresentam frequências similares. Identificar variantes presentes nas populações virais é importante para compreender a progressão da doença, determinar o efeito do genótipo viral, otimizar os modelos de vacina, assim como identificar e detectar mutações resistentes aos ARVs.

Em 2010, Archer e colaboradores [105], desenvolveram um software para a análise de dados virais de NGS obtidos em plataformas como 454 Life Sciences, Illumina, Ion Torrent e Pacific Biosciences. O programa, denominado Segminator II, foi implementado em Java e funciona nos principais sistemas operacionais. O programa permite o mapeamento dos dados de NGS contidos em arquivos FASTQ, a exportação e tradução das sequências, a determinação dos nucleotídeos e resíduos de aminoácidos, assim como suas frequências, a geração de uma sequência consenso (sequência obtida pelas bases de maior frequência em cada posição), o alinhamento múltiplo de leituras e a detecção e anotação de variantes minoritárias. Nesse estudo, o software desenvolvido foi aplicado em dois indivíduos infectados pelo HIV-1 que não respondiam de forma ótima ao fármaco MVC. Em ambos os casos, o programa

detectou grupos distintos de MMRs presentes antes do tratamento, variando em frequências de 2,5 a 15%. O Segminator II já vem sendo utilizado em diferentes estudos, dentre eles, o de Gibson e equipe (2014) [106], que desenvolveram um ensaio genotípico do HIV-1 baseado no NGS, denominado DeepGen HIV, para avaliar simultaneamente as suscetibilidades do HIV-1 a todos os fármacos direcionados para as enzimas virais PR, RT e IN e para predizer o tropismo do correceptor de HIV-1.

Em 2012, Prosperi e Salemi [107] desenvolveram o programa QuRe (Quasispecies Reconstruction algorithm) visando a reconstrução de quasispécies virais a partir de sequências longas de NGS (>100 pb). O algoritmo foi desenvolvido na linguagem de programação Java, com sua utilização em linhas de comando. Ao final das análises, o programa fornece todas as variantes que cobrem o genoma de referência juntamente com a prevalência associada a cada uma delas. Segundo os autores, o QuRe determina uma redução das chances de construir falsos positivos, além de oferecer um mapeamento rápido e robusto.

No mesmo ano, Macalalad e equipe [108] apresentaram a ferramenta V-Phaser com o objetivo de identificar variantes biológicas raras em populações geneticamente heterogêneas a partir de dados de NGS. O método combina informações sobre a covariância entre as variantes observadas, a fim de aumentar a sensibilidade, e um algoritmo de maximização de expectativa que recalibra iterativamente os índices de qualidade das bases, com o intuito de aumentar a especificidade. No estudo, o programa apresentou sensibilidade e especificidade maiores que 97%. Segundo os autores, o V-Phaser é capaz de detectar, com segurança, variantes raras em diversas populações que ocorrem em frequências menores que 1%.

Em 2013, Yang e colaboradores [109] promoveram uma atualização do V-Phaser, desenvolvendo o V-Phaser 2.0. Dentre as novidades, a nova versão apresenta um menor tempo de execução e utilização de memória e permite determinar uma sequência consenso a partir do resultado do alinhamento. A ferramenta foi validada e comparada a outras duas, QuRe e V-Phaser original, apresentando resultados superiores em sensibilidade e especificidade, e menor custo em termos de tempo e utilização de memória. A linguagem de programação utilizada é a Perl e a utilização da ferramenta se dá por meio de linhas de comando.

Watson e equipe (2013) [110] também desenvolveram um programa específico para a análise de dados virais de NGS. O *pipeline*, nomeado QUASR (Quality Assessment of Short Read), visa minimizar os erros de sequenciamento de múltiplas plataformas de NGS, permitindo uma análise pós-mapeamento das variantes minoritárias presentes na população viral. Um *pipeline* é um sistema computacional que executa uma série de programas encadeados entre si, em que a saída de um torna-se a entrada de outro. Segundo os autores, o *pipeline* reduz significativamente o ruído relacionado ao erro em dados de NGS, resultando em um aumento da precisão de

mapeamento e na redução das mutações errôneas. O QUASR pode ser aplicado a qualquer banco de dados de NGS nos formatos SFF ou FASTQ (formatos de arquivos que contém as saídas do sequenciamento). Ele funciona com 2 *scripts* de pré-montagem, que preparam os dados brutos, e um conjunto de *scripts* de pós-montagem, que analisa os dados após o mapeamento. Ele foi escrito na linguagem de programação Python e todo processamento faz uso de linhas de comando.

Pode-se destacar três sistemas recentes voltados mais especificamente para a análise de dados de sequenciamento do genoma do HIV.

Em 2016, Riemenschneider e equipe [111] desenvolveram um serviço web, denominado SHIVA, para a predição de resistência aos fármacos PIs, NNRTIs, NRTIs e INSTIs, e também para a nova classe de fármaco de MI, o Bevirimat. Adicionalmente, é possível determinar o tropismo do HIV, que é essencial para o tratamento com EIs, como o MVC. A ferramenta SHIVA fornece 23 modelos preditivos de resistência aos ARVs e um modelo classificador do tropismo do HIV. Os dados de entrada devem estar expressos no formato FASTA. Os resultados da previsão são fornecidos como relatórios clínicos que são enviados por e-mail ao usuário. A plataforma foi desenvolvida em Java e os modelos de resistência e tropismos foram implementados no software R.

Adicionalmente, um aplicativo de genotipagem do HIV, denominado HyDRA Web, foi desenvolvido pela Agência de Saúde Pública do Canadá como parte da Iniciativa de Pesquisa e Desenvolvimento em Genômica. Esse programa tem como objetivo identificar a resistência do HIV a fármacos a partir de um conjunto de dados de NGS, podendo detectar mutações com até 1% de frequência. O programa foi escrito na linguagem Perl e usa o software Bowtie2 para o mapeamento das sequências, ferramenta especializada em genomas maiores, como os de mamíferos [112]. Ele foi desenvolvido para análise de sequenciamentos tanto do Illumina MiSEQ como também da plataforma Roche 454 e aceita dados nos formatos FASTQ e SFF. O HyDRA Web gera um relatório de resistência aos medicamentos utilizando as mutações listadas pela base de dados Stanford HIVdb. As análises são flexíveis dentro do programa, permitindo que o usuário ajuste alguns parâmetros de mapeamento [113].

Em 2017, Huber e colaboradores [70] desenvolveram um *pipeline* englobando diversas ferramentas disponíveis (bwa [114], GATK [115] e LoFreq2 [116]) na análise de dados de sequenciamento. O *pipeline*, denominado MinVar*,* permite que o usuário avalie sequências brutas geradas pelos sequenciadores e obtenha como resposta final uma lista de mutações de aminoácidos com suas respectivas frequências. Além disso, combinadas com informações da base de dados Stanford HIVdb, a ferramenta destaca as mutações que conferem resistência aos ARVs. O aplicativo foi escrito em Python e seu uso se dá por meio de linhas de comando. Os arquivos de entrada

são no formato FASTQ e o ponto de corte das frequências é de 1,5%. Segundo os autores, o programa é capaz de ser aplicado em sequências de diferentes plataformas de NGS e não há necessidade de especificar opções ou parâmetros de mapeamento, e pode ser expandido para análise de outras regiões do HIV (além do gene *pol*) e de outros vírus.

Conforme retratado acima, muitas são as ferramentas desenvolvidas ao longo dos anos com o intuito de avaliar cada vez mais precisamente os dados de NGS. Assim como os sistemas de detecção de quasispécies virais, os algoritmos de interpretação genotípica de resistência aos ARVs evoluíram, sendo atualizados continuamente e aprimorados na última década. Entretanto, algumas discrepâncias de classificação ainda são observadas entre eles.

Os três sistemas de interpretação de resistência aos ARVs mais comumente utilizados e publicamente disponíveis são o ANRS, HIVdb e Rega, todos atualizados regularmente. Esses algoritmos foram desenvolvidos a partir de diferentes conjuntos de dados, variando desde os subtipos do HIV até os perfis dos pacientes (nunca fizeram uso ou já experimentaram algum ARV), podendo gerar predições conflituosas da resposta terapêutica.

Em 2003, Ravela e colaboradores [15] avaliaram a concordância dos algoritmos ANRS v3-02, HIVdb v8-02, Rega v5.5 e VGI (Visible Genetics) v6, utilizando 2045 sequências. Os níveis de suscetibilidade foram classificados como suscetível (S), intermediário (I) e resistente (R). As interpretações foram consideradas concordantes quando os algoritmos apontavam para o mesmo nível de resistência. A discordância parcial foi considerada quando um algoritmo apresentava resultado suscetível e o outro intermediário, ou intermediário e resistente. Já a discordância total ocorreu quando um algoritmo classificava como suscetível e o outro como resistente. Das 30675 interpretações (2045 sequências $\times$ 15 ARVs avaliados), 4,4% foram completamente discordantes, com pelo menos um algoritmo atribuindo S e o outro R; 29,2% foram parcialmente discordantes, com pelo menos um algoritmo atribuindo S e outro I, ou pelo menos um atribuindo I e outro R; e 66,4% exibiram completa concordância, com todos os quatro algoritmos atribuindo a mesma interpretação.

No estudo de Vergne e colaboradores (2006) [16], os três principais algoritmos de interpretação foram comparados a fim de avaliar seus desempenhos na classificação de resistência em indivíduos infectados pelo HIV-1, subtipos não-B, tanto de indivíduos que não fizeram nenhum tipo de tratamento quanto daqueles com uso prévio de ARVs. As versões dos algoritmos utilizados foram: ANRS v3-02, HIVdb v8-03 e Rega v5.5. Como o HIVdb classifica em cinco níveis, as categorias suscetível e potencial resistência de baixo nível foram consideradas S, resistência de baixo nível foi considerada I e resistência intermediária e resistência de alto nível classificadas como R. As interpretações foram consideradas concordantes para um ARV em par-

ticular quando os três algoritmos atribuíram o mesmo nível de resistência (S, I, ou R) para uma dada sequência, e discordantes quando pelo menos um dos três algoritmos atribuiu um nível diferente de resistência. Além disso, foram analisadas as discordâncias em um sistema de dois níveis, em que o nível I foi considerado alternativamente S ou R. A maioria das sequências dos indivíduos sem e com tratamento prévio foi interpretada como S por todos os algoritmos. O Rega registrou mais sequências como I para PIs, enquanto que quase nenhuma sequência foi classificada como I pelo ANRS. Na população de indivíduos tratados, o número de I e R foi mais elevado para todos os algoritmos. Em indivíduos não tratados previamente, foram observados níveis mais elevados de discordância para PIs (0,60-39%) do que para NNRTIs e NRTIs (0-4%). Em contraste, nos indivíduos tratados, foram observadas mais discordâncias para NNRTIs e NRTIs (5-48%) do que PIs (10-31%).

Poonpiriya e colaboradores (2008) [17] compararam a classificação de seis sistemas de interpretação: ANRS v13.0, DMC (Detroit Medical Centre) v10-04, GAV (Grupo de Aconselhamento Virológico) v01-06, Rega v6.3, CHL (Centre Hospitalier de Luxembourg) v5.0 e HIVdb v4.1.9 para todos os ARVs de todas as classes de medicamentos. As amostras foram provenientes de pacientes tailandeses que já haviam feito uso de alguma terapia e que apresentaram pelo menos uma falha terapêutica, sendo a maioria das amostras pertencentes aos subtipos não-B ou à forma recombinante CRF01_AE. Os algoritmos apresentavam diferentes interpretações que também foram normalizadas em 3 níveis: S, I e R. O maior nível de concordância completa foi observada para PIs e NNRTIs (67%). Para os NRTIs, a concordância completa foi registrada em 52% dos casos, indicando maior variabilidade entre os interpretadores.

Yebra e equipe (2010) [18] avaliaram cinco algoritmos (ANRS v7-08, HIVdb v4.3.7, Rega v7.1.1, Geno2pheno v3.0 e RIS v2008) em termos de desempenho, quando aplicados a diferentes subtipos (B e não-B) e formas recombinantes do HIV-1. Todos os algoritmos foram normalizados em 3 níveis: S, I e R. As interpretações foram consideradas concordantes quando os algoritmos apontavam para a mesma categoria de resistência. A discordância parcial foi considerada quando um algoritmo apresentava resultado S e o outro I, ou I e R. Já a discordância total ocorreu quando um algoritmo classificava como S e o outro como R. As discordâncias (total ou parcial) foram significantemente maiores para os subtipos não-B, quando comparadas ao subtipo B, para os NNRTIs (13% vs. 2,9%) e similares para os NRTIs (16,7% vs. 14,8%) e PIs (excluindo o fármaco TPV) (8,6% vs. 5,7%). A partir desses resultados, os autores levantaram a necessidade de inclusão de mais amostras de subtipos não-B nas bases de dados dos diferentes algoritmos, a fim de melhorar o desempenho dos métodos para o conjunto de subtipos, numa escala global.

No estudo de Frentz e equipe (2010) [117], um grande conjunto de sequências

de indivíduos infectados pelo HIV-1, a maioria pertencente ao subtipo B, foi analisado juntamente com dados virológicos a fim de comparar os três sistemas de interpretação: ANRS v17, HIVdb v5.1.2 e Rega v8.0.1. Diferentemente dos outros trabalhos, neste estudo observou-se que os três sistemas de interpretação não diferiram em sua capacidade de predizer a resposta virológica. Além disso, quando observados os diferentes pontos de acompanhamento dos pacientes (12, 24 e 48 semanas), as predições entre os sistemas foram semelhantes.

Em 2015, Wagner e colaboradores [19] desenvolveram um estudo que visava avaliar a concordância dos sistemas ANRS, HIVdb e Rega e suas evoluções ao longo do tempo. Amostras foram avaliadas utilizando as versões dos sistemas de 2004 (ANRS-AC11-2004, Rega v6.2 e HIVdb v3.9) e de 2013 (ANRS-AC11-2012, Rega v8.0.2 e HIVdb v6.2.0). Os sistemas foram normalizados em 3 categorias (S, I e R) e os mesmos foram considerados discordantes quando um classificava como S e um outro como R. As versões de 2004 classificaram 55% dos vírus como R a pelo menos um ARV. Quando utilizadas as versões de 2013, esse percentual aumentou para 62%. Nas versões de 2004, as discordâncias foram de 24% para todas as amostras e 11% apresentaram discordâncias múltiplas, sendo mais frequentes para os NRTIs (17%), seguido dos PIs (8%) e NNRTIs (2%). Quando aplicada as versões de 2013 no mesmo conjunto de dados, 26% das amostras apresentaram resultados discordantes, com 10% de múltiplas discordâncias, sendo também mais frequente para os NRTIs (15%). PIs e NNRTIs apresentaram, respectivamente, frequências de 9% e 5%. Novas sequências clínicas foram utilizadas para avaliar as versões de 2013. Nas amostras mais recentes, 35% foram classificadas como R em pelo menos um ARV e 49% apresentaram pelo menos uma discordância, que foi maior para os PIs (43%). O estudo também observou que o ANRS se mostrou associado a maioria das discordâncias e que, apesar da evolução dos algoritmos, divergências significativas permanecem atualmente.

Capítulo 4

Materiais e Métodos

4.1 Multiclassificadores

Nesta seção é apresentada a metodologia aplicada no desenvolvimento dos multiclassificadores.

4.1.1 Conjunto de Dados

O conjunto de dados utilizado no desenvolvimento dos multiclassificadores foi obtido do banco de dados Stanford HIV Drug Resistance Database (http://hivdb.stanford.edu/pages/genopheno.dataset.html). Esse banco de correlação genótipo-fenótipo é composto por 1808 sequências da PR e 1498 sequências da RT de indivíduos HIV+, além de apresentar informações sobre o subtipo do HIV-1, o método de fenotipagem aplicado, os valores de suscetibilidade fenotípica e algumas informações de genotipagem, como os aminoácidos identificados em cada posição das enzimas.

Neste trabalho, apenas os resultados de suscetibilidade obtidos pelo teste fenotípico PhenoSense (Monogram Biosciences, South San Francisco, EUA) [118] foram selecionados, uma vez que somente os pontos de corte dessa metodologia estão disponíveis na literatura.

As sequências que apresentaram misturas (posições com dois ou mais aminoácidos) foram eliminadas da análise, a fim de evitar que a presença dessas misturas desencadeasse uma confusão nas correlações entre genótipo e fenótipo. As mutações foram definidas quando havia diferença entre os aminoácidos da sequência consenso do subtipo B e os aminoácidos encontrados no sequenciamento. O presente estudo incluiu 947 sequências de aminoácidos da PR (posições 1-99) e 598 sequências de aminoácidos da RT (posições 1-560).

A distribuição dos subtipos do HIV-1 é mostrada na Tabela 4.1. A maioria das sequências utilizadas neste estudo foi do subtipo B, com uma prevalência de 92,4%

e 94,0% para PR e RT, respectivamente.

Tabela 4.1: Distribuição dos subtipos do HIV-1 das 1545 amostras incluídas no estudo.

Subtipo	Protease	Transcriptase reversa
A	3	29
B	875	562
C	17	0
CRF01_AE	2	1
CRF02_AG	15	1
CRF06_cpx	1	0
D	4	0
F	3	0
G	24	4
H	0	0
J	1	0
U	2	1
Total	947	598

U: subtipo não classificado (*unclassified*)
Fonte: Stanford HIV Drug Resistance Database. Data de acesso: março de 2015.

4.1.1.1 Classificação das Sequências Segundo os Algoritmos de Interpretação Genotípica e a Fenotipagem

Cada sequência foi classificada pelos três principais algoritmos de interpretação genotípica: ANRS v25, HIVdb v7.0 e Rega v9.1.0 (versões disponíveis no período de desenvolvimento do estudo). Os algoritmos foram implementados no software R [119] a fim de agilizar o processo de classificação, permitindo que as sequências fossem categorizadas de uma só vez pelos três algoritmos.

As regras de cada sistema podem ser acessadas em seus respectivos sítios: ANRS (http://www.hivfrenchresistance.org/2016/Algo2016-hiv1.pdf), HIVdb (https://hivdb.stanford.edu/page/release-notes/#algorithms) e Rega (http://rega.kuleuven.be/cev/avd/files/software/rega_algorithm/Rega_HIV1_Rules_v9.1.0.pdf).

Para obter a classificação das sequências segundo a fenotipagem, foram utilizados pontos de corte clínicos e biológicos do ensaio PhenoSense disponibilizados na página http://www.monogrambio.com/hiv-tests/phenotype-assays/phenosense. Segundo o teste fenotípico, valores abaixo do ponto de corte indicam vírus sensíveis ao ARV e valores acima do limiar correspondem à resistência. Para alguns medicamentos, o ponto de corte é dado por uma faixa de valores. Caso o resultado se enquadre dentro desse intervalo, o vírus é considerado parcialmente sensível ao ARV. Na Tabela 4.2, encontram-se os pontos de corte de suscetibilidade para cada ARV.

Tabela 4.2: Pontos de corte/faixa de suscetibilidade do HIV-1 aos antirretrovirais segundo o teste fenotípico PhenoSense. Os valores em negrito correspondem aos pontos de corte clínicos, os demais são biológicos.

Classe	Antirretroviral	Ponto de Corte/Faixa
PI	Atazanavir (ATV)	**5,2**
	Darunavir (DRV)	**(10 - 90)**
	Fosamprenavir (FPV)	**(4 - 11)**
	Indinavir (IDV)	**10**
	Lopinavir (LPV)	**(9 - 55)**
	Nelfinavir (NFV)	3,6
	Saquinavir (SQV)	(2,3 - 12)
	Tipranavir (TPV)	**(2 - 8)**
NRTI	Abacavir (ABC)	**(4,5 - 6,5)**
	Didanosina (ddI)	**(1,3 - 2,2)**
	Estavudina (d4T)	1,7
	Lamivudina (3TC)	**3,5**
	Tenofovir (TDF)	**(1,4 - 4)**
	Zidovudina (AZT)	1,9
NNRTI	Efavirenz (EFV)	3
	Etravirina (ETR)	**(2,9 - 10)**
	Nevirapina (NVP)	4,5
	Rilpivirina (RPV)	2,5

Fonte: PhenoSense®. Data de acesso: março de 2015.

As classificações dos algoritmos de interpretação genotípica foram categorizadas de acordo com o número de níveis apresentados pela fenotipagem. Essa padronização foi realizada a fim de facilitar a comparação dos algoritmos entre si e com o teste fenotípico. Para a maioria dos medicamentos, os algoritmos ANRS e Rega apresentam três níveis de classificação (suscetível, resistência intermediária ou possível resistência e resistência), enquanto que o HIVdb apresenta cinco níveis de suscetibilidade (suscetível, potencial resistência de baixo nível, resistência de baixo nível, resistência intermediária e resistência de alto nível). Na Tabela 4.3, está representada a padronização dos níveis de suscetibilidade para os algoritmos de acordo com o número de categorias da fenotipagem.

Os fármacos FPV e ddI, apesar de apresentarem três níveis de suscetibilidade segundo a fenotipagem, foram categorizados somente em suscetível ou resistente, uma vez que o algoritmo ANRS apresentava regras de classificação referentes a apenas duas classes.

4.1.1.2 Separação em Conjunto de Treinamento e Teste

Para o desenvolvimento e avaliação dos multiclassificadores, o conjunto de dados foi dividido em duas partições: conjunto de treinamento, destinado ao desenvolvimento

Tabela 4.3: Padronização dos níveis de suscetibilidade dos algoritmos de interpretação genotípica de acordo com as categorias da fenotipagem.

Genotipagem		Fenotipagem	
		Duas categorias	Três categorias
ANRS	Suscetível	Suscetível	Suscetível
	Possível resistência	Resistente	Resistência intermediária
	Resistente		Resistente
HIVdb	Suscetível	Suscetível	Suscetível
	Potencial resistência de baixo nível		
	Resistência de baixo nível	Resistente	Resistência intermediária
	Resistência intermediária		
	Resistência de alto nível		Resistente
Rega	Suscetível	Suscetível	Suscetível
	Resistência intermediária	Resistente	Resistência intermediária
	Resistente		Resistente

dos modelos, contemplando 75% do conjunto original; e conjunto de teste, reservado para a avaliação de desempenho das estratégias de multiclassificação aplicadas neste trabalho, contemplando os 25% dos dados restantes. Os dados foram divididos aleatoriamente, respeitando a proporção entre as classes do teste fenotípico.

Na Tabela 4.4, encontram-se o número de sequências presentes no conjunto original para cada ARV e a distribuição dos dados, segundo os níveis de suscetibilidade da fenotipagem.

Como pode ser observado na Tabela 4.4, há um grande desbalanceamento entre as categorias da fenotipagem para os fármacos DRV, TPV e TDF. Um conjunto de dados está desbalanceado quando uma categoria apresenta um número muito maior de instâncias do que a(s) outra(s). Além disso, o número de sequências para a ETR e a RPV é muito reduzido. Dessa forma, os ARVs DRV, TPV, TDF, ETR e RPV foram excluídos do estudo a fim de não comprometerem na divisão dos dados em treinamento e teste.

4.1.1.3 Balanceamento do Conjunto de Treinamento

Após a divisão do conjunto original de dados, o conjunto de treinamento foi balanceado para cada ARV. Essa estratégia foi adotada a fim de evitar o desenvolvimento de classificadores com bom desempenho restrito à classificação dos dados majoritários, comprometendo características importantes dos modelos, como a robustez e a capacidade de generalização. Apenas os fármacos NFV, AZT e NVP não passaram por essa etapa por já apresentarem categorias balanceadas, como pode ser observado na Tabela 4.4.

Para equilibrar as categorias do conjunto de treinamento, foram utilizados

Tabela 4.4: Número de sequências presentes no conjunto de dados para cada antirretroviral e distribuição segundo as classes da fenotipagem. Alguns antirretrovirais não apresentam a classe I (resistência intermediária), uma vez que o ponto de corte da fenotipagem apenas classifica em duas classes: S (suscetível) e R (resistente).

Classe	Antirretroviral	Número de sequências	Proporção S/I/R (%)
PI	Atazanavir (ATV)	541	335/206 (61,9/38,1)
	Darunavir (DRV)	296	222/73/1 (75,0/247/0,3)
	Fosamprenavir (FPV)	868	581/287 (66,9/33,1)
	Indinavir (IDV)	897	633/264 (70,6/29,4)
	Lopinavir (LPV)	704	440/139/125 (62,5/19,7/17,8)
	Nelfinavir (NFV)	926	445/481 (48,1/51,9)
	Saquinavir (SQV)	903	541/131/231 (59,9/14,5/25,6)
	Tipranavir (TPV)	359	276/79/4 (76,9/22,0/1,1)
NRTI	Abacavir (ABC)	442	290/70/82 (65,6/15,8/18,6)
	Didanosina (ddI)	448	177/271 (39,5/60,5)
	Estavudina (d4T)	447	286/161 (64,0/36,0)
	Lamivudina (3TC)	448	167/281 (37,3/62,7)
	Tenofovir (TDF)	358	238/99/21 (66,5/27,7/5,8)
	Zidovudina (AZT)	445	227/218 (51,0/49,0)
NNRTI	Efavirenz (EFV)	497	282/215 (56,7/43,3)
	Etravirina (ETR)	146	106/19/21 (76,6/13,0/14,4)
	Nevirapina (NVP)	506	255/251 (50,4/49,6)
	Rilpivirina (RPV)	79	55/24 (69,6/30,4)

métodos de sobreamostragem (*oversampling*) e subamostragem (*undersampling*). A sobreamostragem é um método que visa balancear a distribuição das classes por meio da replicação de exemplos aleatórios da classe minoritária do conjunto de dados original. A subamostragem, por sua vez, tem como objetivo equilibrar a distribuição de classes removendo aleatoriamente alguns exemplos da classe majoritária do conjunto original [120].

Para os ARVs com três níveis de suscetibilidade, foi aplicada uma metodologia híbrida, com sobreamostragem e subamostragem. Primeiramente, foi identificada o nível que possuía valores intermediários entre as três categorias. Após a identificação, as classes minoritárias e majoritárias foram balanceadas em relação à classe de valor intermediário. Por exemplo, para o SQV, a classe resistente possui um número de exemplos (174 instâncias) situado entre a classe suscetível (406 instâncias) e a resistência intermediária (99 instâncias). Portanto, as amostras suscetíveis (majoritária) foram subamostradas e as de resistência intermediária foram sobreamostradas em relação às amostras resistentes.

Na sobreamostragem, as instâncias adicionadas foram replicadas a partir das existentes e não criadas. Para fármacos com duas categorias, apenas a técnica de subamostragem foi aplicada.

O balanceamento dos dados foi realizado por meio da função *ovun.sample* do pacote *ROSE* do software R [121]. Essa função realiza a reamostragem de modo que a proporção de exemplos da classe positiva seja aproximadamente igual a uma probabilidade p [121]. No presente trabalho, foi adotado $p = 0,5$. Na Tabela 4.5, encontra-se a distribuição dos dados do conjunto de treinamento antes e após as estratégias de balanceamento aplicadas.

Tabela 4.5: Distribuição dos dados no conjunto de treinamento antes e após o balanceamento. Os valores antes do balanceamento correspondem a 75% dos dados originais.

Classe	Antirretroviral	Antes do balanceamento (S/I/R)	Após o balanceamento (S/I/R)
PI	Atazanavir (ATV)	252/155	143/155
	Fosamprenavir (FPV)	436/216	202/216
	Indinavir (IDV)	475/198	182/198
	Lopinavir (LPV)	330/105/94	106/105/106
	Saquinavir (SQV)	406/99/174	161/161/174
NRTI	Abacavir (ABC)	218/53/62	58/58/62
	Didanosina (ddI)	133/204	133/131
	Estavudina (d4T)	215/121	121/121
	Lamivudina (3TC)	126/211	126/126
NNRTI	Efavirenz (EFV)	212/162	149/162

4.1.2 Desenvolvimento dos Multiclassificadores

Para combinar os algoritmos ANRS, HIVdb e Rega, a fim de obter um classificador único, três abordagens foram aplicadas no presente trabalho: um esquema simples de votação por maioria (Voto Majoritário, VM), a escolha do algoritmo com melhor medida de desempenho (denominada aqui como Melhor Sistema, MS) e uma adaptação da técnica *stacking*.

Voto Majoritário A classificação final de resistência de cada instância foi dada pela categoria com maior número de votos entre os três algoritmos de interpretação genotípica.

Melhor Sistema Nesta abordagem, o algoritmo com melhor medida de desempenho foi selecionado para fornecer a classificação final de cada instância.

O desempenho de cada algoritmo de interpretação genotípica foi avaliado utilizando o conjunto de treinamento. A medida-F, que representa a média harmônica ponderada da precisão (*precision*) e revocação (*recall*), foi usada para selecionar a

melhor estratégia. A precisão, equivalente ao valor preditivo positivo neste contexto, indica a proporção de verdadeiros positivos em relação ao total de positivos classificados pelo algoritmo. A revocação, equivalente à sensibilidade nesta abordagem, indica a proporção de verdadeiros positivos em relação ao total de exemplos da classe positiva. Dessa forma, a medida-F tem um significado intuitivo. Ela diz o quão preciso é o classificador (quantas instâncias resistentes ele classifica corretamente), bem como o quanto ele é robusto (não perde um número significativo de instâncias resistentes). Na Figura 4.1 está representado o cálculo da precisão e revocação de maneira ilustrativa.

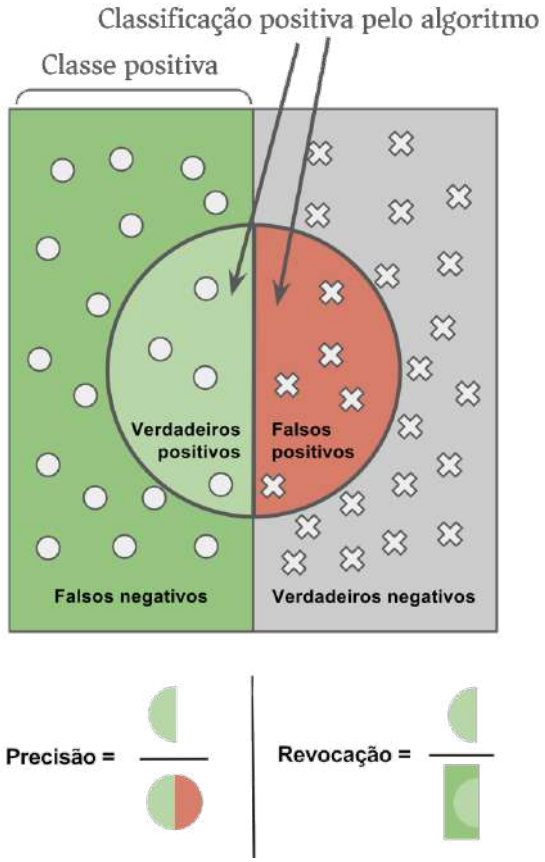


Figura 4.1: Precisão e revocação. Adaptado de Walber, 2014 (https://commons.wikimedia.org/wiki/File:Precisionrecall.svg)

A partir dessas duas medidas de desempenho, a medida-F é dada pela fórmula

$$F = 2\frac{precisão \times revocação}{precisão + revocação} \tag{4.1}$$

Stacking Na estratégia adaptada da técnica *stacking*, descrita na seção 2.5.3, os sistemas de interpretação genotípica ANRS, HIVdb e Rega foram utilizados como classificadores base e as classificações do conjunto de treinamento fornecidas pelos três algoritmos foram utilizadas como entrada de um metaclassificador.

Os metaclassificadores aplicados foram os métodos *naïve Bayes* (NB) e *k-Nearest Neighbors* (k-NN). O ensaio PhenoSense foi utilizado como referência, fornecendo o padrão ouro para os níveis de suscetibilidade para cada sequência e ARV.

O algoritmo NB compreende uma família de classificadores baseada no teorema de Bayes, que considera os atributos condicionalmente independentes entre si, ou seja, a informação de um evento não é informativa sobre nenhum outro. O algoritmo tem como objetivo calcular a probabilidade que uma instância desconhecida pertença a cada uma das possíveis classes [122].

Baseado em um conjunto de treinamento prévio, o algoritmo recebe como entrada uma nova instância desconhecida, que não possui classificação, e retorna como saída a classe mais provável de acordo com cálculos de probabilidade.

Dada uma instância a ser classificada, representada pelo vetor $\mathbf{x} = (x_1, \ldots, x_n)$, sendo n o número de variáveis aleatórias independentes discretas, o NB atribui a essa instância uma probabilidade a posteriori $p(y_k \mid x_1, \ldots, x_n)$ para cada uma das k classes (y). Usando o teorema de Bayes, essa probabilidade condicional pode ser decomposta em:

$$p(y_k \mid \mathbf{x}) = \frac{p(y_k)p(\mathbf{x} \mid y_k)}{p(\mathbf{x})} \tag{4.2}$$

em que $p(y_k)$ representa a probabilidade a priori da classe y_k, $p(\mathbf{x} \mid y_k)$ é a probabilidade condicionada de $\mathbf{x}$ à classe y_k e $p(\mathbf{x})$ é a probabilidade total de $\mathbf{x}$.

Como $\mathbf{x} = (x_1, \ldots, x_n)$, tem-se:

$$p(y_k \mid x_1, \ldots, x_n) = \frac{p(y_k)p(x_1, \ldots, x_n \mid y_k)}{p(x_1, \ldots, x_n)} \tag{4.3}$$

Para calcular a classe mais provável da nova instância, calcula-se a probabilidade de todas as possíveis classes e, no final, escolhe-se a classe com a maior probabilidade como rótulo da nova instância. Para tal, deve-se maximizar o valor do numerador $p(y_k)p(x_1, \ldots, x_n \mid y_k)$ ou minimizar o valor do denominador $p(x_1, \ldots, x_n)$. Como o denominador $p(x_1, \ldots, x_n)$ é uma constante, pois não depende da classe que se está procurando, pode-se desconsiderá-lo (denominador igual a 1) no teorema de Bayes, resultando em:

$$\arg\max p(y_k \mid x_1, \ldots, x_n) = \arg\max p(y_k)p(x_1, \ldots, x_n \mid y_k) \tag{4.4}$$

Como o classificador NB assume que todos os atributos da instância que se deseja classificar são independentes, pode-se simplificar o cálculo do termo $p(x_1, \ldots, x_n \mid y_k)$ por $p(x_1 \mid y_k) \times \ldots \times p(x_n \mid y_k)$. Assim, a fórmula final do classificador é:

$$\arg\max p(y_k \mid x_1, \ldots, x_n) = \arg\max \prod_i p(x_i \mid y_k) \times p(y_k) \tag{4.5}$$

De acordo com a equação 4.5, deve-se calcular os termos $p(x_i \mid y_k)$ e $p(y_k)$. O último, como descrito anteriormente, representa o número de casos pertencentes à classe em questão sobre o número total de casos, e $p(x_i \mid y_k)$, por sua vez, é o número de casos pertencentes à classe em questão com o atributo x_i sobre o número total de casos da classe.

Caso alguma probabilidade condicionada seja nula, ou seja, um determinado atributo x_i não seja observado em alguma classe y_k no conjunto de treinamento, é aplicada a correção de Laplace, normalmente dada por:

$$p(x_i \mid y_k) = \frac{N_{i,k} + 1}{N_c + |x_i|} \tag{4.6}$$

em que $p(x_i \mid y_k)$ é a nova probabilidade condicionada, $N_{i,k}$ é o número de vezes que o atributo x_i surge na classe y_k, N_c é o número de atributos na classe y_k e $|x_i|$ é o número de possíveis valores que o atributo x_n pode tomar.

Neste estudo, o pacote do R *klaR* [123] foi utilizado para desenvolver os modelos com *naïve Bayes*.

O algoritmo k-NN é um dos métodos não paramétricos mais simples utilizado em problemas de classificação. Ele adota uma medida de similaridade para determinar a proximidade de uma instância a outros exemplos. Nesse método, é assumido que os exemplos são independentes e identicamente distribuídos. Dessa forma, as instâncias que apresentam maior proximidade entre si possuem a mesma classificação [124].

Na classificação pelo k-NN, cada instância é definida por um conjunto de atributos $(x_1, x_2, \ldots, x_n)$ e todas as instâncias são representadas pelo mesmo número de atributos, de modo que elas sejam demonstradas como pontos espaciais de dimensão n. Um desses atributos representa a classificação real da instância, cujos valores são previstos para novas instâncias não classificadas.

Cada exemplo a ser classificado recebe uma classe por meio de um voto majoritário dentre os k casos vizinhos mais próximos no espaço de atributos. Se $k = 1$, então o objeto recebe a mesma classe do único vizinho mais próximo. Normalmente, os valores de k escolhidos são número inteiros positivos ímpares a fim de evitar empates e, consequentemente, a indefinição da classe final.

A proximidade de um vizinho é definida com base nos atributos da nova instância e dos exemplos de treinamento. Aqueles que possuem atributos mais semelhantes aos da nova instância serão considerados mais próximos. O algoritmo k-NN calcula a distância entre os atributos da nova instância e dos exemplos anteriores para descobrir a classe. Para calcular a distância entre as instâncias que apresentam atributos do mesmo tipo, algumas funções de distância podem ser empregadas [125]. Nas fórmulas abaixo, há três exemplos de funções em que $\mathbf{x}_i$ e $\mathbf{x}_j$ são duas instâncias de entrada e p é o número de atributos:

- Distância Euclidiana:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{s=1}^{p} (x_{is} - x_{js})^2} \tag{4.7}$$

- Distância Manhattan:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{s=1}^{p} |x_{is} - x_{js}| \tag{4.8}$$

- Distância Minkowski:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{i=1}^{p} |x_{is} - x_{js}|^q \right)^{1/q} \tag{4.9}$$

Como pode ser observado, a distância Manhattan e a distância Euclidiana são casos particulares da distância Minkowski quando $q = 1$ e $q = 2$, respectivamente.

Na Figura 4.2, está representado um exemplo da classificação pelo algoritmo k-NN de acordo com o valor de k escolhido.

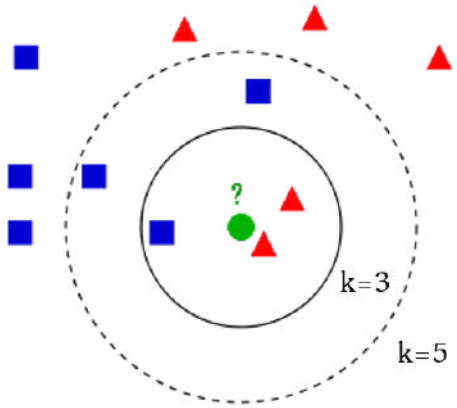


Figura 4.2: Exemplo de classificação pelo algoritmo k-NN. Se $k = 3$ (círculo de linha contínua) é atribuído ao círculo verde (exemplo a ser classificado), pelo voto majoritário, a classe "triângulo". Se $k = 5$ (círculo tracejado) é atribuída a classe "quadrado" (3 quadrados versus 2 triângulos). Adaptado de Antti Ajanki, 2007 (https://commons.wikimedia.org/wiki/File:KnnClassification.svg).

No presente trabalho, foi utilizada uma versão de k-NN ponderado, implementada pelo pacote do R denominado *kknn* [126, 127]. Nesta versão, observações do conjunto de treinamento que são particularmente próximas da nova observação a ser classificada recebem um peso maior na decisão do que os vizinhos que estão longe da nova instância.

No caso das variáveis explicativas nominais, elas são convertidas em variáveis *dummy*. Para atribuir o mesmo peso a cada variável no cálculo das distâncias, é preciso padronizar os valores. A técnica de padronização oferecida por esse pacote se

baseia no traço (soma dos elementos da diagonal principal) da matriz de covariância das *dummies* correspondentes. O termo utilizado como divisor normalizante é:

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m} var\left(v_i\right)} \tag{4.10}$$

em que m é o número de classes das variáveis e v_i são as variáveis *dummy*.

Essa padronização de todas as variáveis *dummy* com o mesmo desvio padrão (médio) é necessária uma vez que as diferenças entre as classes devem ser tratadas simetricamente, independentemente das diferenças nos desvios padrão de uma única variável *dummy*.

Considere $\mathcal{T} = \{(y_i, \mathbf{x}_i), i = 1, ..., n_{\mathcal{T}}\}$ um conjunto de treinamento formado pelas observações $\mathbf{x}_i$ com sua respectiva classe y_i. Considere também $\mathbf{x}$ como uma nova instância cuja classe y precisa ser predita. O algoritmo k-NN ponderado irá normalizar as distâncias e transformá-las em pesos utilizando diferentes funções de kernel.

A largura de banda da função de kernel é automaticamente selecionada de acordo com a distância do primeiro vizinho $\mathbf{x}_{(k+1)}$ que não é mais levado em consideração. Isto é feito implicitamente pela padronização de todas as outras distâncias com a distância do $(k+1)$*ésimo* vizinho. Dessa forma, o algoritmo busca os $k+1$ vizinhos mais próximos de $\mathbf{x}$ de acordo com a função de distância $d(\mathbf{x}, \mathbf{x}_i)$ e utiliza o $(k+1)$*ésimo* vizinho na padronização das k menores distâncias:

$$D_{(i)} = D(\mathbf{x}, \mathbf{x}_{(i)}) = \frac{d(\mathbf{x}, \mathbf{x}_{(i)})}{d(\mathbf{x}, \mathbf{x}_{(k+1)})} \tag{4.11}$$

Em seguida, as distâncias normalizadas $D_{(i)}$ são transformadas, utilizando uma função de kernel $K(.)$, em pesos $w_{(i)} = K(D_{(i)})$. É adicionado um $\epsilon > 0$ a fim de evitar possíveis distâncias iguais a zero. A classe y da observação $\mathbf{x}$ será definida pela classe que apresenta uma maioria ponderada dos k vizinhos mais próximos:

$$\hat{y} = max_r \left(\sum_{i=1}^{k} w_{(i)} I(y(i) = r) \right) \tag{4.12}$$

em que $I(\cdot)$ é uma função identidade e r o índice da classe.

4.1.2.1 Seleção dos Modelos

Nesta etapa, para encontrar o conjunto de hiperparâmetros apropriados para os modelos obtidos a partir dos algoritmos NB e k-NN, foi utilizada a função *train* do pacote *caret* do software R [128]. Essa função tem como objetivo obter um modelo final "ótimo" a partir da escolha de uma combinação de hiperparâmetros

com melhor resultado para uma medida de desempenho. No presente trabalho, dentre as medidas disponíveis, foi utilizada a acurácia, que fornece a proporção de acertos realizados pelo classificador.

Para um algoritmo particular previamente escolhido, diversas combinações de hiperparâmetros são criadas pela função *train* e o modelo é treinado com dados ligeiramente diferentes para cada um desses arranjos. Para cada iteração da reamostragem, o desempenho dos modelos é calculado usando os dados não incluídos no treinamento, sendo obtidos a média e o desvio padrão da medida de desempenho para cada combinação de hiperparâmetros. Por *default*, a função escolhe automaticamente os hiperparâmetros de ajuste associados ao melhor valor da medida de desempenho escolhida. Os hiperparâmetros definidos compõe o modelo final e todo o conjunto de treinamento é usado para ajustar o modelo. Na Figura 4.3, está representado o algoritmo da função *train*.

```
1  Defina um conjunto de valores dos parâmetros do modelo para avaliar
2  para cada conjunto de parâmetros faça
3  |  para cada iteração da reamostragem faça
4  |  |  Separe algumas instâncias específicas
5  |  |  [Opcional] Pré-processe os dados
6  |  |  Ajuste o modelo com os exemplos restantes
7  |  |  Prediga as instâncias separadas anteriormente
8  |  fim
9  |  Calcule o desempenho médio por meio das predições das instâncias separadas
10 fim
11 Determine o melhor conjunto de parâmetros
12 Ajuste o modelo final com todos os dados de treinamento usando o conjunto de parâmetros determinado
```

Figura 4.3: Algoritmo da função *train* do pacote *caret*. Adaptado de https://topepo.github.io/caret/model-training-and-tuning.html.

No presente trabalho, o método de reamostragem utilizado na função *train*, em busca da melhor combinação de hiperparâmetros, foi a validação cruzada *k-fold* com repetição, com k igual a 10 e repetição igual a três vezes. A validação cruzada é um dos principais mecanismos usados na aprendizagem de máquinas para controlar o erro de generalização quando não há uma quantidade suficiente de dados para dividi-lo em subconjuntos de treino, validação e teste independentes.

Na validação cruzada *k-fold*, o conjunto de dados é dividido, aleatoriamente, em k grupos (*folds*) mutuamente exclusivos, com tamanhos aproximadamente iguais. A partir disto, um subconjunto é utilizado para teste e os *k-1* restantes são utilizados para treinar o modelo. Esse processo é realizado k vezes, alternando de forma circular o subconjunto de teste [129]. Este é usado na predição do modelo treinado e os resultados são expressos em algum tipo de medida de desempenho (por exemplo, acurácia). Por fim, é calculada a média das k estimativas de desempenho a fim de obter uma estimativa global. Quando se tem uma validação cruzada *k-fold* com repetição, todo o processo anterior é repetido n vezes.

Os hiperparâmetros de ajuste otimizados para o algoritmo NB, implementado pelo pacote *klaR* do software R, foram:

(i) correção de Laplace (presente ou não);

(ii) estimativa de densidade de *kernel* (se sim, é usada para estimar a densidade); e

(iii) largura de banda.

Esse pacote estima a probabilidade condicional $p(x_i \mid y_k)$ usando a estimativa de densidade de *kernel* (EDK) a partir de um conjunto de dados de treinamento com classe. A EDK é uma forma não-paramétrica para estimar a função de densidade de probabilidade de uma variável aleatória [130]. Assim, a fórmula para $p(x_i \mid y_k)$ fica

$$p(x_i \mid y_k) = \frac{1}{M_y h} \sum_{j=1}^{M_y} K(x_i, x_{j|i|y}) \tag{4.13}$$

em que K é a função *kernel*, M_y é o número de dados de entrada pertencentes à classe y_k, $x_{j|i|y}$é o valor do i-ésimo atributo da j-ésima entrada na classe y_k, e h é a largura de banda (parâmetro de alisamento).

Para o k-NN, os hiperparâmetros de ajuste da função *kknn* foram:

(i) o valor de p da distância Minkowski;

(ii) a função *kernel* usada; e

(iii) o número de vizinhos.

4.1.3 Avaliação dos Multiclassificadores

4.1.4 Testes Estatísticos

Para comparar o desempenho dos multiclassificadores desenvolvidos, de maneira independente, para cada ARV, o teste estatístico de Friedman [131, 132] foi aplicado neste estudo. Esse é um teste estatístico não-paramétrico que ranqueia diferentes classificadores para cada conjunto de dados separadamente. O modelo com melhor desempenho obtém uma colocação (*rank*) igual a um, o segundo melhor obtém uma colocação igual a dois, e assim por diante.

Para realizar o teste de Friedman, os dados são dispostos no formato apresentado na Tabela 4.6, em que as k linhas representam os grupos (classificadores) e as n colunas representam os blocos (conjuntos de dados) [133].

Considerando-se r_{ij} como a colocação do j-ésimo classificador no i-ésimo conjunto de dados, o teste de Friedman compara as colocações médias dos algoritmos $R_j = \frac{1}{n}\sum_i r_i^j$ e tem como hipótese nula a afirmação de que não há distinções entre os desempenhos dos classificadores (as colocações R_j são equivalentes) [132, 133]. A

Tabela 4.6: Tabelas cruzada dos dados utilizados no teste de Friedman.

Grupo	Bloco			
	1	2	$\cdots$	n
1	x_{11}	x_{12}	$\cdots$	x_{1n}
2	x_{21}	x_{22}	$\cdots$	x_{2n}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
k	x_{k1}	x_{k2}	$\cdots$	x_{kn}

Bloco				Soma dos
1	2	$\cdots$	n	*ranks*
r_{11}	r_{12}	$\cdots$	r_{1n}	R_1
r_{21}	r_{22}	$\cdots$	r_{2n}	R_2
$\vdots$	$\vdots$		$\vdots$	$\vdots$
r_{k1}	r_{k2}	$\cdots$	r_{kn}	R_n

estatística do teste tem uma distribuição qui-quadrado (χ^2) e é dada por:

$$\chi_F^2 = \left[\frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3n(k+1) \tag{4.14}$$

em que n representa o número de conjuntos de dados e k o número de classificadores.

Se o valor-p está abaixo do nível de significância escolhido, pode-se inferir que pelo menos dois dos classificadores apresentam desempenhos significativamente diferentes um do outro.

Neste estudo, para obter as colocações, a medida-F foi utilizada como resultado de cada classificador e a validação cruzada *10-fold* foi aplicada para obter 10 conjuntos de dados diferentes a partir do treinamento. O nível de significância considerado foi de 5% e a função *friedmanTest* do pacote *scmamp* do software R [134] foi a escolhida para executar o teste de Friedman.

Quando a hipótese nula é rejeitada, outro teste estatístico deve ser aplicado para determinar quais classificadores apresentaram uma diferença significativa. No presente trabalho, o teste de Holm [135] foi utilizado como teste estatístico *post hoc*. Para realizar esse teste, o classificador com a melhor colocação é selecionado como classificador-controle e uma comparação dois a dois dos demais classificadores com o controle é realizada. A hipótese nula afirma que o classificador-controle e o segundo utilizado na comparação pareada possuem a mesma colocação média.

Determinado o valor-p para cada comparação pareada, esses são ranqueados do menor para o maior e cada valor-p é multiplicado por um fator dado por $(n - k)$, em que n é o número de comparações e k é a colocação do valor-p. Se o valor-p final for menor que $0,05$, rejeita-se a hipótese nula e considera-se que há uma diferença de colocação entre os classificadores, indicando que o classificador-controle apresenta um melhor desempenho. O teste de Holm foi executado por meio da função *postHocTest* do pacote *scmamp* [134].

4.1.5 Medidas de Desempenho

O conjunto de teste independente foi utilizado para avaliar o desempenho das estratégias propostas por meio das seguintes medidas de desempenho: acurácia, sensibilidade, especificidade, valores preditivos positivo (VPP) e negativo (VPN) e medida-F.

As principais medidas podem ser calculadas a partir de uma matriz de confusão para duas classes, onde são representados quatro tipos de classificação, segundo o resultado do preditor (Tabela 4.7):

- Classificados como positivos e pertencentes à classe positiva (verdadeiros positivos, VP);
- Classificados como negativos, mas pertencentes à classe positiva (falsos negativos, FN);
- Classificados como positivos, mas pertencentes à classe negativa (falsos positivos, FP);
- Classificados como negativos e pertencentes à classe negativa (verdadeiros negativos, VN);

Tabela 4.7: Matriz de confusão para duas classes. VP: verdadeiros positivos; VN: verdadeiros negativos; FP: falsos positivos; FN: falsos negativos.

Classificador	Referência	
	Positivo	Negativo
Positivo	VP	FP
Negativo	FN	VN

VP: verdadeiros positivos; VN: verdadeiros negativos; FP: falsos positivos; FN: falsos negativos

A acurácia é definida como a proporção de acertos do classificador. No caso de duas classes, ela é dada pela fórmula:

$$Acurácia = \frac{VN + VP}{VP + FN + FP + VN} \tag{4.15}$$

Quando há mais de duas classes, a acurácia é calculada pela soma dos valores da diagonal principal dividida pelo número total de casos.

A sensibilidade é dada pela proporção de verdadeiros positivos em relação ao total da classe positiva. Nos casos de mais de duas classes, para o cálculo dessa medida, uma classe é definida como positiva e as demais são agrupadas e consideradas como negativas. A sensibilidade é calculada por:

$$Sensibilidade = \frac{VP}{VP + FN} \tag{4.16}$$

A especificidade, por sua vez, compreende a proporção de verdadeiros negativos em relação ao total da classe negativa. O mesmo procedimento para o cálculo da sensibilidade é feito para a especificidade quando há mais de duas classes. A fórmula da especificidade é dada por:

$$Especificidade = \frac{VN}{FP + VN} \tag{4.17}$$

Os VPP e VPN indicam, respectivamente, a probabilidade de um indivíduo com teste positivo ter realmente o problema diagnosticado e a probabilidade de um indivíduo com teste negativo ser realmente normal. Suas fórmulas são:

$$VPP = \frac{VP}{VP + FP} \tag{4.18}$$

$$VPN = \frac{VN}{FN + VN} \tag{4.19}$$

A medida-F, como mencionada anteriormente, é dada pela média harmônica do VPP (também chamado de *precision*) e da sensibilidade (também denominada *recall*) cuja fórmula pode ser escrita como

$$F = 2\,\frac{VPP \times S}{VPP + S} \tag{4.20}$$

Os ARVs que se mostraram associados a três desfechos (suscetível, resistência intermediária e resistente) tiveram suas categorias agrupadas, gerando três subconjuntos com classe binária:

(i) Suscetível e não suscetível (resistência intermediária + resistente);

(ii) Resistência intermediária e não intermediária (suscetível + resistente); e

(iii) Resistente e não resistente (suscetível + resistência intermediária).

Esse agrupamento foi realizado para o cálculo da medida-F, sensibilidade, especificidade, VPP e VPN. Para os ARVs com três níveis de suscetibilidade, a medida-F final para cada classificador foi dada pela média das três medidas encontradas quando utilizada a categorização binária. O pacote do R utilizado nesta etapa foi o *caret* [128].

Na Figura 4.4 encontra-se um fluxograma das etapas desempenhadas no desenvolvimento e avaliação dos multiclassificadores.

4.2 Desenvolvimento do Sistema de Identificação de Resistência aos Antirretrovirais - SIRA-HIV

Esta seção apresenta as fases de desenvolvimento do ambiente integrado destinado à identificação dos aminoácidos presentes nas posições de resistência e à classificação do nível de suscetibilidade do HIV-1 aos ARVs.

4.2.1 Organização do Sistema

O SIRA-HIV tem como objetivo realizar uma análise geral dos dados obtidos pelo sequenciamento de nova geração de amostras de pacientes HIV+.

O sistema se divide, basicamente, em três fases:

(i) Mapeamento das leituras geradas pelo NGS em relação a um genoma de referência do HIV-1;

(ii) Análise dos resultados do mapeamento e identificação dos aminoácidos presentes nas posições de resistência e;

(iii) Classificação do nível de suscetibilidade do vírus pelos algoritmos de interpretação genotípica.

Para o mapeamento das leituras, o programa Segminator II foi escolhido. Esse software, desenvolvido por Archer e colaboradores (2012) [136], realiza a análise de dados virais de NGS de diversas plataformas, como 454 Life Sciences, Illumina, Ion Torrent e Pacific Biosciences, permitindo um mapeamento das sequências. O mapeamento está relacionado com a identificação da região da sequência de referência a qual uma leitura pertence. O Segminator II foi implementado em Java, no formato de interface gráfica, sendo de fácil uso pelos usuários (Figura 4.5).

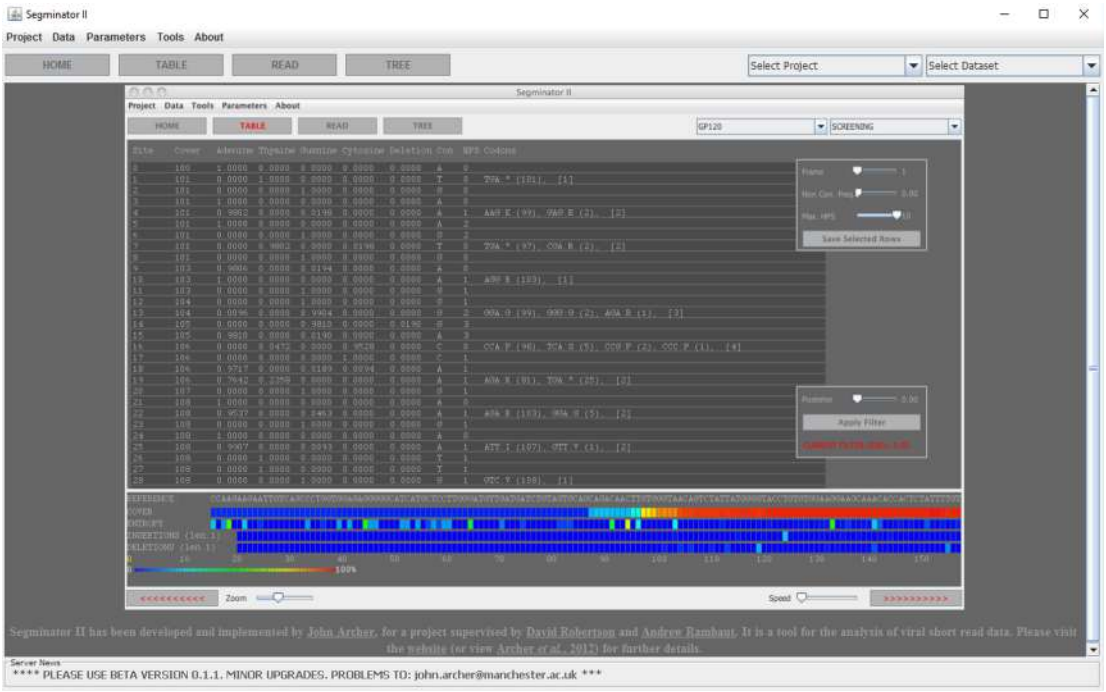


Figura 4.5: Interface gráfica do software Segminator II desenvolvido por Archer *et al.* (2012).

O Segminator II gera como saída um arquivo, denominado VEMETable, contendo dados referentes a cada posição da sequência de referência, como a cobertura (quantas vezes cada base do genoma foi coberta), a quantidade de cada base nitrogenada na mesma posição, a base presente na sequência consenso, entre outras variáveis.

Para a análise dos resultados do mapeamento contidos no arquivo VEMETable, apenas as informações referentes às posições das regiões que compreendem a PR (posições 2253 a 2549), a RT (posições 2550 a 4229) e a IN (posições 4230 a 5096) foram selecionadas. Foram desenvolvidos *scripts* para montar todos os possíveis códons - sequência de três bases nitrogenadas consecutivas (trinca) - a partir dos dados de cobertura e de bases nitrogenadas presentes em cada posição contidos no arquivo VEMETable. Os códons são interpretados e traduzidos, gerando todas as possibilidades de aminoácidos e suas respectivas frequências para cada posição da proteína. Apenas as posições de resistência e contendo aminoácidos com mais de 1% de frequência são selecionadas para formarem a tabela a ser exibida na tela do sistema. Esse valor de ponto de corte foi escolhido uma vez que novas tecnologias de NGS têm apresentado taxas de erros de sequenciamento que podem variar de 0,1 a 1%, dependendo do tipo de plataforma, do método de preparação da biblioteca e do tipo de sequenciamento [137].

As posições selecionadas para exibição no sistema foram as mesmas da lista de mutações do HIVdb, encontrada na página https://hivdb.stanford.edu/hivdb/by-

mutations/, com a inclusão de outras citadas na literatura [138, 139]. As posições escolhidas, acompanhadas do aminoácido da sequência original do HIV-1 subtipo B, foram:

- PR: L10, V11, I13, G16, K20, L23, L24, D30, V32, L33, E34, E35, M36, K43, M46, I47, G48, I50, F53, I54, Q58, D60, I62, L63, I64, H69, A71, G73, T74, L76, V77, V82, N83, I84, I85, N88, L89, L90 e I93;
- RT: E40, M41, E44, A62, K65, D67, T69, K70, L74, V75, F77, V90, A98, L100, K101, K103, V106, V108, Y115, F116, V118, E138, Q151, V179, Y181, M184, Y188, G190, L210, T215, K219, H221, P225, F227, M230, P236, K238, Y318 E N348;
- IN: H51, T66, L74, E92, Q95, T97, L101, H114, F121, A128, E138, G140, Y143, P145, Q146, S147, Q148, V151, S153, N155, E157, G163, G193, S230 e R263.

Adicionalmente, foi desenvolvido um *script* para a exibição de um gráfico contendo os valores de cobertura para cada posição da região analisada.

Todo o código para a análise das informações contidas no arquivo VEMETable foi desenvolvido no software R. Para a implementação do gráfico de cobertura foi utilizado o pacote *plotly* do R [140].

Para realizar a classificação do nível de suscetibilidade do HIV-1 aos ARVs, os algoritmos de interpretação ANRS v26, HIVdb v8.3, Rega v9.1.0 e Algoritmo Brasileiro v13, todos em suas versões atuais, foram implementados em linguagem R. As regras de classificação dos algoritmos foram obtidas nos sítios: http://www.hivfrenchresistance.org/2016/Algo2016-hiv1.pdf (ANRS), https://hivdb.stanford.edu/dr-summary/mut-scores/PI/ (HIVdb) e http://rega.kuleuven.be/cev/avd/files/software/rega_algorithm/Rega_HIV1_Rules_v9.1.0.pdf (Rega).

4.2.2 Interface Gráfica do SIRA-HIV

A interface gráfica do sistema foi implementada por meio do software R, utilizando o pacote *Shiny* [141]. O *Shiny* foi projetado com o objetivo de transformar as análises desenvolvidas no ambiente R em aplicativos interativos, não exigindo por parte do usuário final conhecimentos de programação. Detalhes e exemplos desse pacote podem ser encontrados em http://shiny.rstudio.com/.

Para desenvolver uma ferramenta por meio do Shiny, dois componentes precisam ser implementados, o *server*, local de construção dos objetos R a serem exibidos pelo sistema, e o *ui*, onde se ajustam os *layouts* de controle e aparência do aplicativo.

Com o objetivo de incrementar o *layout* da interface do SIRA-HIV foi utilizada a estrutura de código aberto *Bootstrap*, que permite maior personalização da aparência, com uma flexibilidade estética profissional. O modelo *Bootstrap* utilizado no presente trabalho foi obtido na página http://bootswatch.com/. Algumas alterações, como aumento da fonte e mudança de cores, foram realizadas no código-fonte.

O ambiente foi desenvolvido nas línguas portuguesa e inglesa, a fim de atingir um maior público-alvo.

4.2.3 Avaliação do SIRA-HIV

Neste estudo, a fim de validar o sistema desenvolvido, foram utilizados dez sequenciamentos gerados pela plataforma Ion Torrent PGM e cedidos pelo Laboratório de Virologia Molecular do Centro de Ciências da Saúde da Universidade Federal do Rio de Janeiro (CCS - UFRJ/Brasil), integrante da RENAGENO. Essas amostras foram analisadas previamente pelo programa DeepGen HIV, desenvolvido por Gibson e colaboradores, em 2014 [106]. O DeepGen HIV consiste em um teste de genotipagem do HIV altamente sensível baseado no sequenciamento de nova geração para determinar a suscetibilidade aos PIs, NRTIs, NNRTIs, INSTIs e inibidores de maturação, assim como o tropismo do vírus.

As mutações encontradas pelo SIRA-HIV foram comparadas àquelas definidas pelo DeepGen HIV. As mesmas amostras foram analisadas por ambos os sistemas e apenas as mutações com frequência maior ou igual a 1% foram consideradas na comparação.

Tanto o DeepGen HIV quanto o SIRA-HIV utilizaram o mesmo software de mapeamento das sequências (Segminator II) e o mesmo tamanho da sequência de referência, compreendendo as posições 1807 à 5096 do genoma do vírus.

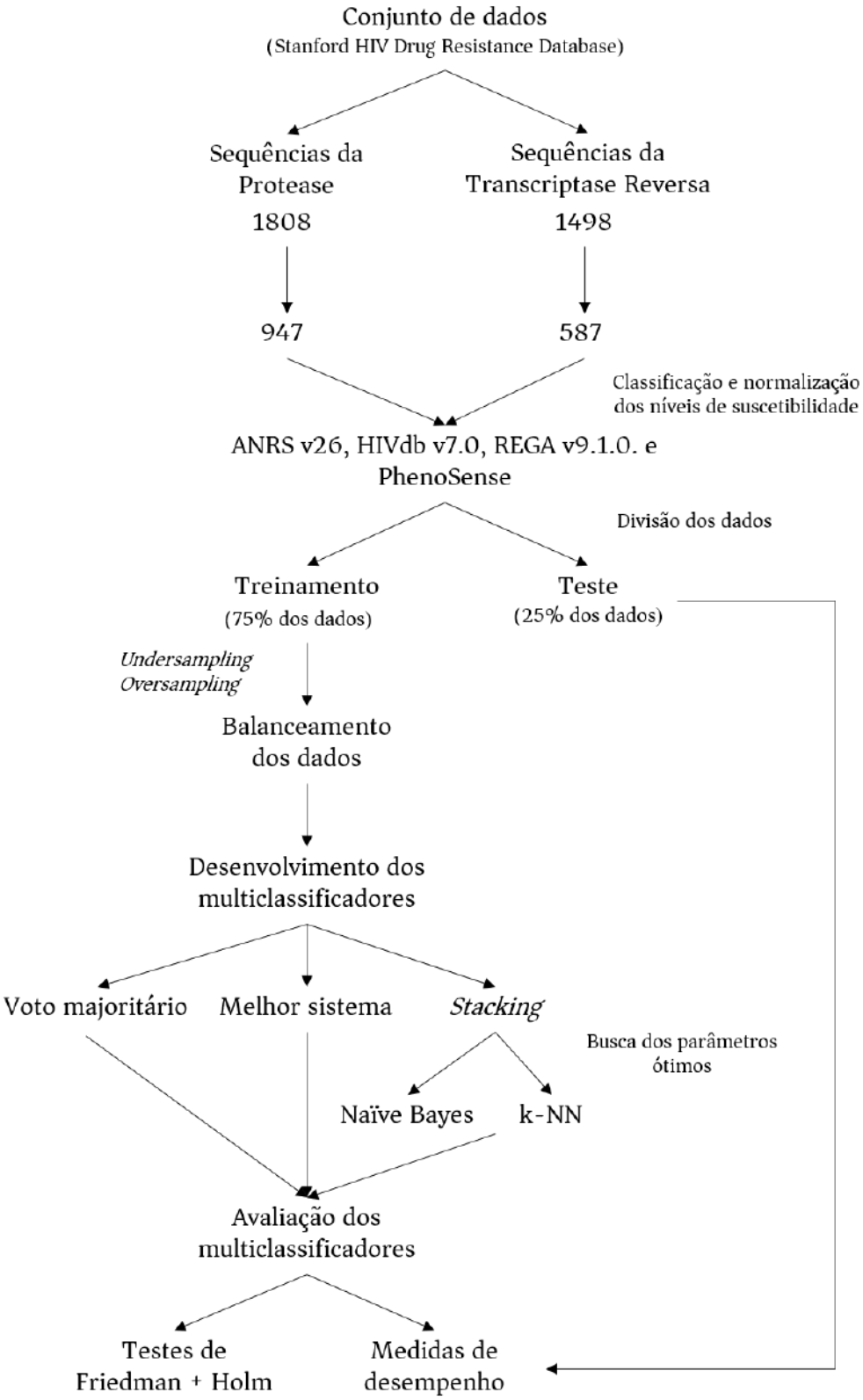


Figura 4.4: Etapas realizadas no desenvolvimento e avaliação dos multiclassificadores.

Capítulo 5

Resultados

5.1 Multiclassificadores

Neste trabalho, a fim de obter um classificador único de resistência a partir dos sistemas de interpretação genotípica, três estratégias foram aplicadas: voto majoritário (VM), melhor sistema (MS) e uma adaptação da técnica *stacking*. Esta última utilizou como metaclassificadores os algoritmos *naïve Bayes* (NB) e k-NN. Os hiperparâmetros definidos pela função *train* do pacote *caret* para cada estratégia estão apresentados no Anexo 1.

Para constituir a estratégia MS, o algoritmo com maior medida-F foi escolhido para cada ARV. A Tabela 5.1 mostra os valores das medidas-F para todos os ARVs incorporados neste estudo, segundo os três algoritmos de interpretação genotípica. Para todos os PIs, o Rega foi identificado como o melhor algoritmo. Para os NRTIs, o Rega também apresentou os melhores valores das medidas-F para os fármacos ABC, 3TC e AZT, enquanto que o HIVdb foi superior para o ddI e o ANRS para o d4T. Para os NNRTIs, o Rega e o HIVdb foram caracterizados como melhor sistema uma vez cada.

A comparação do desempenho dos multiclassificadores foi realizada por meio do teste de Friedman. Conforme mostrado na Tabela 5.2, para a maioria dos PIs, o algoritmo NB apresentou os melhores ranqueamentos médios. A abordagem MS foi superior para os fármacos FPV e NFV, enquanto que o k-NN, junto com o NB, apresentou melhor valor para o IDV. Para os NRTIs, o MS apresentou ranqueamento médio superior para os ARVs ddI, 3TC e AZT, porém, foi semelhante ao k-NN para o ddI e aos modelos k-NN e VM para o 3TC. O NB apresentou melhores colocações médias para o ABC e o d4T. Para a classe dos NNRTIs, o EFV apresentou ranqueamento similar para o NB, k-NN e VM, e o MS foi o mais bem colocado para o fármaco NVP. As melhores colocações foram realçadas em negrito.

A Tabela 5.3 apresenta os valores-p encontrados no teste de Friedman. Apenas

Tabela 5.1: Medidas-F para os algoritmos de interpretação genotípica. O algoritmo com maior valor foi selecionado para compor o melhor sistema (Os melhores valores da medida-F para cada antirretroviral estão em negrito).

Classe	Antirretroviral	ANRS (%)	HIVdb (%)	Rega (%)
PI	Atazanavir (ATV)	86,36	86,38	**88,73**
	Fosamprenavir (FPV)	81,35	83,23	**89,72**
	Indinavir (IDV)	84,24	82,79	**85,39**
	Lopinavir (LPV)	55,10	57,30	**74,23**
	Nelfinavir (NFV)	93,68	94,36	**95,10**
	Saquinavir (SQV)	55,50	70,48	**71,61**
NRTI	Abacavir (ABC)	49,30	49,45	**57,66**
	Didanosina (ddI)	49,48	**79,45**	76,47
	Estavudina (d4T)	**79,81**	71,20	76,92
	Lamivudina (3TC)	89,74	91,19	**92,05**
	Zidovudina (AZT)	87,13	88,58	**89,46**
NNRTI	Efavirenz (EFV)	90,37	92,26	**92,93**
	Nevirapina (NVP)	84,24	**96,08**	95,30

Tabela 5.2: Colocação média dos classificadores segundo os antirretrovirais. Em cada uma das 10 rodadas, o classificador com melhor medida-F recebia a colocação 1, o segundo melhor, a colocação 2 e assim sucessivamente. Após 10 rodadas, foram calculadas as colocações médias. Os valores em negrito correspondem aos melhores ranqueamentos para cada antirretroviral.

Classe	Antirretroviral	VM	MS	NB	k-NN
PI	Atazanavir (ATV)	2,30	3,25	**2,15**	2,30
	Fosamprenavir (FPV)	2,65	**1,90**	2,65	2,80
	Indinavir (IDV)	3,00	2,80	**2,10**	**2,10**
	Lopinavir (LPV)	4,00	2,00	**1,95**	2,05
	Nelfinavir (NFV)	2,65	**2,05**	2,65	2,65
	Saquinavir (SQV)	3,40	2,85	**1,55**	2,20
NRTI	Abacavir (ABC)	3,25	2,55	**1,35**	2,85
	Didanosina (ddI)	3,20	**1,80**	3,20	**1,80**
	Estavudina (d4T)	3,10	2,15	**1,65**	3,10
	Lamivudina (3TC)	**2,45**	**2,45**	2,65	**2,45**
	Zidovudina (AZT)	2,55	**2,35**	2,55	2,55
NNRTI	Efavirenz (EFV)	**2,40**	2,80	**2,40**	**2,40**
	Nevirapina (NVP)	2,80	**1,95**	2,80	2,45

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*

cinco ARVs (LPV, SQV, ABC, ddI e d4T) apresentaram valor-p abaixo do nível de significância escolhido ($\alpha = 5\%$), indicando que pelo menos dois dos classificadores possuem desempenhos estatisticamente diferentes um do outro. Para esses ARVs, foi realizado o teste de Holm para identificar quais multiclassificadores apresentavam diferenças de desempenho em relação ao classificador-controle.

Tabela 5.3: Valores-p do teste de Friedman usado na comparação do desempenho dos multiclassificadores. Valores menores que o nível de significância de 5% (marcados em negrito) indicam haver diferença de desempenho entre pelo menos dois dos classificadores comparados. Os resultados foram arredondados em quatro casas decimais.

Classe	Antirretroviral	Valor-p
PI	Atazanavir (ATV)	0,2044
	Fosamprenavir (FPV)	0,3963
	Indinavir (IDV)	0,2658
	Lopinavir (LPV)	**0,0004**
	Nelfinavir (NFV)	0,6549
	Saquinavir (SQV)	**0,0091**
NRTI	Abacavir (ABC)	**0,0072**
	Didanosina (ddI)	**0,0083**
	Estavudina (d4T)	**0,0245**
	Lamivudina (3TC)	0,9808
	Zidovudina (AZT)	0,9808
NNRTI	Efavirenz (EFV)	0,8685
	Nevirapina (NVP)	0,4057

Na Tabela 5.4 encontram-se os valores-p das comparações pareadas obtidos pelo teste de Holm. Para os ARVs LPV, SQV, ABC e d4T, o classificador-controle foi o NB e para o ddI o controle foi o algoritmo k-NN (o MS também poderia ter sido escolhido, visto que apresentou o mesmo ranqueamento médio). Em todos os casos analisados, o classificador-controle diferiu da estratégia VM. O classificador-controle NB teve desempenho superior em relação à abordagem MS para o fármaco SQV e em relação ao k-NN para o d4T. No caso do ABC, o NB se mostrou estatisticamente diferente nas três comparações. Para o ddI, além do VM, o k-NN também diferiu estatisticamente em relação ao NB.

Tabela 5.4: Valores-p das comparações pareadas dos diferentes classificadores. Valores abaixo de $0,05$ são rejeitados e foram marcados em negrito. Os resultados foram arredondados em quatro casas decimais.

Comparação	LPV	SQV	ABC	d4T	Comparação	ddI
VM vs. NB	**0,0012**	**0,0041**	**0,0030**	**0,0361**	VM vs. k-NN	**0,0459**
MS vs. NB	1,0000	**0,0487**	**0,0376**	0,3865	MS vs. k-NN	1,0000
k-NN vs. NB	1,0000	0,2602	**0,0187**	**0,0361**	NB vs. k-NN	**0,0459**

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*

Nas subseções seguintes serão apresentados os desempenhos dos multiclassificadores para cada ARV em termos de medida-F, acurácia, sensibilidade, especificidade, VPP e VPN.

5.1.1 Inibidores da Protease

5.1.1.1 Atazanavir (ATV)

Na Tabela 5.5 e na Figura 5.1 encontram-se os valores de desempenho dos multiclassificadores para o ATV. As abordagens VM, NB e k-NN apresentaram desempenhos similares paras as medidas avaliadas. O MS, por sua vez, apresentou os maiores valores de medida-F (86,95%), acurácia (88,81%), especificidade (83,13%), VPP (78,12%) e VPN (98,57%).

Tabela 5.5: Desempenho dos multiclassificadores para o antirretroviral atazanavir.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	80,64	86,95	81,30	80,99
Acurácia	82,09	88,81	82,84	82,84
Sensibilidade	98,04	98,04	98,04	96,08
Especificidade	72,29	83,13	73,49	74,70
Valor preditivo positivo	68,49	78,12	69,44	70,00
Valor preditivo negativo	98,36	98,57	98,39	96,88

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*

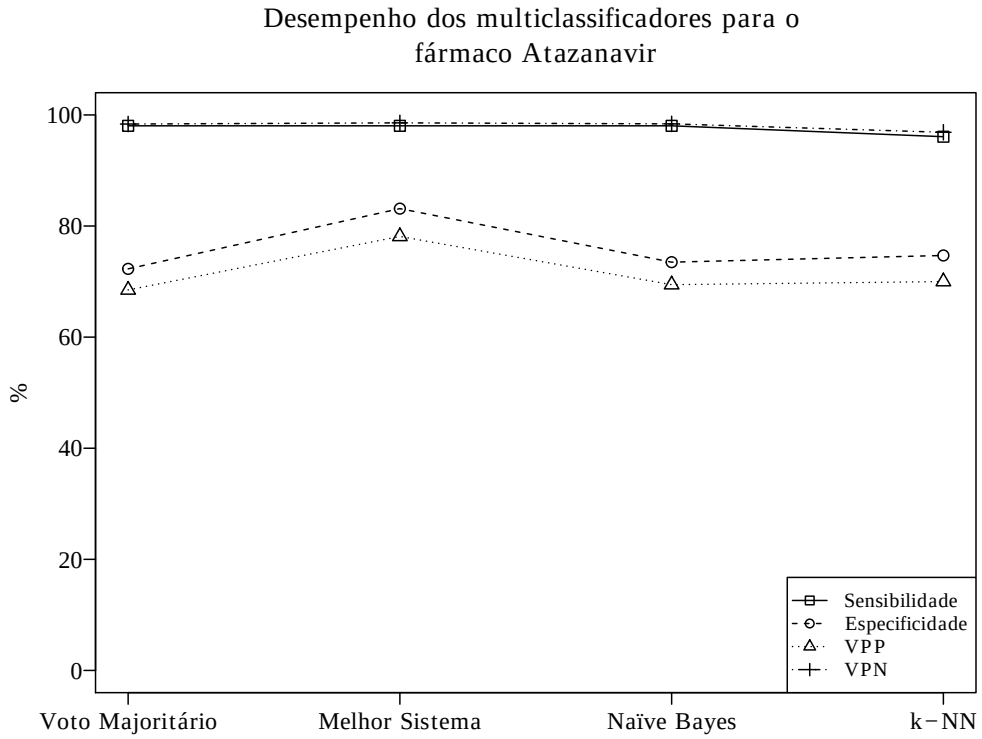


Figura 5.1: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados ao atazanavir.

5.1.1.2 Fosamprenavir (FPV)

Para o fármaco FPV, os multiclassificadores VM, NB e k-NN apresentaram resultados idênticos para todas as medidas de desempenho. O MS apresentou resultados próximos aos dos outros multiclassificadores, com valores não ultrapassando uma diferença de 2% (Tabela 5.6 e Figura 5.2). A medida-F foi aproximadamente a mesma para todas as estratégias.

Tabela 5.6: Desempenho dos multiclassificadores para o antirretroviral fosamprenavir.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	88,31	88,74	88,31	88,31
Acurácia	91,67	92,13	91,67	91,67
Sensibilidade	95,77	94,37	95,77	95,77
Especificidade	89,66	91,03	89,66	89,66
Valor preditivo positivo	81,93	83,75	81,93	81,93
Valor preditivo negativo	97,74	97,06	97,74	97,74

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*

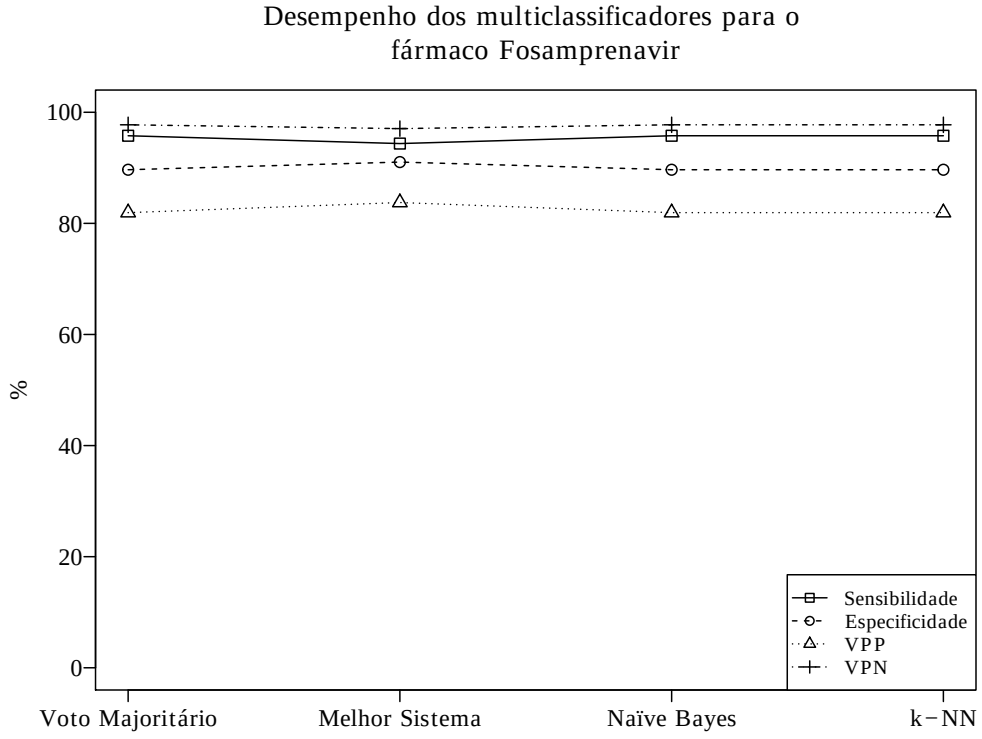


Figura 5.2: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados ao fosamprenavir.

5.1.1.3 Indinavir (IDV)

A Tabela 5.7 e a Figura 5.3 apresentam o desempenho das estratégias de multiclassificação para o IDV. Os classificadores MS e NB apresentaram os melhores desempenhos, com medidas-F aproximadamente iguais. O k-NN errou todos os casos de resistência (sensibilidade igual a 0%), porém acertou todos os casos classificados como suscetível (especificidade igual a 100%). Os demais multiclassificadores apresentaram sensibilidade e VPN iguais a 100%.

Tabela 5.7: Desempenho das abordagens de multiclassificação para o antirretroviral indinavir.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	74,58	77,19	77,65	NaN
Acurácia	79,91	82,59	83,04	70,54
Sensibilidade	100	100	100	0,00
Especificidade	71,52	75,32	75,95	100
Valor preditivo positivo	59,46	62,86	63,46	NaN
Valor preditivo negativo	100	100	100	70,54

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*; NaN: *not a number*

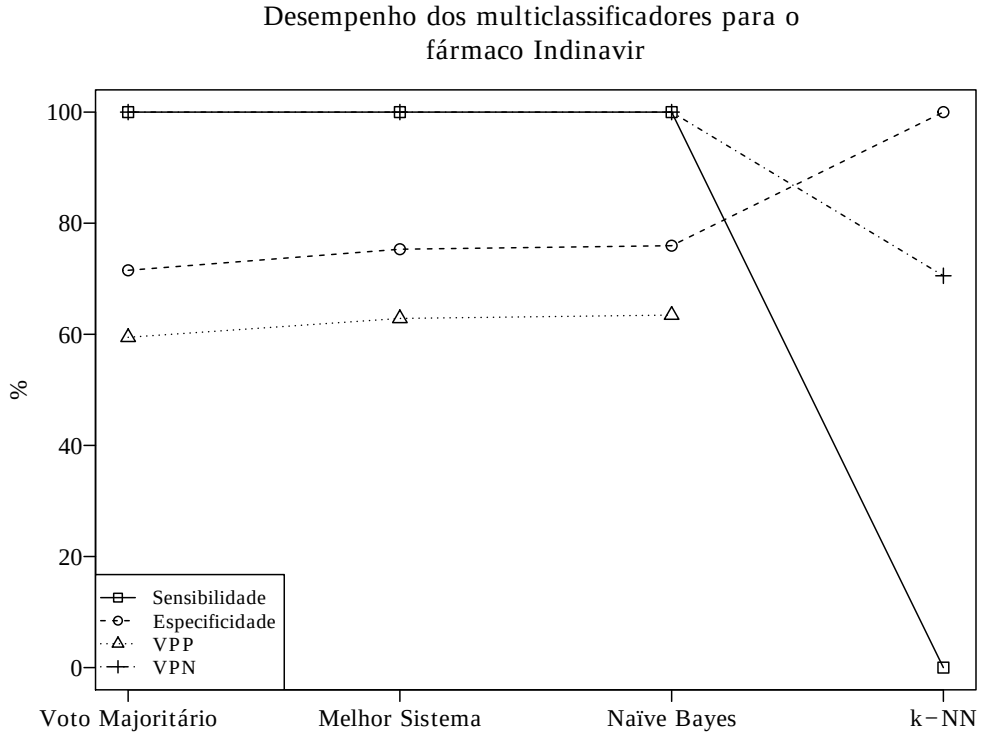


Figura 5.3: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados ao indinavir. O k-NN não apresenta VPP por ter apresentado uma sensibilidade igual a zero.

5.1.1.4 Lopinavir (LPV)

O multiclassificador NB apresentou os melhores resultados de desempenho para o fármaco LPV, com medida-F igual a 72,42%, acurácia igual a 81,71%, sensibilidade média igual a 76,28%, especificidade média de 92,51%, VPP e VPN médios de 72,06% e 89,98%, respectivamente. O VM apresentou valores reduzidos de sensibilidade (2,94%) e VPP (4,76%) para a categorização resistência intermediária vs. não intermediária (I vs. NI). O k-NN, por sua vez, teve um desempenho ruim para a sensibilidade (3,23%) na categorização resistência vs. não resistência (R vs. NR). Na Tabela 5.8 e na Figura 5.4 estão indicados os desempenhos dos multiclassificadores para o LPV.

Tabela 5.8: Desempenho dos multiclassificadores para o antirretroviral lopinavir. Por apresentar três níveis de suscetibilidade, os dados foram categorizados binariamente para o cálculo das medidas de desempenho.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	50,90	74,23	72,42	51,93
Acurácia	66,86	80,00	81,71	75,43
Sensibilidade				
S vs. NS	77,27	88,18	89,09	88,18
I vs. NI	2,94	50,00	55,88	100
R vs. NR	100	83,87	83,87	3,23
Especificidade				
S vs. NS	100	100	100	100
I vs. NI	85,81	87,94	88,65	68,50
R vs. NR	73,61	87,50	88,89	100
Valor preditivo positivo				
S vs. NS	100	100	100	100
I vs. NI	4,76	50,00	54,29	44,16
R vs. NR	44,93	59,09	61,90	100
Valor preditivo negativo				
S vs. NS	72,22	83,33	84,42	83,33
I vs. NI	78,57	87,94	89,29	100
R vs. NR	100	96,18	96,24	82,76

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*; S: suscetível; NS: não suscetível; I: resistência intermediária; NI: não resistência intermediária; R: resistência; NR: não resistência; NaN: *not a number*

5.1.1.5 Nelfinavir (NFV)

Na Tabela 5.9 e na Figura 5.5 pode-se observar o desempenho dos multiclassificadores para o NFV. As estratégias apresentaram resultados próximos para todas as medidas avaliadas.

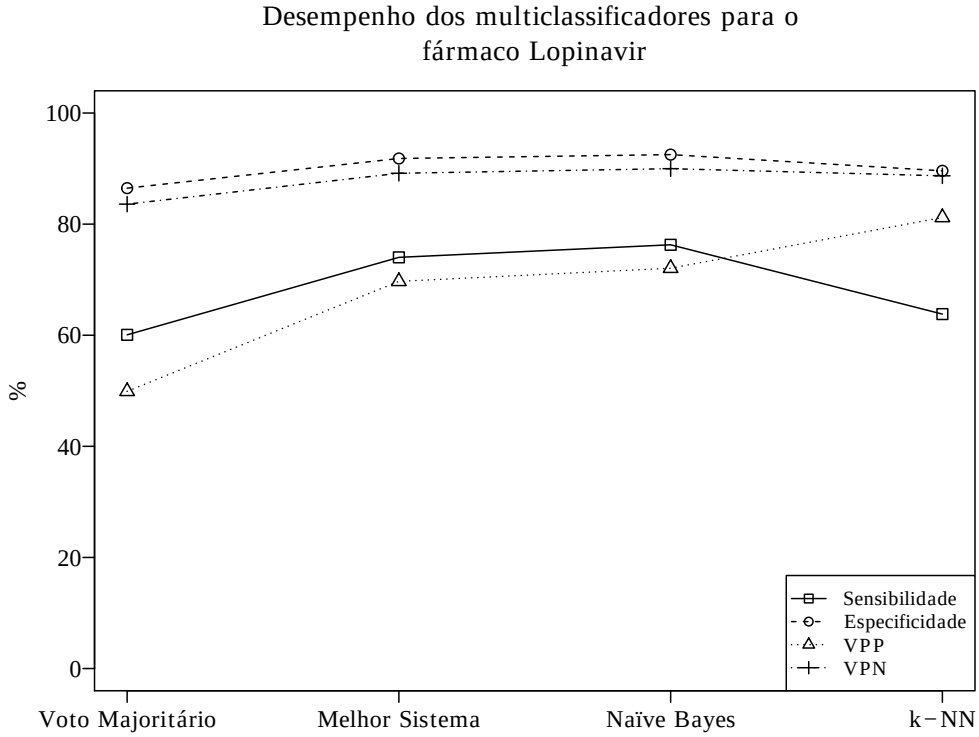


Figura 5.4: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados ao lopinavir.

Tabela 5.9: Desempenho das abordagens de multiclassificação para o antirretroviral nelfinavir.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	92,49	92,86	92,49	92,49
Acurácia	91,77	92,21	91,77	91,77
Sensibilidade	97,50	97,50	97,50	97,50
Especificidade	85,59	86,49	85,59	85,59
Valor preditivo positivo	87,97	88,64	87,97	87,97
Valor preditivo negativo	96,94	96,97	96,94	96,94

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*

5.1.1.6 Saquinavir (SQV)

Os resultados dos multiclassificadores para o SQV encontram-se na Tabela 5.10. Em média, as abordagens apresentaram desempenhos semelhantes, com variações de valores em algumas medidas, como pode ser observado na Figura 5.6. O VM e o MS apresentaram os melhores valores de sensibilidade (96,49%) e VPN (98,66%), para a categorização R vs. NR, e de especificidade (96,63%) e VPP (97,44% e 97,50%), para a categorização suscetível vs. não suscetível (S vs. NS). O NB apresentou melhor valor de sensibilidade (53,12%) e VPN (91,89%) para a categorização I vs. NI. O k-NN, por sua vez, se destacou na acurácia (83,93%), na sensibilidade (89,63%)

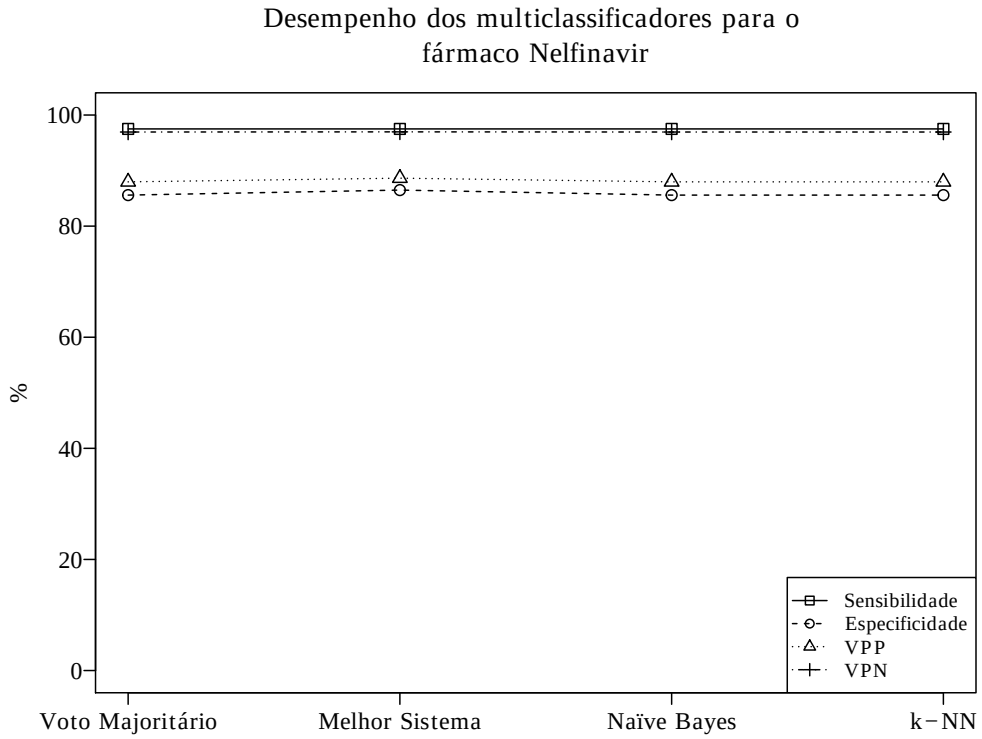


Figura 5.5: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados ao nelfinavir.

e no VPN (85,42%), para a categorização S vs. NS, e no VPP (79,10%) para a categorização R vs. NR. Os metaclassificadores NB e k-NN apresentaram os melhores valores de medida-F.

5.1.2 Inibidores Análogos de Nucleosídeos da Transcriptase Reversa

5.1.2.1 Abacavir (ABC)

Para o ABC, o NB apresentou melhor desempenho médio que os outros multiclassificadores, com medida-F igual a 61,51%, acurácia de 76,15%, sensibilidade média igual a 65,17%, especificidade média de 88,82%, VPP e VPN médios de 61,26% e 86,48%, respectivamente (Figura 5.7). A sensibilidade e o VPP para todas as abordagens apresentaram valores reduzidos para a categorização I vs. NI. Na Tabela 5.11 encontram-se os valores das medidas de desempenho das quatro estratégias utilizadas no estudo.

5.1.2.2 Didanosina (ddI)

Para este ARV, o VM e NB apresentaram o mesmo desempenho e o MS e o k-NN também foram semelhantes entre si. A primeira dupla de multiclassificadores

Tabela 5.10: Desempenho dos multiclassificadores para o antirretroviral saquinavir.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	70,77	71,83	73,50	73,63
Acurácia	81,25	82,59	82,14	83,93
Sensibilidade				
S vs. NS	84,44	86,67	85,19	89,63
I vs. NI	40,62	40,62	53,12	43,75
R vs. NR	96,49	96,49	91,23	92,98
Especificidade				
S vs. NS	96,63	96,63	95,51	92,13
I vs. NI	90,10	91,67	88,54	92,19
R vs. NR	88,02	88,02	91,62	91,62
Valor preditivo positivo				
S vs. NS	97,44	97,50	96,64	94,53
I vs. NI	40,62	44,83	43,59	48,28
R vs. NR	73,33	73,33	78,79	79,10
Valor preditivo negativo				
S vs. NS	80,37	82,69	80,95	85,42
I vs. NI	90,10	90,26	91,89	90,77
R vs. NR	98,66	98,66	96,84	97,45

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*; S: suscetível; NS: não suscetível; I: resistência intermediária; NI: não resistência intermediária; R: resistente; NR: não resistente

apresentou melhores valores de especificidade (81,82%) e VPP (87,30%), enquanto que os últimos foram superiores na sensibilidade (88,06%) e VPN (79,49%). A acurácia foi praticamente a mesma nas quatro abordagens. O desempenho dos multiclassificadores está retratado na Tabela 5.12 e Figura 5.8.

5.1.2.3 Estavudina (d4T)

Na Tabela 5.13 e Figura 5.9 encontram-se os valores de desempenho para o d4T. O NB apresentou os melhores valores de medida-F (79,57%), acurácia (82,88%) e VPP (69,81%). O VM foi superior aos outros multiclassificadores na sensibilidade (95,00%) e no VPN (96,08%). O k-NN, por sua vez, se sobressaiu na especificidade (100%), porém sua sensibilidade foi igual a 0%, com nenhum caso de resistência identificado.

5.1.2.4 Lamivudina (3TC)

Para o 3TC, as quatro estratégias de multiclassificação apresentaram valores muito similares, com as técnicas VM, NB e k-NN com os mesmos valores de medida-F (89,92%), acurácia (88,29%), sensibilidade (82,86%), especificidade (97,56%),

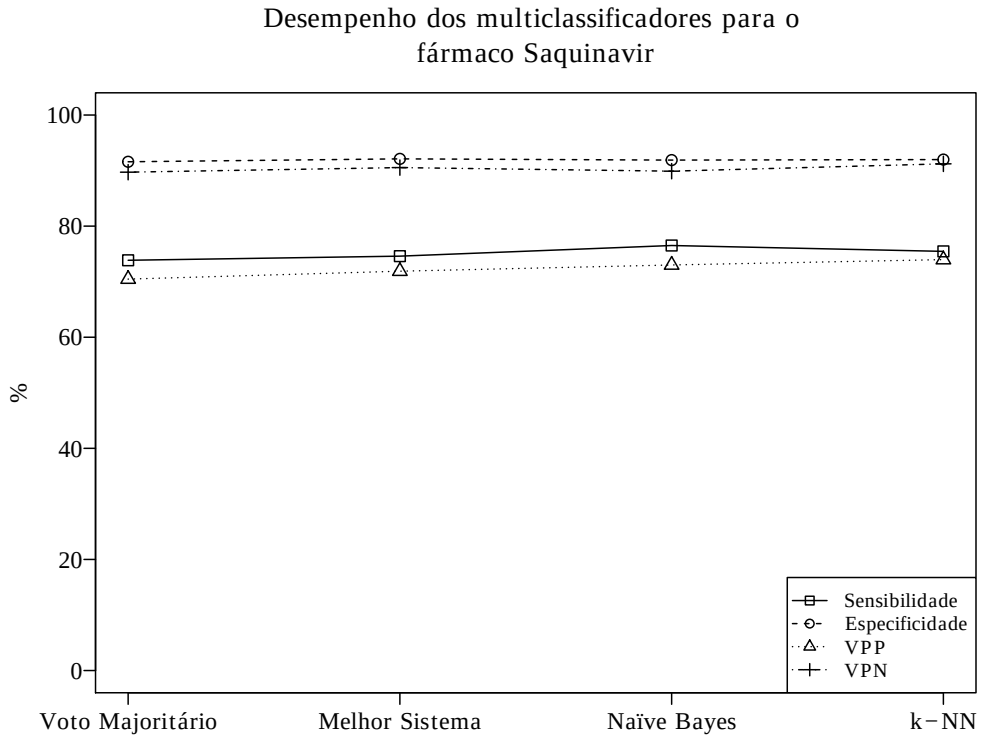


Figura 5.6: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados ao saquinavir.

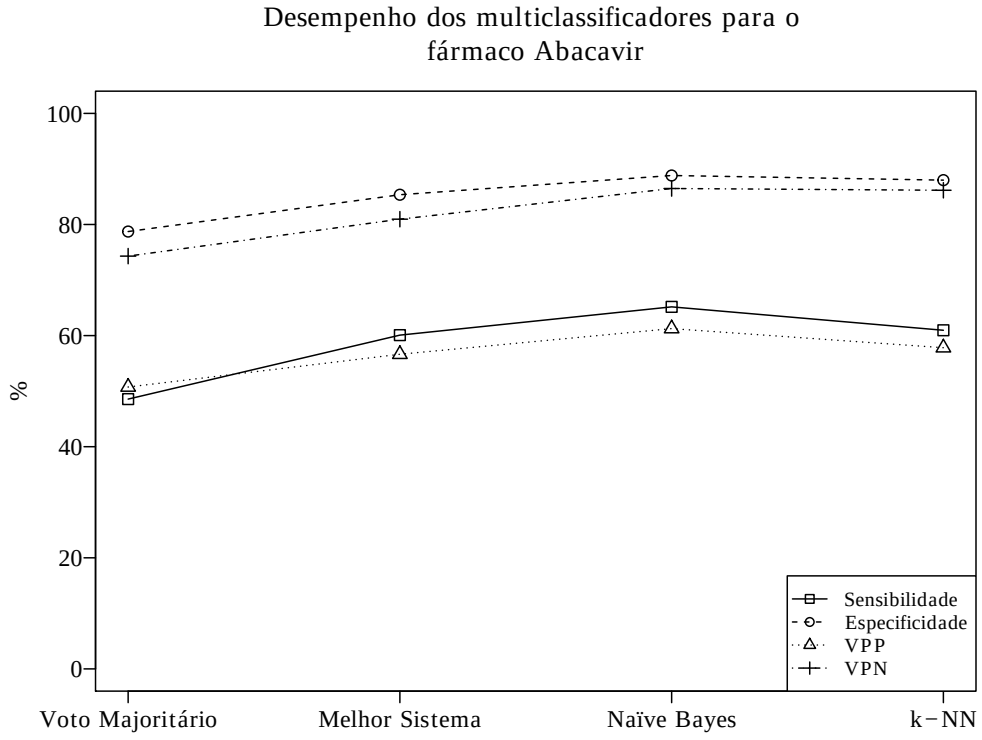


Figura 5.7: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados ao abacavir.

Tabela 5.11: Desempenho dos multiclassificadores para o antirretroviral abacavir.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	42,83	57,66	61,52	55,30
Acurácia	46,79	66,06	76,15	74,31
Sensibilidade				
S vs. NS	43,06	70,83	86,11	86,11
I vs. NI	17,65	29,41	29,41	11,76
R vs. NR	85,00	80,00	80,00	85,00
Especificidade				
S vs. NS	100	94,59	91,89	91,89
I vs. NI	59,78	76,09	88,04	95,65
R vs. NR	76,40	85,39	86,52	76,40
Valor preditivo positivo				
S vs. NS	100	96,23	95,38	95,38
I vs. NI	7,50	18,52	31,25	33,33
R vs. NR	44,74	55,17	57,14	44,74
Valor preditivo negativo				
S vs. NS	47,44	62,50	77,27	77,27
I vs. NI	79,71	85,37	87,10	85,44
R vs. NR	95,77	95,00	95,06	95,77

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*; S: suscetível; I: resistência intermediária; R: resistência

Tabela 5.12: Desempenho dos multiclassificadores para o antirretroviral didanosina.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	84,62	84,89	84,62	84,89
Acurácia	81,98	81,08	81,98	81,08
Sensibilidade	82,09	88,06	82,09	88,06
Especificidade	81,82	70,45	81,82	70,45
Valor preditivo positivo	87,30	81,94	87,30	81,94
Valor preditivo negativo	75,00	79,49	75,00	79,49

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*

VPP (98,31%) e VPN (76,92%). O MS diferiu na medida-F (89,06%), na acurácia (87,39%), na sensibilidade (81,43%) e no VPN (75,47%) (Tabela 5.14 e Figura 5.10).

5.1.2.5 Zidovudina (AZT)

Os multiclassificadores VM, NB e k-NN apresentaram resultados idênticos para o AZT. O MS, por sua vez, apresentou resultados próximos aos das outras estratégias, com porcentagens levemente menores para os valores de medida-F (93,69%), acurácia (93,64%), sensibilidade (96,30%) e VPN (96,23%) (Tabela 5.15 e Figura 5.11).

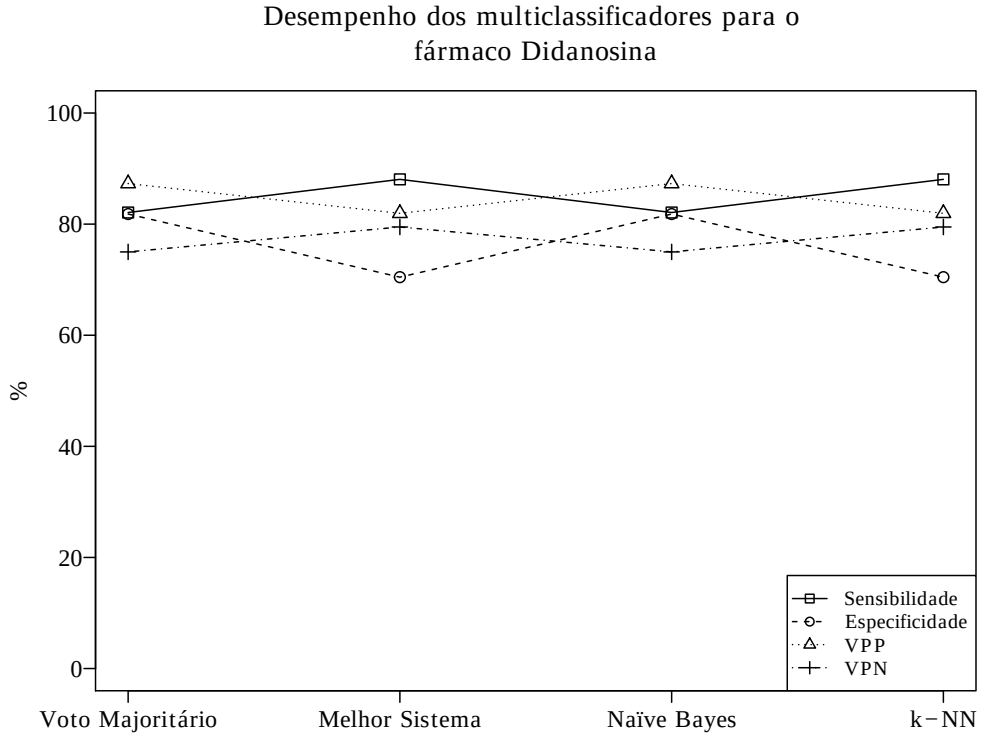


Figura 5.8: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados à didanosina.

Tabela 5.13: Desempenho dos multiclassificadores para o antirretroviral estavudina.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	76,00	76,29	79,57	NaN
Acurácia	78,38	79,28	82,88	63,96
Sensibilidade	95,00	92,50	92,50	0,00
Especificidade	69,01	71,83	77,46	100
Valor preditivo positivo	63,33	64,91	69,81	NaN
Valor preditivo negativo	96,08	94,44	94,83	63,96

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors;* NaN: *not a number*

5.1.3 Inibidores Não Análogos de Nucleosídeos da Transcriptase Reversa

5.1.3.1 Efavirenz (EFV)

Para o conjunto de teste, o EFV apresentou multiclassificadores com desempenhos muito próximos, com as técnicas VM, NB e k-NN apresentando os mesmos valores para as medidas de desempenho calculadas. O MS apresentou desempenho superior aos demais multiclassificadores, com medida-F igual a 86,44%, acurácia de 86,99%, sensibilidade igual a 96,23%, especificidade de 80,00% e VPP e VPN iguais a 78,46%

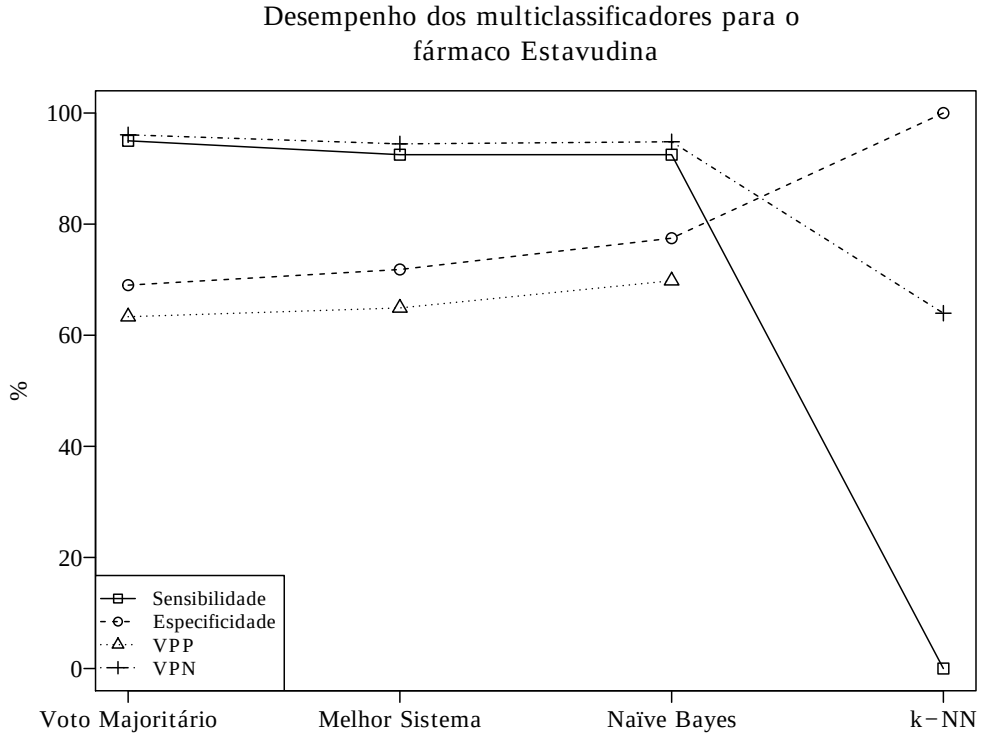


Figura 5.9: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados à estavudina.

Tabela 5.14: Desempenho dos multiclassificadores para o antirretroviral lamivudina.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	89,92	89,06	89,92	89,92
Acurácia	88,29	87,39	88,29	88,29
Sensibilidade	82,86	81,43	82,86	82,86
Especificidade	97,56	97,56	97,56	97,56
Valor preditivo positivo	98,31	98,28	98,31	98,31
Valor preditivo negativo	76,92	75,47	76,92	76,92

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*

e 96,55%, respectivamente. A Tabela 5.16 e a Figura 5.12 mostram os resultados encontrados para o EFV.

5.1.3.2 Nevirapina (NVP)

Na Tabela 5.17 e Figura 5.13 estão representados os desempenhos das abordagens de multiclassificação para a NVP. O MS obteve os melhores valores de desempenho, compartilhando com o k-NN os mesmos valores de sensibilidade (98,39%). Os demais multiclassificadores também apresentaram resultados satisfatórios.

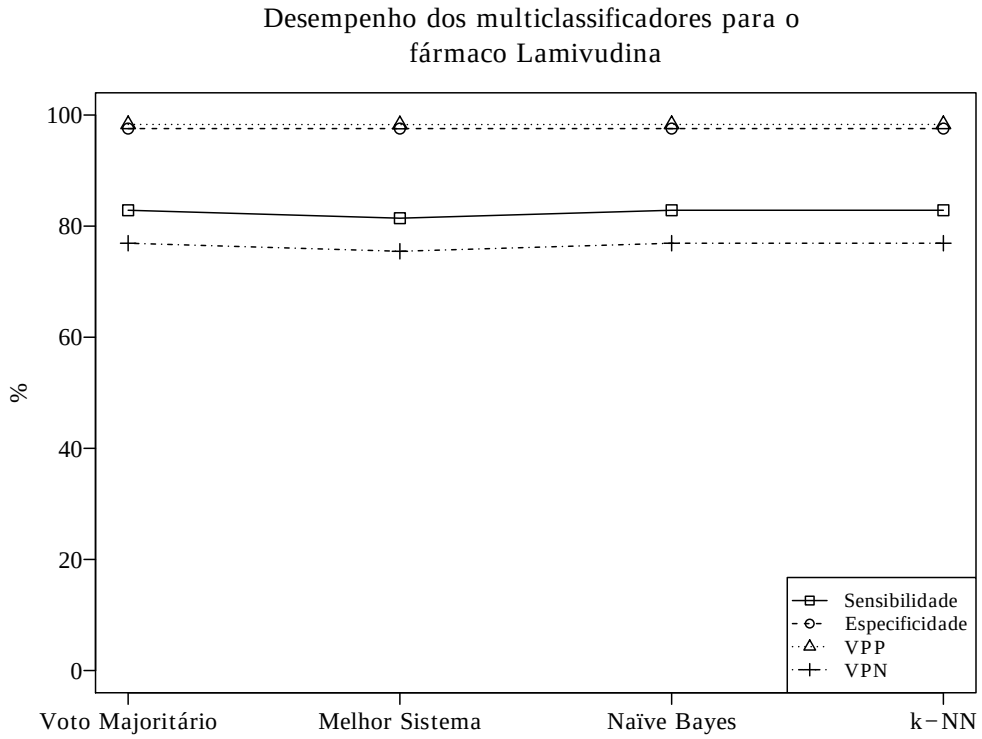


Figura 5.10: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados à lamivudina.

Tabela 5.15: Desempenho dos multiclassificadores para o antirretroviral zidovudina.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	94,74	93,69	94,74	94,74
Acurácia	94,55	93,64	94,55	94,55
Sensibilidade	100	96,30	100	100
Especificidade	89,29	91,07	89,29	89,29
Valor preditivo positivo	90,00	91,23	90,00	90,00
Valor preditivo negativo	100	96,23	100	100

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*

Tabela 5.16: Desempenho dos multiclassificadores para o antirretroviral efavirenz.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	85,00	86,44	85,00	85,00
Acurácia	85,37	86,99	85,37	85,37
Sensibilidade	96,23	96,23	96,23	96,23
Especificidade	77,14	80,00	77,14	77,14
Valor preditivo positivo	76,12	78,46	76,12	76,12
Valor preditivo negativo	96,43	96,55	96,43	96,43

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*

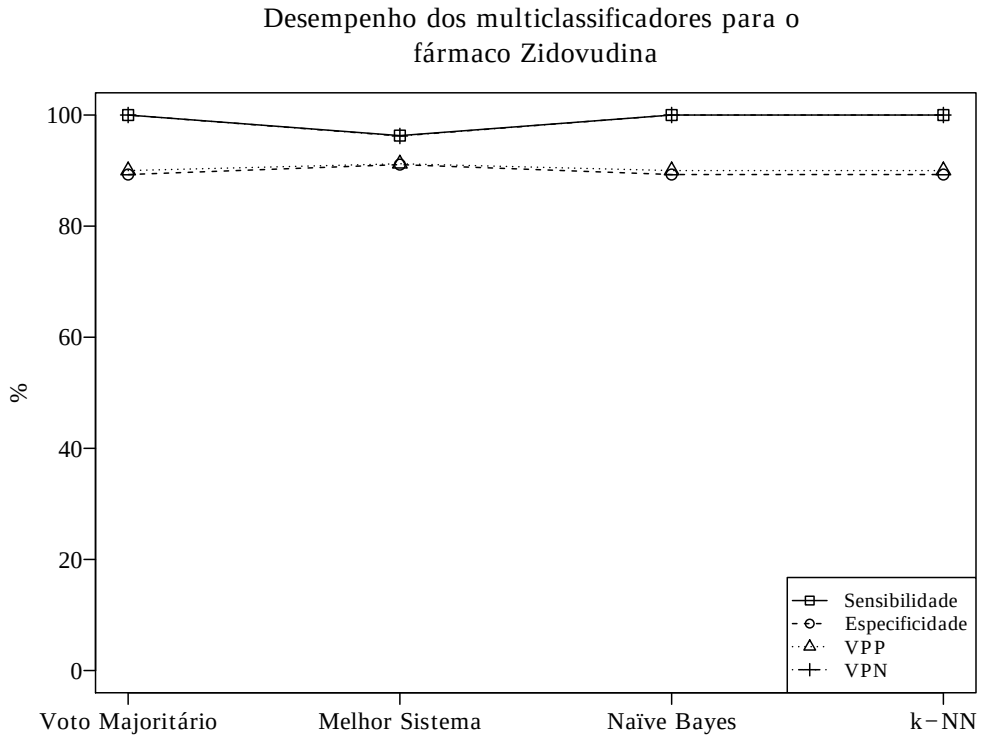


Figura 5.11: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados à zidovudina.

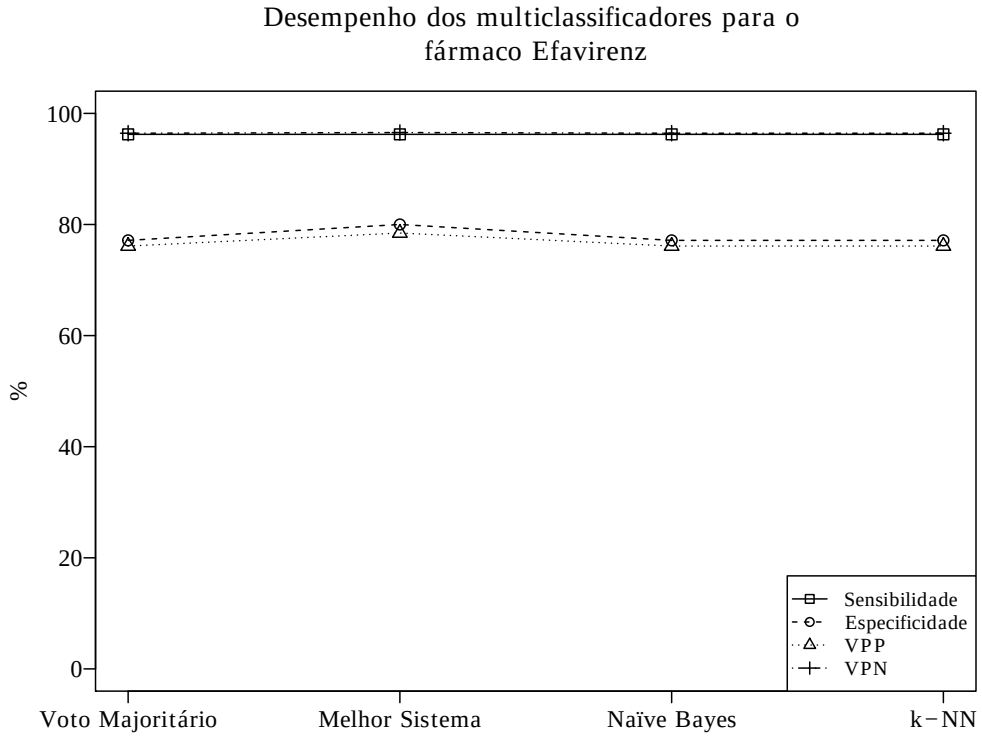


Figura 5.12: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados ao efavirenz.

Tabela 5.17: Desempenho das abordagens de multiclassificação para o antirretroviral nevirapina.

	VM (%)	MS (%)	NB (%)	k-NN (%)
Medida-F	91,47	93,85	91,47	91,73
Acurácia	91,20	93,60	91,20	91,20
Sensibilidade	95,16	98,39	95,16	98,39
Especificidade	87,30	88,89	87,30	84,13
Valor preditivo positivo	88,06	89,71	88,06	85,92
Valor preditivo negativo	94,83	98,25	94,83	98,15

VM: voto majoritário; MS: melhor sistema; NB: *naïve Bayes*; k-NN: *k-nearest neighbors*

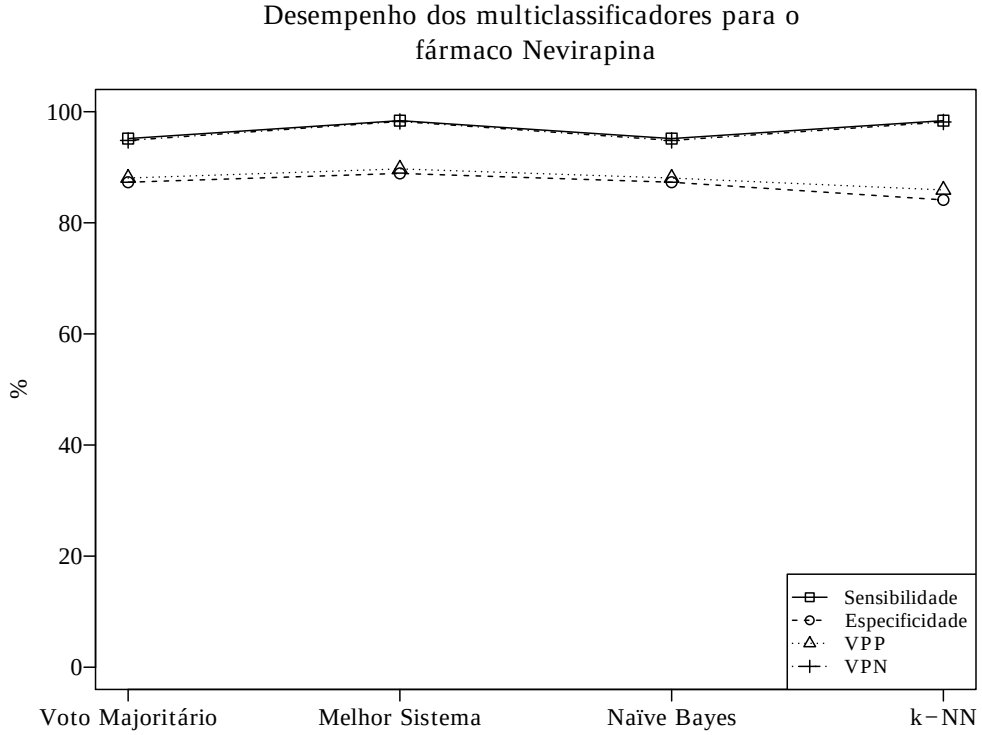


Figura 5.13: Sensibilidade, especificidade e valores preditivos positivo e negativo dos multiclassificadores aplicados à nevirapina.

5.2 Sistema de Identificação de Resistência aos Antirretrovirais - SIRA-HIV

Nesta seção são apresentados os módulos e funções do Sistema de Identificação de Resistência aos Antirretrovirais - SIRA-HIV. O SIRA-HIV é um sistema desenvolvido para análise de dados de sequenciamento de nova geração provenientes de amostras de indivíduos HIV+.

5.2.1 Módulos do Sistema

O SIRA-HIV é composto por dois módulos principais denominados Posições de Resistência e Algoritmos de Interpretação Genotípica de Resistência. O primeiro módulo é responsável pela análise do NGS e identificação das mutações presentes no genoma do HIV-1. O segundo módulo, por sua vez, realiza a classificação do nível de suscetibilidade do vírus aos ARVs por meio de algoritmos de interpretação genotípica. Toda a análise executada pelo SIRA-HIV dura em torno de 5-10 minutos. Maiores detalhes sobre cada um dos módulos serão descritos no decorrer desta subseção. No Anexo 2, encontra-se uma descrição detalhada do uso do SIRA-HIV.

Inicialmente, o ambiente possui uma interface com uma tela de Início onde estão contidas algumas informações referentes ao sistema. Nas duas telas seguintes, encontram-se os módulos e, na última tela, intitulada Sobre, estão as informações sobre a equipe desenvolvedora do projeto. A Figura 5.14 mostra as telas Início e Sobre do SIRA-HIV.

5.2.1.1 Posições de Resistência

Este módulo envolve o mapeamento das leituras geradas pelo NGS em relação a um genoma de referência do HIV-1, a análise dos resultados do mapeamento e a identificação dos aminoácidos presentes nas posições de resistência.

Neste módulo, a análise dos dados foi dividida em três etapas. Na etapa 1, o usuário inicializa o Segminator II e insere os dados de sequenciamento a serem caracterizados pelo programa. É necessário fornecer ao programa um arquivo contendo a sequência de referência no formato FASTA e um arquivo com os resultados do sequenciamento no formato FASTQ. Há ainda a opção de ajuste de parâmetros de mapeamento no programa, caso o usuário queira utilizar valores diferentes do padrão. O software gera automaticamente uma montagem por meio de um mapeamento inicial e alinha cada fragmento de DNA em relação à sequência de referência. Os resultados são exportados por meio do arquivo VEMETable gerado pelo programa. Caso o usuário já possua esses arquivos, a etapa 1 pode ser pulada.

Na etapa 2, o usuário fornece os resultados do mapeamento contidos no arquivo VEMETable ao SIRA-HIV. Esse arquivo pode ser o gerado na etapa 1 ou qualquer outro arquivo VEMETable gerado em análise anterior.

Na etapa 3, a região do gene *pol* a ser analisada é escolhida. No presente trabalho, apenas as análises das regiões correspondentes às enzimas PR, RT e IN foram incorporadas no sistema. Cada região é avaliada separadamente, de acordo com a opção selecionada. O SIRA-HIV retorna ao usuário uma tabela contendo as posições dos aminoácidos na enzima selecionada, o aminoácido original (à esquerda da posição), o aminoácido presente nas sequências (à direita da posição), as frequências de cada

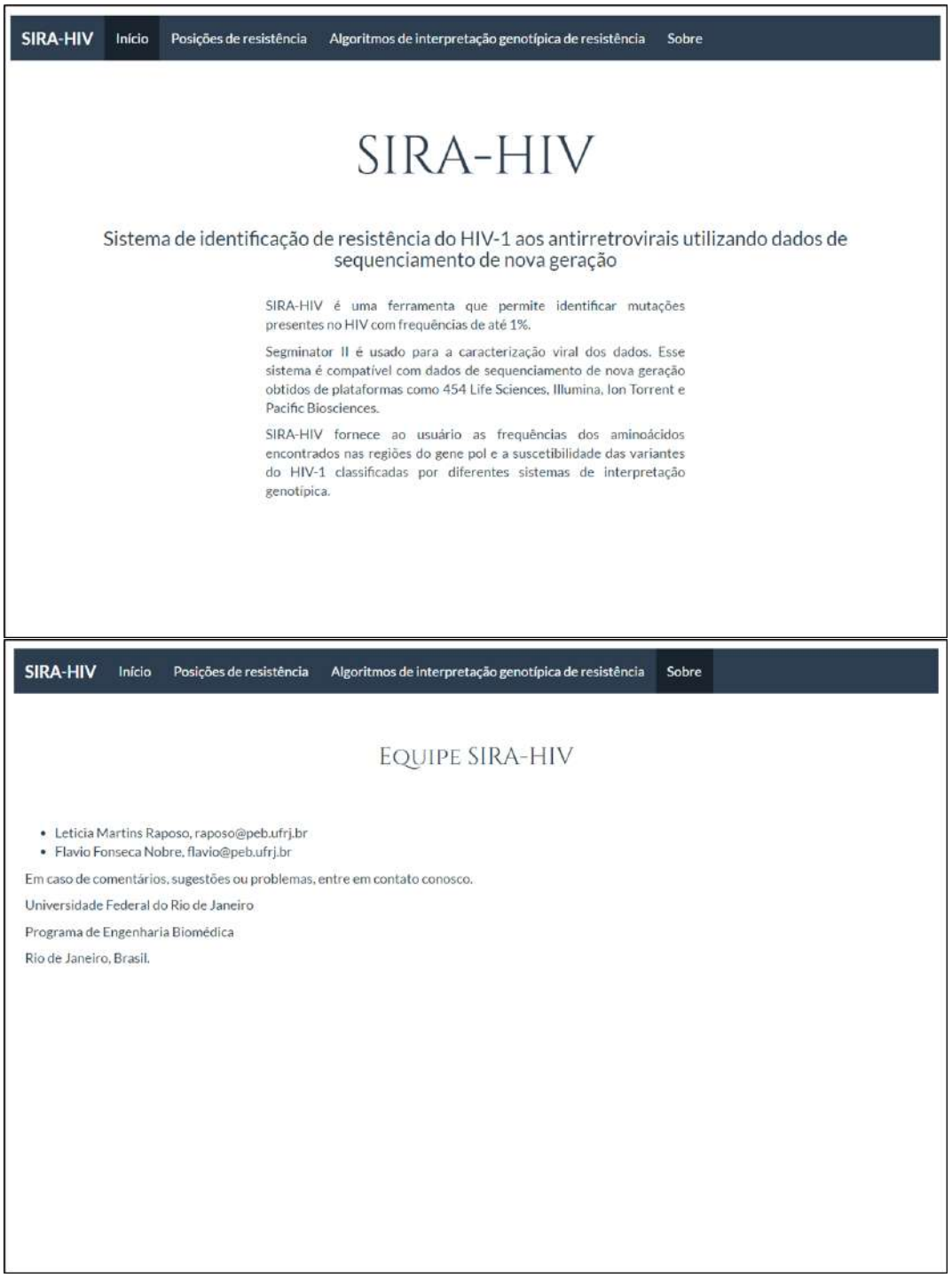


Figura 5.14: Telas inicial e final do ambiente integrado SIRA-HIV.

aminoácido e a cobertura de cada posição. Um gráfico de cobertura também foi adicionado ao sistema a fim de facilitar a visualização do número de aminoácidos por posição (Figura 5.15).

O usuário pode conferir um nome à sua análise, que é atribuído aos arquivos a serem salvos ao longo do processo. Tanto as informações contidas na tabela quanto o gráfico de cobertura podem ser salvos pelos usuários. Para a tabela, há três formatos

de saída (.xls, .csv e .pdf) e, para o gráfico, a saída é no formato .png.

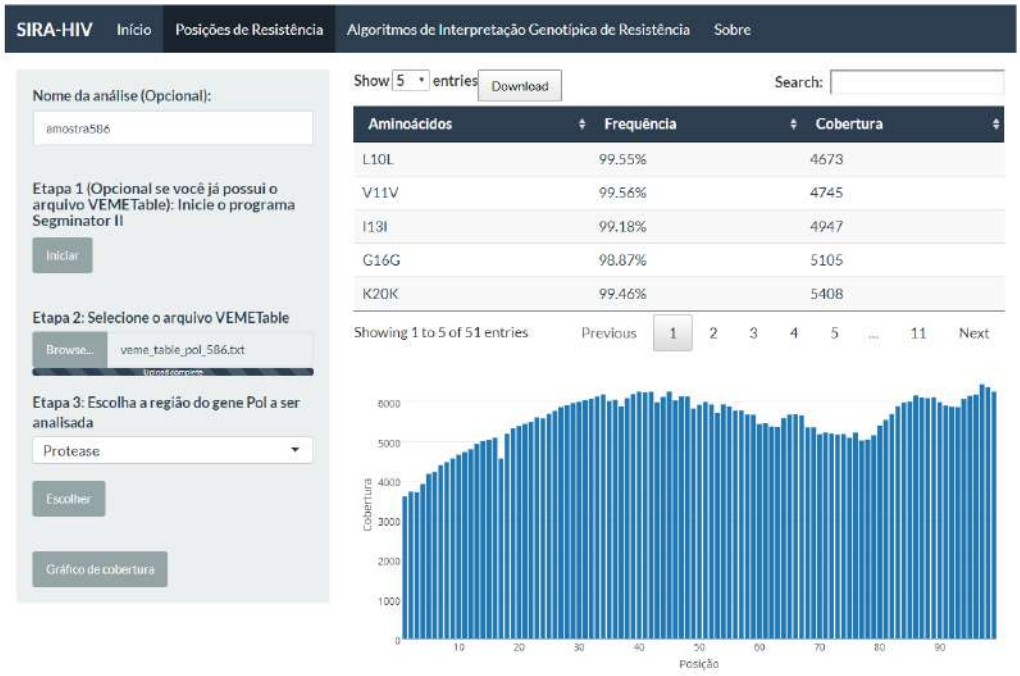


Figura 5.15: Resultado da análise de uma amostra de exemplo exibido pelo SIRA-HIV. A tabela apresenta os aminoácidos encontrados nas posições de resistência, suas frequências e a cobertura. No gráfico de cobertura, o usuário é capaz de visualizar rapidamente se alguma posição possui cobertura insuficiente.

5.2.1.2 Algoritmos de Interpretação Genotípica de Resistência

Este módulo fornece ao usuário o nível de suscetibilidade do HIV-1 aos ARVs atuantes sobre as enzimas PR, RT e IN. Ele ficou estruturado em duas etapas.

Na etapa 1, o usuário escolhe os algoritmos de interpretação genotípica. Quatro sistemas de classificação (ANRS, HIVdb, Rega e Algoritmo Brasileiro) foram implementados e podem ser selecionados individualmente ou em conjunto.

Na etapa 2, o nível de frequência dos aminoácidos deve ser escolhido. Dois pontos de corte foram definidos: $\geq 20\%$ e $\geq 1\%$. A categoria $\geq 20\%$ equivale ao ponto de corte do sequenciamento tradicional de Sanger e apenas os aminoácidos com esse limiar de frequência são incluídos na análise de resistência (Figura 5.16). Na categoria $\geq 1\%$, além das mutações majoritárias, as minoritárias também são levadas em consideração pelos algoritmos de interpretação genotípica (Figura 5.17).

A classificação de resistência aos ARVs está limitada à região do gene *pol* escolhida no primeiro módulo. Caso o usuário tenha selecionado a enzima RT, aparecerão apenas os níveis de suscetibilidade aos NRTIs e NNRTIs. Os resultados podem ser salvos nos formatos .xls, .csv e .pdf.

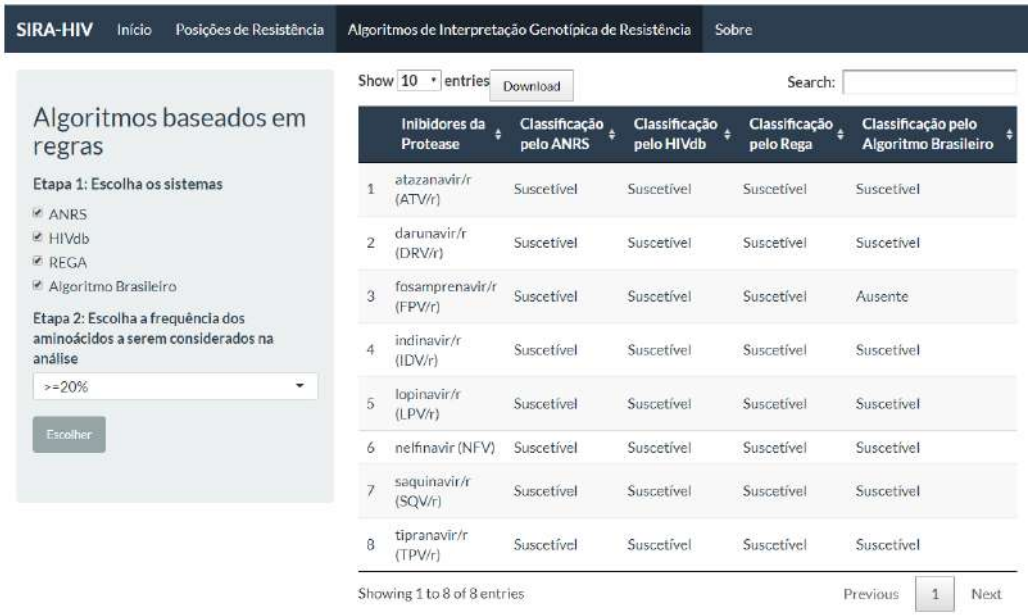


Figura 5.16: Classificação de uma amostra segundo os algoritmos de interpretação genotípica ANRS, HIVdb, Rega e Algoritmo Brasileiro. O ponto de corte da frequência de seleção das mutações foi maior ou igual a 20%.

5.2.2 Validação do Sistema

A fim de validar o sistema desenvolvido, foram utilizadas 10 amostras de indivíduos HIV+, disponibilizados pelo Laboratório de Virologia do CCS/UFRJ, previamente analisadas pelo sistema DeepGen HIV. Apenas as mutações com frequência maior ou igual a 1% foram consideradas. Nas Tabelas 5.18, 5.19 e 5.20, encontram-se as diferenças entre os resultados do SIRA-HIV e do DeepGen HIV. São apresentadas apenas as posições, com seus respectivos aminoácidos e frequências, que variaram entre os programas. Em negrito encontram-se as variações com frequência superior a 2%.

Ao considerar apenas as mutações discordantes e utilizando voto majoritário das classificações pelo ANRS, HIVdb e Rega, para a PR, apenas a amostra 3 apresentaria uma mudança do nível de suscetibilidade. As mutações encontradas pelo SIRA-HIV tornariam a amostra suscetível em resistente ao ATV. Poucas mutações discordantes apareceram com frequências acima de 2%. Destacam-se as mutações 50M, 88D e 88S, encontradas pelo SIRA-HIV, com frequências próximas ou maiores a 20% (Tabela 5.18).

Em relação à RT, o número de mutações discordantes foi maior e, na maioria dos casos, ocorreu mudança no perfil de resistência quando consideradas apenas essas mutações. Entretanto, apenas duas mutações apresentaram porcentagens detectáveis pelo sequenciamento de Sanger (próximas de 20%), a 181F e 181S, identificadas pelo SIRA-HIV (Tabela 5.19).

Figura 5.17: Classificação de uma amostra segundo os algoritmos de interpretação genotípica ANRS, HIVdb, Rega e Algoritmo Brasileiro. O ponto de corte da frequência de seleção das mutações foi maior ou igual a 1%.

Para a IN, em dois casos ocorreu mudança do nível de suscetibilidade quando utilizado o voto majoritário das classificações pelo ANRS, HIVdb e Rega. A amostra 4 apresentaria resistência intermediária ao EVG para os dois sistemas e, em relação ao RAL, o SIRA-HIV classificaria como resistência intermediária. Já a amostra 10 apresentaria mudança de classificação apenas para o SIRA-HIV. Nenhuma mutação discordante apresentou frequência próxima ou maior que 20% (Tabela 5.20).

Em todos os casos, a mudança de classificação corresponde ao voto majoritário das classificações pelo ANRS, HIVdb e Rega.

No Anexo 3, encontra-se a lista completa dos aminoácidos encontrados em cada posição pelos sistemas DeepGen HIV e SIRA-HIV.

Tabela 5.18: Discordâncias entre o DeepGen HIV e o SIRA-HIV na determinação dos aminoácidos e suas frequências presentes na protease. Em negrito, encontram-se as variações com frequência superior a 2%. Na última coluna, está registrada a mudança no nível de suscetibilidade (de suscetível para resistência intermediária ou resistência) quando consideradas apenas as mutações encontradas por cada sistema.

Amostra	DeepGen HIV	SIRA-HIV	Mudança de nível de suscetibilidade
1	43I (1,16%), 47G (1,42%), 48R (1,47%), 69Y (1,2%)	-	Não
2	16R (1,18%), 47S (1,06%), 48R (1,46%)	-	Não
3	36K (1,99%)	33S (1,18%), **50M (17,98%)**, **88S (24,79%)**, **88D (24,67%)**	SIRA-HIV: ATV (R)
4	48R (1,36%)	-	Não
5	16R (1,96%)	-	Não
6	20I (1,64%)	20M (1,66%)	Não
7	**36K (2,44%)**	-	Não
8	**10P (5,81%)**	-	Não
9	**10P (2,5%)**	-	Não
10	-	-	-

R: resistência

Fonte: Laboratório de Virologia Molecular - Centro de Ciências da Saúde/Universidade Federal do Rio de Janeiro

Tabela 5.19: Discordâncias entre o DeepGen HIV e o SIRA-HIV na determinação dos aminoácidos e suas frequências presentes na transcriptase reversa. Em negrito, encontram-se as variações com frequência superior a 2%. Na última coluna, está registrada a mudança no nível de suscetibilidade (de suscetível para resistência intermediária ou resistência), quando consideradas apenas as mutações encontradas por cada sistema.

Amostra	DeepGen HIV	SIRA-HIV	Mudança de Nível de Suscetibilidade
1	-	-	-
2	67N (1,66%), 98D (1,32%)	115F (1,26%), 115* (1,07%)	SIRA-HIV: ABC (R)
3	**69N (3,56%)**, **101G (6,2%)**, 103E (1,45%)	**69T (3,46%)**, **69Y (3,41%)**, 100V (1,69%), 100F (1,49%), **101R (6,68%)**, **101E (6,47%)**, 225T (1,15%)	SIRA-HIV: EFV(I), NVP (R), RPV (R)
4	221N (1,09%)	100V (1,06%), 100F (1,03%)	Não
5	67N (1,27%), 221N (1,71%)	-	Não
6	67N (1,45%), **215Y (2,59%)**	**179I (3,86%)**, **215S (2,46%)**, **215N (2,4%)**	DeepGen HIV: ABC (I), AZT (I), d4T (I)) SIRA-HIV (AZT (I), d4T (I))
7	101K (1,06%), 103E (1,6%)	-	Não
8	227I (1,37%)	**100V (2,06%)**	Não
9	-	100V (1,74%), 225L (1,18%)	Não
10	-	65R (1,4%), **181F (23,99%)**, **181S (19,17%)**, **181N (3,21%)**, 181G (1,34%)	SIRA-HIV: ABC (I), d4T (R), ddI (R), FTC (I), 3TC (I), TDF (R)

*: códon de parada; R: resistência; I: resistência intermediária

Fonte: Laboratório de Virologia Molecular - Centro de Ciências da Saúde/Universidade Federal do Rio de Janeiro

Tabela 5.20: Discordâncias entre o DeepGen HIV e o SIRA-HIV na determinação dos aminoácidos e suas frequências presentes na integrase. Em negrito, encontram-se as variações com frequência superior a 2%. Na última coluna, está registrada a mudança no nível de suscetibilidade (de suscetível para resistência intermediária ou resistência) quando consideradas apenas as mutações encontradas por cada sistema.

Indivíduo	DeepGen HIV	SIRA-HIV	Mudança de Nível de Suscetibilidade
1	-	-	-
2	-	-	-
3	-	-	-
4	**140C (2,99%)**	**140S (3,12%)**	DeepGen HIV: EVG (I) SIRA-HIV: EVG (I), RAL (I)
5	-	-	-
6	-	-	-
7	-	-	-
8	**101F (4,37%)**	-	Não
9	**101F (4,97%)**	193E (1,03%)	Não
10	**143Y (2,19%)**	**143H (2,66%), 143C (2,1%)**	SIRA-HIV: RAL (R)

R: resistência; I: resistência intermediária

Fonte: Laboratório de Virologia Molecular - Centro de Ciências da Saúde/Universidade Federal do Rio de Janeiro

Capítulo 6

Discussão

Diversos estudos têm apontado discordâncias de classificação do nível de suscetibilidade do HIV-1 aos ARVs entre os três principais algoritmos de interpretação genotípica (ANRS, HIVdb e Rega). Esses sistemas são amplamente utilizados na prática clínica como ferramenta auxiliar na escolha do melhor regime terapêutico a ser recomendado para cada indivíduo HIV+. O presente estudo propôs três estratégias de combinação de classificadores a fim de fornecer um único perfil de resistência a partir dos resultados dos algoritmos ANRS, HIVdb e Rega. As abordagens utilizadas foram o voto majoritário, a escolha do melhor algoritmo de interpretação para cada ARV e uma adaptação da técnica *stacking*.

Neste trabalho, a abordagem *stacking* foi aplicada com o objetivo de obter um multiclassificador, utilizando as saídas dos algoritmos de interpretação como dados de entrada. Uma vez que classificadores com desempenhos de generalização similares podem trabalhar diferentemente, a ideia por trás do *stacking* é desenvolver um metaclassificador que irá aprender a melhor forma de combinar as saídas dos classificadores a fim de alcançar resultados mais acurados. Para tal, foram utilizados dois tipos de algoritmos, o NB e o k-NN, métodos simples e de rápido aprendizado empregados também como metaclassificadores em outros estudos [142–147].

A escolha do melhor algoritmo de interpretação genotípica para cada ARV foi realizada com base nos valores da medida-F, também aplicada como critério de ranqueamento no teste de Friedman. Essa medida foi escolhida por representar, em um único valor, a combinação da sensibilidade e do VPP, sumarizando achados relativos a amostras resistentes.

O teste de Friedman mostrou que, para cinco ARVs, os multiclassificadores apresentaram diferença estatisticamente significante entre seus desempenhos. Desses, três fármacos (LPV, SQV e ABC) fazem parte do grupo de ARVs com três possíveis desfechos de classificação. Essa diferença provavelmente está relacionada à grande variabilidade dos valores das medidas de desempenho para a categorização resistência intermediária vs. não intermediária. Para alguns multiclassificadores,

medidas como a sensibilidade e o VPP apresentaram valores abaixo de 10% para essa categorização. É esperado que a classe intermediária apresente um maior número de erros por se localizar entre duas classificações que correspondem a eventos opostos. Adicionalmente, é possível observar que esses ARVs foram os que apresentaram maiores flutuações nos valores das medidas de desempenho, com o NB apresentando o melhor desempenho, sendo superior ao VM para os três fármacos e ao MS para o SQV e ABC.

Na comparação pareada, o teste de Holm indicou diferença estatisticamente significante entre o classificador-controle NB e o VM para o LPV, SQV, ABC e d4T. O NB apresentou desempenho superior também em relação ao MS para o SQV e ABC e se mostrou superior ao k-NN para os fármacos ABC e d4T. Para o ddI, o classificador-controle k-NN se mostrou superior ao VM e NB, entretanto o valor-p encontrado situa-se muito próximo do nível de significância considerado. A partir desses resultados, é possível observar que o NB apresentou os melhores desempenhos, ou resultados pelo menos comparáveis, às outras estratégias aplicadas.

As estratégias que mais se destacaram em relação às medidas de desempenho avaliadas com o conjunto de teste foram o MS e o NB, principalmente quando levada em consideração a medida-F. De uma forma geral, o VM foi a metodologia que apresentou os menores valores de desempenho. Para alguns casos, os metaclassificadores NB e k-NN apresentaram desempenho equivalente ao do VM, indicando um comportamento similar a uma simples votação dos desfechos dos classificadores-base.

Em relação às abordagens MS e NB, cabe observar que a última possui a vantagem de apresentar, em sua composição, informações dos três algoritmos de interpretação, enquanto que, para fazer uso da MS, é necessário ter o conhecimento do algoritmo com melhor desempenho para cada ARV.

Cabe destacar também que, neste presente trabalho, antes do desenvolvimento dos multiclassificadores, o conjunto de dados de treinamento de cada ARV foi balanceado em relação aos níveis de suscetibilidade. Esse procedimento foi adotado a fim de que a classe sub-representada (classe resistente para a maioria dos ARVs) não fosse secundarizada ou mesmo ignorada no treinamento dos metaclassificadores NB e k-NN. O desequilíbrio entre as classes pode gerar classificadores com boa acurácia na predição da classe majoritária, mas com um desempenho muito baixo com relação à(s) classe(s) minoritária(s). Estudos mostram que, para diferentes classificadores, a aplicação de conjuntos de dados balanceados geralmente melhoram o desempenho geral do modelo em comparação a um conjunto de dados não balanceados [148–150].

Em 2012, Yashik e colaboradores [14] também desenvolveram multiclassificadores a partir dos algoritmos ANRS, HIVdb e Rega usando técnicas de votação ponderada. Eles encontraram acurácias significativas para as estratégias aplicadas. No entanto, não apresentaram outras medidas relevantes de desempenho, nem mesmo forneceram

informações mais detalhadas sobre a modelagem dos dados, limitando a comparação dos seus resultados com os encontrados no presente estudo. Outro detalhe a ser destacado é que, por não indicarem se os dados estavam balanceados ou não, a medida de desempenho escolhida pode não ser adequada para avaliar o desempenho dos modelos, uma vez que a acurácia é apenas influenciada pelo número de acertos. Por exemplo, se um conjunto possui 98% dos dados formado pela classe majoritária e 2% pela classe minoritária, caso um classificador acerte todos os casos da classe majoritária, mas classifique erroneamente todos os casos minoritários, o modelo apresentará uma ótima acurácia, igual a 98%. Porém, a generalidade dos seus achados estará aquém do desejável.

Algumas limitações a respeito dos multiclassificadores devem ser mencionadas. Primeiro, a maioria das amostras empregadas pertencem ao subtipo B, o que não contribui para a generalidade do desempenho das estratégias aplicadas para outros subtipos existentes. É possível que os multiclassificadores desenvolvidos não possam ser extrapolados para amostras de subtipos não B, uma vez que alguns estudos [16, 151] observaram diferentes valores de desempenho para os algoritmos ANRS, HIVdb e Rega para esses subtipos. Subtipos não B são um desafio para esses sistemas de interpretação, visto que a maioria foi projetada usando informações de genótipo, fenótipo e resposta terapêutica derivadas da experiência com o subtipo B, Esse subtipo predomina exatamente nos locais onde são constituídos grandes repositórios e se dispõe de recursos computacionais e *expertise* (EUA e Europa).

A categorização dos níveis de resistência também deve ser apontada como uma possível limitação deste estudo. Os mesmos níveis de suscetibilidade do teste fenotípico foram utilizados para os algoritmos de interpretação. Alguns ARVs foram categorizados em apenas duas classes (suscetível e resistente), enquanto que a minoria foi classificada em três níveis, incluindo a classe de resistência intermediária. Essa categorização pode ter reduzido o desempenho global das estratégias de classificação, principalmente o do sistema HIVdb, uma vez que esse algoritmo reconhece mais de três níveis de resistência.

Outra limitação deste estudo pode estar relacionada aos métodos de balanceamento de dados utilizados. A subamostragem, ao remover exemplos da classe majoritária, pode subtrair algumas informações úteis dos conjuntos de dados, fazendo com que o classificador perca conceitos importantes pertencentes a essa classe. Em relação à sobreamostragem, uma vez que ela simplesmente adiciona dados replicados ao conjunto original, várias instâncias se repetem, podendo levar a um sobreajustamento do modelo [152].

Adicionalmente aos multiclassificadores, este trabalho desenvolveu um ambiente integrado capaz de identificar mutações presentes no genoma do HIV-1 e categorizar o nível de suscetibilidade do vírus a cada ARV a partir de dados de NGS. O sistema

recebeu o nome de SIRA-HIV e foi implementado em R.

Nesse sistema, foi desenvolvida uma plataforma em que usuários não familiarizados com linhas de comando e outros conhecimentos de programação pudessem facilmente analisar as sequências de nova geração obtidas a partir de amostras de indivíduos HIV+. Para realizar o mapeamento e alinhamento das sequências, o software Segminator II foi escolhido por já ter sido aplicado em outros estudos [106, 108, 136], ser específico para a caracterização de dados virais e apresentar uma interface gráfica de fácil manuseio.

Para validar o SIRA-HIV, o sistema DeepGen HIV foi escolhido por utilizar também o Segminator II no mapeamento das leituras de DNA. Apesar das diferenças encontradas entre o SIRA-HIV e o DeepGen HIV, pode-se considerar que ambos os sistemas apresentaram desempenho similares, com um maior número de mutações discordantes para a região da RT. Entretanto, pode-se notar que apenas em dois casos as mutações apresentavam uma frequência acima de 20%, que seria, de fato, considerado pelo teste genotípico padrão. Assim, segundo o SIRA-HIV, a amostra 3 seria classificada como resistente ao ATV e a amostra 10 seria resistente ao d4T, ddI e TDF e teria resistência intermediária ao ABC, FTC e 3TC. As mudanças do nível de suscetibilidade foram registradas levando em consideração apenas as mutações discordantes e o voto majoritário dos algoritmos ANRS, HIVdb e Rega. Não foram avaliadas as classificações considerando todas as mutações indicadas por cada sistema, o que constitui uma limitação. Contudo, tal alternativa, em prol da exaustividade, não é exequível dentro dos marcos de uma tese.

É importante destacar que, embora as mutações resistentes minoritárias estejam presentes, por haver ainda muito debate sobre sua relevância clínica, elas ainda não são consideradas nas tomadas de decisão sobre o melhor esquema terapêutico [153]. Mas não há como distinguir por ora, se isso se deve à sua pouca relevância clínica ou à dificuldade de sua detecção pelos procedimentos padrão. Entretanto, é esperado que essas mutações minoritárias, em contato com os medicamentos após o início do tratamento, sejam selecionadas naturalmente, aumentando suas frequências na população e levando à falha terapêutica.

Uma das prováveis explicações para as diferenças encontradas entre os sistemas pode estar centrada nas definições dos parâmetros de mapeamento do Segminator II. Neste estudo, foram utilizados os valores *default* do programa, com exceção da opção *Replace Template with Con. During Mapping*, em *Miscellaneous*, que foi selecionada. Essa função tem como objetivo remapear as leituras, em relação a uma sequência consenso gerada a partir do primeiro mapeamento, a fim de reduzir a diversidade entre elas. Essa opção também é utilizada pelo DeepGen HIV, entretanto, não foi possível obter informações a respeito dos demais parâmetros utilizados por esse sistema. Alterações nos valores podem ocasionar mudanças significativas no

mapeamento e, consequentemente, gerar resultados parcialmente diferentes entre as análises.

Outra possível fonte de discordância entre as mutações identificadas pode estar relacionada à sequência de referência utilizada no mapeamento. No DeepGen HIV, a sequência de referência é escolhida a partir do banco de dados *Los Alamos HIV Sequence Database*. A sequência mais similar a 100 leituras selecionadas aleatoriamente do conjunto de dados de NGS é utilizada como referência. Neste estudo, foi utilizada a sequência HXB2, correspondente ao genoma do vírus selvagem do subtipo B do HIV-1.

Em relação à interface do sistema, o SIRA-HIV está estruturado em quatro telas, compreendendo dois módulos de análise. No módulo Posições de Resistência, os resultados são exibidos em forma de tabela, contendo as posições de resistência, seus aminoácidos originais e os presentes na amostra, suas frequências e a cobertura. Uma vez que é sugerido uma cobertura mínima em torno de 450 nucleotídeos em regiões não homopoliméricas (sem repetições dos nucleotídeos) para garantir a detecção de variantes minoritárias presentes em mais de 1% da população [24], torna-se importante a visualização dessa variável na tomada de decisão acerca da confiabilidade dos dados. Portanto, foi incluído um gráfico de cobertura por posição para facilitar na visualização dos valores. No módulo Algoritmos de Interpretação Genotípica, além do Algoritmo Brasileiro, disponibilizado pelo Ministério da Saúde, os algoritmos internacionais ANRS, HIVdb e Rega foram incluídos no SIRA-HIV. Esses sistemas são os mais utilizados mundialmente, estão disponíveis publicamente e são atualizados regularmente.

Diversas outras ferramentas de análise de dados de NGS de amostras de HIV já foram desenvolvidas. É o caso dos softwares MinVar [70] e HyDRA Web [113]. O MinVar foi implementado em Python, logo é necessária a utilização de linhas de comando em sua análise, o que limita o número de possíveis usuários do programa. Além disso, o software não fornece nenhum resultado do nível de suscetibilidade aos ARVs, exibindo como resultado apenas as mutações encontradas e suas respectivas frequências. Em relação ao HyDRA Web, apesar de ser um sistema robusto e bem completo, apresentando-se também em um formato amigável, esse software também não disponibiliza dados acerca da suscetibilidade do HIV aos fármacos. Assim, para os dois sistemas mencionados, é necessário que o usuário utilize manualmente algum algoritmo de interpretação genotípica. Diferentemente de outros sistemas disponíveis, o SIRA-HIV contempla, além da identificação das mutações, quatro algoritmos de interpretação genotípica e classifica segundo dois pontos de corte; um considerando apenas as mutações majoritárias (aquelas que aparecem com frequências maiores ou iguais a 20%) e o outro levando em consideração também as mutações minoritárias (todas as mutações com frequências de até 1%).

Algumas limitações em relação ao desenvolvimento e validação do SIRA-HIV devem ser apontadas. O programa de mapeamento Segminator II utilizado pelo SIRA-HIV permite como entrada do conjunto de dados a ser analisado apenas arquivos no formato FASTQ. Isso pode limitar o uso do sistema, uma vez que algumas plataformas geram como saídas arquivos em outros formatos, como o .SFF. Entretanto, há ferramentas de bioinformática que convertem diferentes tipos de arquivos em FASTQ, superando essa limitação. Em relação à validação do sistema, poucas amostras foram utilizadas na comparação do SIRA-HIV com o DeepGen HIV. Essas amostras foram disponibilizadas pelo Laboratório de Virologia Molecular do CCS-UFRJ e uma explicação para o número reduzido reside no fato de que o programa DeepGen HIV é pago, o que acaba limitando o seu uso. Outro ponto a ser levantado diz respeito à avaliação da usabilidade do SIRA-HIV, o que ainda está por ser desenvolvida. Porém, é particularmente importante que essa etapa seja realizada a fim de verificar a facilidade/dificuldade de uso e a compreensibilidade do sistema por parte do usuário.

Capítulo 7

Conclusão

7.1 Contribuição

Neste estudo, foram desenvolvidos multiclassificadores, a partir dos algoritmos ANRS, HIVdb e Rega, com o objetivo de fornecer um único perfil de resistência para cada ARV. Três estratégias diferentes foram utilizadas: voto majoritário (VM), escolha do melhor sistema de interpretação genotípica (MS) e técnica *stacking*, com metaclassificadores *naïve Bayes* (NB) e k-NN. No geral, as abordagens NB e MS obtiveram os melhores resultados, com o NB apresentando desempenho estatisticamente superior a pelo menos uma das outras três estratégias para os fármacos LPV, SQV, ABC e d4T.

O metaclassificador NB se mostra vantajoso ao apresentar em sua formação os três algoritmos de interpretação genotípica mais utilizados na prática clínica, diferentemente da estratégia MS, ancorada em apenas um dos sistemas, havendo ainda a necessidade de se conhecer qual algoritmo de interpretação apresenta melhor desempenho para cada ARV.

Ao sugerir o uso dos multiclassificadores desenvolvidos neste trabalho, o objetivo não é propor uma substituição ou redução do uso dos algoritmos originais, mas apresentar uma abordagem metodológica diferente e, consequentemente, fornecer uma nova ferramenta alternativa baseada nos três principais sistemas utilizados em todo o mundo.

Em relação ao ambiente integrado SIRA-HIV, cabe observar que ele se mostrou adequado, funcionando de maneira simples e rápida, não exigindo, por parte do usuário, conhecimentos de programação e/ou linhas de comando, comumente requisitados no uso de ferramentas de bioinformática. Assim, usuários não familiarizados com a criação de *scripts* podem facilmente analisar as sequências de nova geração. O sistema proposto possui ainda a vantagem de apresentar, em uma única plataforma, uma avaliação abrangente dos dados de NGS, fornecendo ao usuário uma lista dos

aminoácidos (e suas frequências) encontrados nas regiões analisadas, além da classificação de resistência do HIV-1 aos ARVs. Outro benefício é a capacidade de o sistema considerar dois níveis de frequência dos aminoácidos na classificação de resistência. No limiar de 20%, o usuário tem acesso à classificação utilizada na prática clínica. Quando consideradas todas as variantes com frequência maior ou igual a 1%, o impacto das mutações resistentes minoritárias é levado em consideração. Essa informação pode ajudar o clínico a decidir sobre o melhor regime de tratamento a ser adotado para cada paciente infectado pelo HIV-1, uma vez que essas variantes minoritárias podem, sob pressão contínua de medicamentos diversos, se tornar a principal população viral e aumentar o risco de falha virológica precoce.

Na validação do SIRA-HIV, o sistema identificou, em sua grande maioria, os mesmos aminoácidos por posição relatados pelo DeepGen HIV. Em apenas dois casos as mutações discordantes apresentaram uma frequência acima de 20%.

Dessa forma, o ambiente SIRA-HIV se mostra uma ferramenta promissora na área da bioinformática voltada para análise de dados de NGS, permitindo que clínicos e laboratórios identifiquem as populações virais, tanto majoritárias quanto minoritárias, presentes no organismo do paciente e tenham acesso aos níveis de suscetibilidade do HIV-1 aos ARVs. O programa poderá auxiliar na melhor tomada de decisão quanto aos medicamentos a serem administrados para cada paciente, proporcionando, consequentemente, benefícios às pessoas vivendo com HIV.

7.2 Trabalhos Futuros

Como trabalhos futuros, sugere-se a utilização de novos algoritmos como metaclassificadores na técnica *stacking* em busca de modelos com melhores desempenhos e a incorporação de outros sistemas de interpretação. Uma possível abordagem é a utilização de modelos log-lineares ou modelos logísticos multicategóricos. Outra proposta é o desenvolvimento de classificadores, por meio de diferentes técnicas de aprendizagem de máquina, utilizando dados de sequenciamento (por exemplo, informações sobre as mutações).

Em relação aos dados, métodos mais robustos de balanceamento, como a técnica SMOTE (*Synthetic Minority Over-sampling Technique*), podem ser empregados a fim de obter classes melhor representadas.

Sobre o SIRA-HIV, sugere-se:

- Incorporar futuramente modelos preditivos de resistência a fim de atuarem como uma ferramenta auxiliar na classificação do nível de suscetibilidade do HIV-1;
- Desenvolver modelos de classificação de tropismo do HIV-1 e inseri-los como

um novo módulo do sistema;

- Incluir uma análise do subtipo viral;
- Comparar o SIRA-HIV a outros softwares disponíveis;
- Realizar uma avaliação de usabilidade do sistema em condições concretas de operação e;
- Disponibilizar o sistema por meio de um repositório online para fins acadêmicos.

Referências Bibliográficas

[1] UNAIDS. *UNAIDS Data 2017*. Relatório técnico, 2017. 1, 2.2

[2] DA SAÚDE, M. *Boletim Epidemiológico - Aids e DST*. Relatório técnico, Departamento de DST/Aids e Hepatites Virais, 2017. 1

[3] HYMES, K., GREENE, J., MARCUS, A., et al. "Kaposi's sarcoma in homosexual men - a report of eight cases", *The Lancet*, v. 318, n. 8247, pp. 598–600, 1981. 1

[4] FISCHL, M. A., RICHMAN, D. D., GRIECO, M. H., et al. "The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex", *New England Journal of Medicine*, v. 317, n. 4, pp. 185–191, 1987. 1, 2.2

[5] DEEKS, S. G. "Treatment of antiretroviral-drug-resistant HIV-1 infection", *The Lancet*, v. 362, n. 9400, pp. 2002–2011, 2003. 1

[6] VAN DE VIJVER, D., WENSING, A. M., BOUCHER, C. A., et al. "The epidemiology of transmission of drug resistant HIV-1", *HIV sequence compendium*, v. 2007, pp. 17–36, 2006. 1

[7] BRONZE, M., STEEGEN, K., WALLIS, C., et al. "HIV-1 Phenotypic Reverse Transcriptase Inhibitor Drug Resistance Test", *In vitro HIV-1 drug resistance phenotyping, genotyping and novel virological failure detection tools for clinical patient management*, p. 63, 2013. 1, 2.2.2

[8] BEERENWINKEL, N., SCHMIDT, B., WALTER, H., et al. "Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype", *Proceedings of the National Academy of Sciences*, v. 99, n. 12, pp. 8271–8276, 2002. 1, 2.2.2

[9] VERCAUTEREN, J., VANDAMME, A.-M. "Algorithms for the interpretation of HIV-1 genotypic drug resistance information", *Antiviral research*, v. 71, n. 2, pp. 335–342, 2006. 2.4.1, 2.4.2, 2.4.3

[10] SCHUTTEN, M. *Antiretroviral Resistance in Clinical Practice*, v. Chapter 5. Mediscript, 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/books/NBK2252/>. 1, 2.2.2

[11] MEYNARD, J.-L., VRAY, M., MORAND-JOUBERT, L., et al. "Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial", *Aids*, v. 16, n. 5, pp. 727–736, 2002. 1, 2.4.1

[12] VAN LAETHEM, K., DE LUCA, A., ANTINORI, A., et al. "A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients", *Antiviral therapy*, v. 7, n. 2, pp. 123–9, 2002. 1, 2.4.3

[13] SHAFER, R. W., STEVENSON, D., CHAN, B. "Human immunodeficiency virus reverse transcriptase and protease sequence database", *Nucleic acids research*, v. 27, n. 1, pp. 348–352, 1999. 1, 2.4.2

[14] YASHIK, S., MAURICE, M. "Predicting a single HIV drug resistance measure from three international interpretation gold standards", *Asian Pacific journal of tropical medicine*, v. 5, n. 7, pp. 566–572, 2012. 1, 2.4, 2.4.1, 2.4.2, 6

[15] RAVELA, J., BETTS, B. J., BRUN-VÉZINET, F., et al. "HIV-1 protease and reverse transcriptase mutation patterns responsible for discordances between genotypic drug resistance interpretation algorithms", *Journal of acquired immune deficiency syndromes (1999)*, v. 33, n. 1, pp. 8, 2003. 1, 3

[16] VERGNE, L., SNOECK, J., AGHOKENG, A., et al. "Genotypic drug resistance interpretation algorithms display high levels of discordance when applied to non-B strains from HIV-1 naive and treated patients", *FEMS Immunology & Medical Microbiology*, v. 46, n. 1, pp. 53–62, 2006. 3, 6

[17] POONPIRIYA, V., SUNGKANUPARPH, S., LEECHANACHAI, P., et al. "A study of seven rule-based algorithms for the interpretation of HIV-1 genotypic resistance data in Thailand", *Journal of virological methods*, v. 151, n. 1, pp. 79–86, 2008. 3

[18] YEBRA, G., DE MULDER, M., DEL ROMERO, J., et al. "HIV-1 non-B subtypes: High transmitted NNRTI-resistance in Spain and impaired genotypic resistance interpretation due to variability", *Antiviral research*, v. 85, n. 2, pp. 409–417, 2010. 3

[19] WAGNER, S., KURZ, M., KLIMKAIT, T. "Algorithm evolution for drug resistance prediction: comparison of systems for HIV-1 genotyping." *Antiviral therapy*, 2015. 1, 3

[20] GIBSON, R. M., SCHMOTZER, C. L., QUIÑONES-MATEU, M. E. "Next-Generation Sequencing to Help Monitor Patients Infected with HIV: Ready for Clinical Use?" *Current infectious disease reports*, v. 16, n. 4, pp. 1–9, 2014. 1, 2.2.2, 2.3.2, 3

[21] ERALI, M., PAGE, S., REIMER, L. G., et al. "Human immunodeficiency virus type 1 drug resistance testing: a comparison of three sequence-based methods", *Journal of clinical microbiology*, v. 39, n. 6, pp. 2157–2165, 2001. 1, 2.2.2, 2.3.1, 3

[22] ESHLEMAN, S. H., CRUTCHER, G., PETRAUSKENE, O., et al. "Sensitivity and specificity of the ViroSeq human immunodeficiency virus type 1 (HIV-1) genotyping system for detection of HIV-1 drug resistance mutations by use of an ABI PRISM 3100 genetic analyzer", *Journal of clinical microbiology*, v. 43, n. 2, pp. 813–817, 2005.

[23] PALMER, S., KEARNEY, M., MALDARELLI, F., et al. "Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis." *J Clin Microbiol*, v. 43, n. 1, pp. 406–413, Jan 2005. doi: 10.1128/JCM.43.1.406-413.2005. Disponível em: <`http://dx.doi.org/10.1128/JCM.43.1.406-413.2005`>. 1, 2.2.2, 2.3.1, 3

[24] WANG, C., MITSUYA, Y., GHARIZADEH, B., et al. "Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance." *Genome Res*, v. 17, n. 8, pp. 1195–1201, Aug 2007. doi: 10.1101/gr.6468307. Disponível em: <`http://dx.doi.org/10.1101/gr.6468307`>. 1, 2.2.2, 2.3.2, 3, 6

[25] JOHNSON, J. A., LI, J.-F., WEI, X., et al. "Minority HIV-1 drug resistance mutations are present in antiretroviral treatment - naive populations and associate with reduced treatment efficacy", *PLoS medicine*, v. 5, n. 7, pp. e158, 2008. 1, 2.3.2, 3

[26] SIMEN, B. B., SIMONS, J. F., HULLSIEK, K. H., et al. "Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment–naive patients significantly impact treatment outcomes", *Journal of Infectious Diseases*, v. 199, n. 5, pp. 693–701, 2009. 3

[27] METZNER, K. J., GIULIERI, S. G., KNOEPFEL, S. A., et al. "Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and-adherent patients", *Clinical infectious diseases*, v. 48, n. 2, pp. 239–247, 2009. 3

[28] LE T, C. J., SIMEN, B., HANCZARUK, B., et al. "Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use", *PloS one*, v. 4, n. 6, pp. e6079, 2009. 1, 3

[29] BOLTZ, V. F., ZHENG, Y., LOCKMAN, S., et al. "Role of low-frequency HIV-1 variants in failure of nevirapine-containing antiviral therapy in women previously exposed to single-dose nevirapine", *Proceedings of the National Academy of Sciences*, v. 108, n. 22, pp. 9202–9207, 2011. 3

[30] VANDENHENDE, M.-A., BELLECAVE, P., RECORDON-PINSON, P., et al. "Prevalence and evolution of low frequency HIV drug resistance mutations detected by ultra deep sequencing in patients experiencing first line antiretroviral therapy failure." *PLoS One*, v. 9, n. 1, pp. e86771, 2014. doi: 10.1371/journal.pone.0086771. Disponível em: <`http://dx.doi.org/10.1371/journal.pone.0086771`>. 1, 3

[31] FONAGER, J., LARSSON, J. T., HUSSING, C., et al. "Identification of minority resistance mutations in the HIV-1 integrase coding region using next generation sequencing", *Journal of Clinical Virology*, v. 73, pp. 95–100, 2015. 1, 3

[32] LATAILLADE, M., CHIARELLA, J., YANG, R., et al. "Virologic failures on initial boosted-PI regimen infrequently possess low-level variants with major PI resistance mutations by ultra-deep sequencing", *PloS one*, v. 7, n. 2, pp. e30118, 2012. 1

[33] NEKRUTENKO, A., TAYLOR, J. "Next-generation sequencing data interpretation: enhancing reproducibility and accessibility", *Nature Reviews Genetics*, v. 13, n. 9, pp. 667–672, 2012. 1

[34] BARRÉ-SINOUSSI, F. "HIV as the cause of AIDS", *The Lancet*, v. 348, n. 9019, pp. 31–35, 1996. 2.1.1

[35] NYAMWEYA, S., HEGEDUS, A., JAYE, A., et al. "Comparing HIV-1 and HIV-2 infection: Lessons for viral immunopathogenesis", *Reviews in medical virology*, v. 23, n. 4, pp. 221–240, 2013. 2.1.1

[36] SIMON, F., MAUCLÈRE, P., ROQUES, P., et al. "Identification of a new human immunodeficiency virus type 1 distinct from group M and group O", *Nature medicine*, v. 4, n. 9, pp. 1032–1037, 1998. 2.1.1

[37] PLANTIER, J.-C., LEOZ, M., DICKERSON, J. E., et al. "A new human immunodeficiency virus derived from gorillas." *Nat Med*, v. 15, n. 8, pp. 871–872, Aug 2009. doi: 10.1038/nm.2016. Disponível em: <http://dx.doi.org/10.1038/nm.2016>. 2.1.1

[38] BUONAGURO, L., TORNESELLO, M. L., BUONAGURO, F. M. "Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications." *J Virol*, v. 81, n. 19, pp. 10209–10219, Oct 2007. doi: 10.1128/JVI.00872-07. Disponível em: <http://dx.doi.org/10.1128/JVI.00872-07>. 2.1.1, 2.1.1

[39] HEMELAAR, J. "The origin and diversity of the HIV-1 pandemic", *Trends in molecular medicine*, v. 18, n. 3, pp. 182–192, 2012. 2.1.1, 2.1.1

[40] BURKE, D. S. "Recombination in HIV: an important viral evolutionary strategy." *Emerg Infect Dis*, v. 3, n. 3, pp. 253–259, 1997. doi: 10.3201/eid0303.970301. Disponível em: <http://dx.doi.org/10.3201/eid0303.970301>. 2.1.1

[41] DA SAÚDE, M. *Boletim Epidemiológico - AIDS e DST*. Relatório técnico, Departamento de DST/Aids e Hepatites Virais, 2015. 2.1.1

[42] SIERRA, S., KUPFER, B., KAISER, R. "Basics of the virology of HIV-1 and its replication." *J Clin Virol*, v. 34, n. 4, pp. 233–244, Dec 2005. doi: 10.1016/j.jcv.2005.09.004. Disponível em: <http://dx.doi.org/10.1016/j.jcv.2005.09.004>. 2.1.2, 2.1.4

[43] FRANKEL, A. D., YOUNG, J. A. "HIV-1: fifteen proteins and an RNA." *Annu Rev Biochem*, v. 67, pp. 1–25, 1998. doi: 10.1146/annurev.biochem.67.1.1. 2.1.2, 2.1.2, 2.1.2

[44] KUIKEN, C., LEITNER, T., FOLEY, B., et al. "HIV sequence compendium 2009", *Los Alamos, New Mexico: Los Alamos National Laboratory, Theoretical Biology and Biophysics*, 2009. 2.1.2

[45] CLAPHAM, P. R., MCKNIGHT, A. "HIV-1 receptors and cell tropism." *Br Med Bull*, v. 58, pp. 43–59, 2001. 2.1.3

[46] BERGER, E. A., DOMS, R. W., FENYÖ, E. M., et al. "A new classification for HIV-1." *Nature*, v. 391, n. 6664, pp. 240, Jan 1998. doi: 10.1038/34571. Disponível em: <http://dx.doi.org/10.1038/34571>. 2.1.3

[47] WYATT, R., SODROSKI, J. "The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens." *Science*, v. 280, n. 5371, pp. 1884–1888, Jun 1998. 2.1.4

[48] HARRICH, D., HOOKER, B. "Mechanistic aspects of HIV-1 reverse transcription initiation", *Reviews in medical virology*, v. 12, n. 1, pp. 31–45, 2002. 2.1.4

[49] JUNG, A., MAIER, R., VARTANIAN, J.-P., et al. "Recombination: Multiply infected spleen cells in HIV patients", *Nature*, v. 418, n. 6894, pp. 144–144, 2002. 2.1.4

[50] VAN MAELE, B., DEBYSER, Z. "HIV-1 integration: an interplay between HIV-1 integrase, cellular and viral proteins", *AIDS rev*, v. 7, n. 1, pp. 26–43, 2005. 2.1.4

[51] PALELLA JR, F. J., DELANEY, K. M., MOORMAN, A. C., et al. "Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection", *New England Journal of Medicine*, v. 338, n. 13, pp. 853–860, 1998. 2.2

[52] WHO. "Policy brief: consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: what's new". November 2015. 2.2

[53] PORTAL DA SAÚDE, S. "Brasil bate recorde de pessoas em tratamento contra o HIV e aids". 2016. Disponível em: <http://portalarquivos.saude.gov.br/images/pdf/2016/janeiro/28/Apresentao-campanha-de-carnaval.pdf>. 2.2

[54] DOURADO, I., DE SM VERAS, M. A., BARREIRA, D., et al. "Tendências da epidemia de Aids no Brasil após a terapia anti-retroviral", *Revista de Saúde Pública*, v. 40, n. supl., pp. 9–17, 2006. 2.2

[55] ARTS, E. J., HAZUDA, D. J. "HIV-1 antiretroviral drug therapy", *Cold Spring Harbor perspectives in medicine*, v. 2, n. 4, pp. a007161, 2012. 2.2

[56] YOUNG, F. E. "The role of the FDA in the effort against AIDS." *Public Health Reports*, v. 103, n. 3, pp. 242, 1988. 2.2.1

[57] SHAFER, R. W. "Genotypic testing for human immunodeficiency virus type 1 drug resistance", *Clinical microbiology reviews*, v. 15, n. 2, pp. 247–277, 2002. 2.2.1

[58] OF HEALTH, N. I. "FDA-Approved HIV Medicines". 2017. Disponível em: <https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/19/58/fda-approved-hiv-medicines>. 2.2.1

[59] DE CLERCQ, E. "The role of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection", *Antiviral research*, v. 38, n. 3, pp. 153–179, 1998. 2.2.1

[60] HWANG, C., SCHÜRMANN, D., SOBOTHA, C., et al. "Antiviral Activity, Safety, and Exposure-Response relationships of GSK3532795, a Second-Generation HIV-1 Maturation Inhibitor, administered as Monotherapy or in Combination with Atazanavir±Ritonavir in a Phase 2a Randomized, Dose-Ranging, Controlled Trial (AI468002)." *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 2017. 2.2.1

[61] GÜNTHARD, H. F., SAAG, M. S., BENSON, C. A., et al. "Antiretroviral drugs for treatment and prevention of HIV infection in adults: 2016 recommendations of the International Antiviral Society–USA panel", *Jama*, v. 316, n. 2, pp. 191–210, 2016. 2.2.1

[62] "Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents". April 2015. Disponível em: <http://www.aidsinfo.nih.gov/ContentFiles/AdultandAdolescentGL.pdf>. 2.2.1

[63] CLAVEL, F., HANCE, A. J. "HIV drug resistance", *New England Journal of Medicine*, v. 350, n. 10, pp. 1023–1035, 2004. 2.2.2

[64] WHO. "HIV drug resistance". 2015. Disponível em: <http://www.who.int/hiv/topics/drugresistance/en/>. 2.2.2

[65] DIAZ, R. S., ACCETTURI, C. A., SUCUPIRA, M. C. "Emergence of resistance mutations preceding virologic failure in patients receiving antiretroviral therapy", *JAIDS Journal of Acquired Immune Deficiency Syndromes*, v. 49, n. 1, pp. 111–113, 2008. 2.2.2

[66] D'AQUILA, R., INTERNATIONAL, A., SCHAPIRO, J., et al. "Drug resistance mutations in HIV-1." *Topics in HIV medicine: a publication of the International AIDS Society, USA*, v. 10, n. 5, pp. 21, 2002. 2.2.2

[67] PARKIN, N., HELLMANN, N., WHITCOMB, J., et al. "Natural variation of drug susceptibility in wild-type human immunodeficiency virus type 1", *Antimicrobial agents and chemotherapy*, v. 48, n. 2, pp. 437–443, 2004. 2.2.2

[68] THOMPSON, M. A., ABERG, J. A., HOY, J. F., et al. "Antiretroviral treatment of adult HIV infection: 2012 recommendations of the International Antiviral Society–USA panel", *Jama*, v. 308, n. 4, pp. 387–402, 2012. 2.2.2

[69] PARIKH, U. M., MCCORMICK, K., VAN ZYL, G., et al. "Future technologies for monitoring HIV drug resistance and cure", *Current Opinion in HIV and AIDS*, v. 12, n. 2, pp. 182–189, 2017. 2.2.2

[70] HUBER, M., METZNER, K. J., GEISSBERGER, F. D., et al. "MinVar: A rapid and versatile tool for HIV-1 drug resistance genotyping by deep sequencing", *Journal of Virological Methods*, v. 240, pp. 7–13, 2017. 2.2.2, 3, 6

[71] METZKER, M. L. "Sequencing technologies - the next generation", *Nature reviews genetics*, v. 11, n. 1, pp. 31–46, 2010. 2.2.2, 2.3, 2.3.2

[72] SANGER, F., NICKLEN, S., COULSON, A. R. "DNA sequencing with chain-terminating inhibitors", *Proceedings of the National Academy of Sciences*, v. 74, n. 12, pp. 5463–5467, 1977. 2.3, 2.3.1

[73] ARTEM E. MEN, PETER WILSON, K. S. S. F. *Next-Generation Genome Sequencing: Towards Personalized Medicine*, v. DNA Representation. 2008. 2.3.1

[74] SMITH, L. M., SANDERS, J. Z., KAISER, R. J., et al. "Fluorescence detection in automated DNA sequence analysis", 1986. 2.3.1

[75] KIRCHER, M., KELSO, J. "High-throughput DNA sequencing–concepts and limitations", *Bioessays*, v. 32, n. 6, pp. 524–536, 2010. 2.3.1

[76] KUMAR, B. R. *DNA Sequencing - Methods and Applications*. InTech, 2012. 2.3.1

[77] TREANGEN, T. J., SALZBERG, S. L. "Repetitive DNA and next-generation sequencing: computational challenges and solutions", *Nature Reviews Genetics*, v. 13, n. 1, pp. 36–46, 2012. 2.3.1

[78] ANSORGE, W. J. "Next-generation DNA sequencing techniques", *New biotechnology*, v. 25, n. 4, pp. 195–203, 2009. 2.3.2, 2.3.2

[79] REIS-FILHO, J. S., OTHERS. "Next-generation sequencing", *Breast Cancer Res*, v. 11, n. Suppl 3, pp. S12, 2009. 2.3.2

[80] GOODWIN, S., MCPHERSON, J. D., MCCOMBIE, W. R. "Coming of age: ten years of next-generation sequencing technologies", *Nature Reviews Genetics*, v. 17, n. 6, pp. 333–351, 2016. 2.3.2

[81] MARDIS, E. R. "Next-generation sequencing platforms", *Annual review of analytical chemistry*, v. 6, pp. 287–303, 2013. 2.3.2

[82] BERGLUND, E. C., KIIALAINEN, A., SYVÄNEN, A.-C. "Next-generation sequencing technologies and applications for human genetic history and forensics", *Investigative Genetics*, v. 2, n. 1, pp. 23, 2011. 2.3.2

[83] GRADA, A., WEINBRECHT, K. "Next-generation sequencing: methodology and application", *Journal of Investigative Dermatology*, v. 133, n. 8, pp. 1–4, 2013. 2.3.2

[84] GOGOL-DÖRING, A., CHEN, W. "An overview of the analysis of next generation sequencing data", *Next Generation Microarray Bioinformatics: Methods and Protocols*, pp. 249–257, 2012. 2.3.2

[85] MACLEAN, D., JONES, J. D., STUDHOLME, D. J. "Application of'next-generation'sequencing technologies to microbial genetics", *Nature Reviews Microbiology*, v. 7, n. 4, pp. 287–296, 2009. 2.3.2

[86] BARZON, L., LAVEZZO, E., MILITELLO, V., et al. "Applications of next-generation sequencing technologies to diagnostic virology", *International journal of molecular sciences*, v. 12, n. 11, pp. 7861–7884, 2011. 2.3.2

[87] TANG, P., CHIU, C. "Metagenomics for the discovery of novel human viruses", *Future microbiology*, v. 5, n. 2, pp. 177–189, 2010. 2.3.2

[88] BRUSELLES, A., ROZERA, G., BARTOLINI, B., et al. "Use of massive parallel pyrosequencing for near full-length characterization of a unique HIV Type 1 BF recombinant associated with a fatal primary infection", *AIDS research and human retroviruses*, v. 25, n. 9, pp. 937–942, 2009. 2.3.2

[89] EBERLE, J., GÜRTLER, L. "The evolution of drug resistance interpretation algorithms: ANRS, REGA and extension of resistance analysis to HIV-1 group O and HIV-2", *Intervirology*, v. 55, n. 2, pp. 128–133, 2012. 2.4

[90] CLEMEN, R. T. "Combining forecasts: A review and annotated bibliography", *International journal of forecasting*, v. 5, n. 4, pp. 559–583, 1989. 2.5

[91] WOLPERT, D. H. "Stacked generalization", *Neural networks*, v. 5, n. 2, pp. 241–259, 1992. 2.5, 2.5.3

[92] BREIMAN, L. "Stacked regressions", *Machine learning*, v. 24, n. 1, pp. 49–64, 1996. 2.5, 2.5

[93] DIETTERICH, T. G. "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization", *Machine learning*, v. 40, n. 2, pp. 139–157, 2000. 1

[94] MERZ, C. J. "Using correspondence analysis to combine classifiers", *Machine Learning*, v. 36, n. 1-2, pp. 33–58, 1999. 1

[95] BREIMAN, L. "Bagging predictors", *Machine learning*, v. 24, n. 2, pp. 123–140, 1996. 2.5, 2.5.1, 2.5.1

[96] SCHAPIRE, R. E. "The strength of weak learnability", *Machine learning*, v. 5, n. 2, pp. 197–227, 1990. 2.5, 2.5.2

[97] FREUND, Y. "Boosting a weak learning algorithm by majority", *Information and computation*, v. 121, n. 2, pp. 256–285, 1995. 2.5.2

[98] FREUND, Y., SCHAPIRE, R. E., OTHERS. "Experiments with a new boosting algorithm". In: *ICML*, v. 96, pp. 148–156, 1996. 2.5

[99] EFRON, B., TIBSHIRANI, R. J. *An introduction to the bootstrap*. CRC press, 1994. 2.5.1

[100] QUINLAN, J. R. "Bagging, boosting, and C4. 5". In: *AAAI/IAAI, Vol. 1*, pp. 725–730, 1996. 2.5.1, 2.5.2

[101] BAUER, E., KOHAVI, R. "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants", *Machine learning*, v. 36, n. 1-2, pp. 105–139, 1999. 2.5.2

[102] WITTEN, I. H., FRANK, E., HALL, M. A., et al. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016. 2.5.3

[103] COZZI-LEPRI, A., NOGUERA-JULIAN, M., DI GIALLONARDO, F., et al. "Low-frequency drug-resistant HIV-1 and risk of virological failure to first-line NNRTI-based ART: a multicohort European case–control study using centralized ultrasensitive 454 pyrosequencing", *Journal of Antimicrobial Chemotherapy*, p. dku426, 2014. 3

[104] POP, M., SALZBERG, S. L. "Bioinformatics challenges of new sequencing technology", *Trends in Genetics*, v. 24, n. 3, pp. 142–149, 2008. 3

[105] ARCHER, J., RAMBAUT, A., TAILLON, B. E., et al. "The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time - an ultra-deep approach", *PLoS computational biology*, v. 6, n. 12, 2010. 3

[106] GIBSON, R. M., MEYER, A. M., WINNER, D., et al. "Sensitive deep-sequencing-based HIV-1 genotyping assay to simultaneously determine susceptibility to protease, reverse transcriptase, integrase, and maturation inhibitors, as well as HIV-1 coreceptor tropism", *Antimicrobial agents and chemotherapy*, v. 58, n. 4, pp. 2167–2185, 2014. 3, 4.2.3, 6

[107] PROSPERI, M. C., SALEMI, M. "QuRe: software for viral quasispecies reconstruction from next-generation sequencing data", *Bioinformatics*, v. 28, n. 1, pp. 132–133, 2012. 3

[108] MACALALAD, A. R., ZODY, M. C., CHARLEBOIS, P., et al. "Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data", *PLoS Comput Biol*, v. 8, n. 3, pp. e1002417, 2012. 3, 6

[109] YANG, X., CHARLEBOIS, P., MACALALAD, A., et al. "V-Phaser 2: variant inference for viral populations", *BMC genomics*, v. 14, n. 1, pp. 674, 2013. 3

[110] WATSON, S. J., WELKERS, M. R., DEPLEDGE, D. P., et al. "Viral population analysis and minority-variant detection using short read next-generation sequencing", *Philosophical Transactions of the Royal Society B: Biological Sciences*, v. 368, n. 1614, pp. 20120205, 2013. 3

[111] RIEMENSCHNEIDER, M., HUMMEL, T., HEIDER, D. "SHIVA-a web application for drug resistance and tropism testing in HIV", *BMC bioinformatics*, v. 17, n. 1, pp. 314, 2016. 3

[112] LANGMEAD, B., SALZBERG, S. L. "Fast gapped-read alignment with Bowtie 2", *Nature methods*, v. 9, n. 4, pp. 357–359, 2012. 3

[113] ENNS, E., LIANG, B., JI, H., et al. "HyDRA – A novel bioinformatics tool for next generation sequencing-based HIV drug resistance data analysis". In: *25e Congrès annuel canadien de recherche sur le VIH/sida*, 2016. 3, 6

[114] LI, H. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM", *arXiv preprint arXiv:1303.3997*, 2013. 3

[115] DEPRISTO, M. A., BANKS, E., POPLIN, R., et al. "A framework for variation discovery and genotyping using next-generation DNA sequencing data", *Nature genetics*, v. 43, n. 5, pp. 491–498, 2011. 3

[116] WILM, A., AW, P. P. K., BERTRAND, D., et al. "LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets", *Nucleic acids research*, p. gks918, 2012. 3

[117] FRENTZ, D., BOUCHER, C., ASSEL, M., et al. "Comparison of HIV-1 genotypic resistance test interpretation systems in predicting virological outcomes over time", *PloS one*, v. 5, n. 7, pp. e11505, 2010. 3

[118] ZHANG, J., RHEE, S.-Y., TAYLOR, J., et al. "Comparison of the precision and sensitivity of the Antivirogram and PhenoSense HIV drug susceptibility assays", *Journal of acquired immune deficiency syndromes (1999)*, v. 38, n. 4, pp. 439, 2005. 4.1.1

[119] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. Disponível em: <`https://www.R-project.org/`>. 4.1.1.1

[120] GARCIA, E., HE, H. "Learning from Imbalanced Data", *IEEE Transactions on Knowledge & Data Engineering*, v. 21, pp. 1263–1284, 2008. ISSN: 1041-4347. doi: doi.ieeecomputersociety.org/10.1109/TKDE.2008.239. 4.1.1.3

[121] LUNARDON, N., MENARDI, G., TORELLI, N. "ROSE: a Package for Binary Imbalanced Learning", *R Journal*, v. 6, n. 1, pp. 82–92, 2014. 4.1.1.3

[122] RISH, I. "An empirical study of the naive Bayes classifier". In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*, v. 3, pp. 41–46. IBM New York, 2001. 4.1.2

[123] WEIHS, C., LIGGES, U., LUEBKE, K., et al. "klaR Analyzing German Business Cycles". In: Baier, D., Decker, R., Schmidt-Thieme, L. (Eds.), *Data Analysis and Decision Support*, pp. 335–343, Berlin, 2005. Springer-Verlag. 4.1.2

[124] COVER, T., HART, P. "Nearest neighbor pattern classification", *IEEE transactions on information theory*, v. 13, n. 1, pp. 21–27, 1967. 4.1.2

[125] WILSON, D. R., MARTINEZ, T. R. "Improved heterogeneous distance functions", *Journal of artificial intelligence research*, v. 6, pp. 1–34, 1997. 4.1.2

[126] HECHENBICHLER, K., SCHLIEP, K. "Weighted k-nearest-neighbor techniques and ordinal classification", 2004. 4.1.2

[127] SCHLIEP, K., HECHENBICHLER, K., LIZEE, A. "kknn: Weighted k-nearest neighbors", *R package version*, pp. 1–0, 2010. 4.1.2

[128] KUHN, M. "The caret Package". 2009. 4.1.2.1, 4.1.5

[129] KOHAVI, R., OTHERS. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Ijcai*, v. 14, pp. 1137–1145. Stanford, CA, 1995. 4.1.2.1

[130] PARZEN, E. "On Estimation of a Probability Density Function and Mode", *Ann. Math. Statist.*, v. 33, n. 3, pp. 1065–1076, 09 1962. doi: 10.1214/aoms/1177704472. Disponível em: <`http://dx.doi.org/10.1214/aoms/1177704472`>. 4.1.2.1

[131] GARCIA, S., HERRERA, F. "An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets"for all Pairwise Comparisons", *Journal of Machine Learning Research*, v. 9, n. Dec, pp. 2677–2694, 2008. 4.1.4

[132] DEMŠAR, J. "Statistical comparisons of classifiers over multiple data sets", *Journal of Machine learning research*, v. 7, n. Jan, pp. 1–30, 2006. 4.1.4, 4.1.4

[133] EISINGA, R., HESKES, T., PELZER, B., et al. "Exact p-values for pairwise comparison of Friedman rank sums, with application to comparing classifiers", *BMC bioinformatics*, v. 18, n. 1, pp. 68, 2017. 4.1.4, 4.1.4

[134] CALVO, B., SANTAFÉ RODRIGO, G. "scmamp: statistical comparison of multiple algorithms in multiple problems", *The R Journal, Vol. 8/1, Aug. 2016*, 2016. 4.1.4

[135] HOLM, S. "A simple sequentially rejective multiple test procedure", *Scandinavian journal of statistics*, pp. 65–70, 1979. 4.1.4

[136] ARCHER, J., WEBER, J., HENRY, K., et al. "Use of four next-generation sequencing platforms to determine HIV-1 coreceptor tropism", 2012. 4.2.1, 6

[137] FOX, E. J., REID-BAYLISS, K. S., EMOND, M. J., et al. "Accuracy of Next Generation Sequencing Platforms", *Next generation, sequencing & applications*, v. 1, 2014. 4.2.1

[138] SHAFER, R. W., JUNG, D. R., BETTS, B. J. "Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries", *Nature medicine*, v. 6, n. 11, pp. 1290–1292, 2000. 4.2.1

[139] RHEE, S.-Y., KANTOR, R., KATZENSTEIN, D. A., et al. "HIV-1 pol mutation frequency by subtype and treatment experience: extension of the HIVseq program to seven non-B subtypes", *AIDS (London, England)*, v. 20, n. 5, pp. 643, 2006. 4.2.1

[140] SIEVERT, C., PARMER, C., HOCKING, T., et al. "plotly: Create Interactive Web Graphics via plotly. js", *R package version*, v. 3, n. 0, 2016. 4.2.1

[141] CHANG, W., CHENG, J., ALLAIRE, J., et al. *shiny: Web Application Framework for R*, 2015. Disponível em: <`http://CRAN.R-project.org/package=shiny`>. R package version 0.12.2. 4.2.2

[142] COSTA, J. A. F., BITTENCOURT, V. G., DE SOUTO, M. C. "Aplicação de Multi-classificadores no Reconhecimento de Classes Estruturais de Proteínas". In: *Anais do Congresso Nacional de Matemática Aplicada e Computacional (CNMAC)*, 2005. 6

[143] TANWANI, A. K., AFRIDI, J., SHAFIQ, M. Z., et al. "Guidelines to select machine learning scheme for classification of biomedical datasets". In: *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pp. 128–139. Springer, 2009.

[144] NIEVES, J., SANTOS, I., BRINGAS, P. "Combination of machine-learning algorithms for fault prediction in high-precision foundries". In: *Database and Expert Systems Applications*, pp. 56–70. Springer, 2012.

[145] PUN, J., LAWRYSHYN, Y. "Improving credit card fraud detection using a meta-classification strategy", *International Journal of Computer Applications*, v. 56, n. 10, 2012.

[146] SU, Y., ZHANG, Y., JI, D., et al. "Ensemble learning for sentiment classification", *Lect. Notes Comput. Sci*, v. 7717, pp. 84–93, 2013.

[147] HU, Y.-J., LIN, S.-C., LIN, Y.-L., et al. "A meta-learning approach for B-cell conformational epitope prediction", *BMC bioinformatics*, v. 15, n. 1, pp. 378, 2014. 6

[148] JAPKOWICZ, N., OTHERS. "Learning from imbalanced data sets: a comparison of various strategies". In: *AAAI workshop on learning from imbalanced data sets*, v. 68, pp. 10–15. Menlo Park, CA, 2000. 6

[149] WEISS, G. M., PROVOST, F. "The effect of class distribution on classifier learning: an empirical study", *Rutgers Univ*, 2001.

[150] LEE, P. H. "Resampling methods improve the predictive power of modeling in class-imbalanced datasets", *International journal of environmental research and public health*, v. 11, n. 9, pp. 9776–9789, 2014. 6

[151] SNOECK, J., KANTOR, R., SHAFER, R. W., et al. "Discordances between interpretation algorithms for genotypic resistance to protease and reverse transcriptase inhibitors of human immunodeficiency virus are subtype dependent", *Antimicrobial agents and chemotherapy*, v. 50, n. 2, pp. 694–701, 2006. 6

[152] HE, H., GARCIA, E. A. "Learning from imbalanced data", *IEEE Transactions on knowledge and data engineering*, v. 21, n. 9, pp. 1263–1284, 2009. 6

[153] LI, J. Z., KURITZKES, D. R. "Clinical implications of HIV-1 minority variants", *Clinical infectious diseases*, v. 56, n. 11, pp. 1667–1674, 2013. 6

Anexo 1

Tabela com os hiperparâmetros definidos pela função *train* do pacote *caret* para os metaclassificadores *naïve Bayes* do pacote klaR e k-NN.

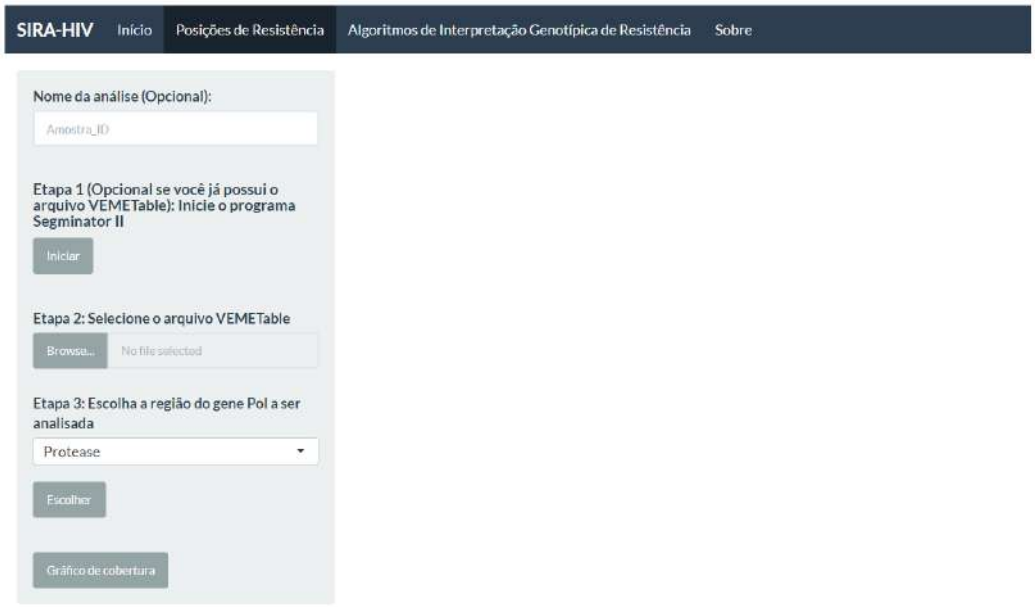
Classe	Antirretroviral	*Naïve Bayes*			k-NN		
		fL	*usekernel*	*adjust*	*k*	*distance*	*kernel*
PI	Atazanavir (ATV)	0	Sim	1	79	2	Ótimo
	Fosamprenavir (FPV)	0	Sim	1	89	2	Ótimo
	Indinavir (IDV)	0	Sim	1	9*	2	Ótimo
	Lopinavir (LPV)	0	Sim	1	45	2	Ótimo
	Nelfinavir (NFV)	0	Sim	1	67	2	Ótimo
	Saquinavir (SQV)	0	Sim	1	31	2	Ótimo
NRTI	Abacavir (ABC)	0	Sim	1	51	2	Ótimo
	Didanosina (ddI)	0	Sim	1	57	2	Ótimo
	Estavudina (d4T)	0	Sim	1	9*	2	Ótimo
	Lamivudina (3TC)	0	Sim	1	53	2	Ótimo
	Zidovudina (AZT)	0	Sim	1	59	2	Ótimo
NNRTI	Efavirenz (EFV)	0	Sim	1	65	2	Ótimo
	Nevirapina (NVP)	0	Sim	1	29	2	Ótimo

fL: correção de Laplace (presente ou não; o fator padrão é 0, ou seja, nenhuma correção); *usekernel*: estimativa de densidade de *kernel* (se sim, é usada para estimar a densidade, se não, a densidade normal é estimada); *adjust*: largura de banda; k: número de vizinhos; *distance*: valor de p da distância Minkowski; *kernel*: função *kernel* usada (O número de vizinhos usado para o kernel "ótimo" deve ser $[(2\,(d+4)\,/\,(d+2)) \wedge (d/\,(d+4))\,k]$, onde k é o número que seria usado para uma classificação não ponderada, ou seja, *kernel* = "retangular". Este fator $[(2\,(d+4)\,/\,(d+2)) \wedge (d/\,(d+4))\,k]$ está entre $1, 2$ e 2).*O desempenho do k-NN não mudou ao variar o número de vizinhos.

Anexo 2

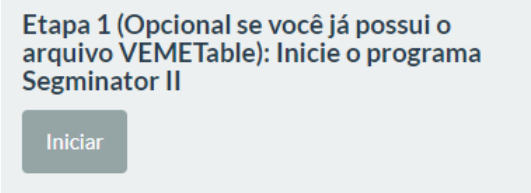
Neste item encontra-se uma descrição do uso do SIRA-HIV.

A análise dos dados no SIRA-HIV tem início na segunda tela do sistema.

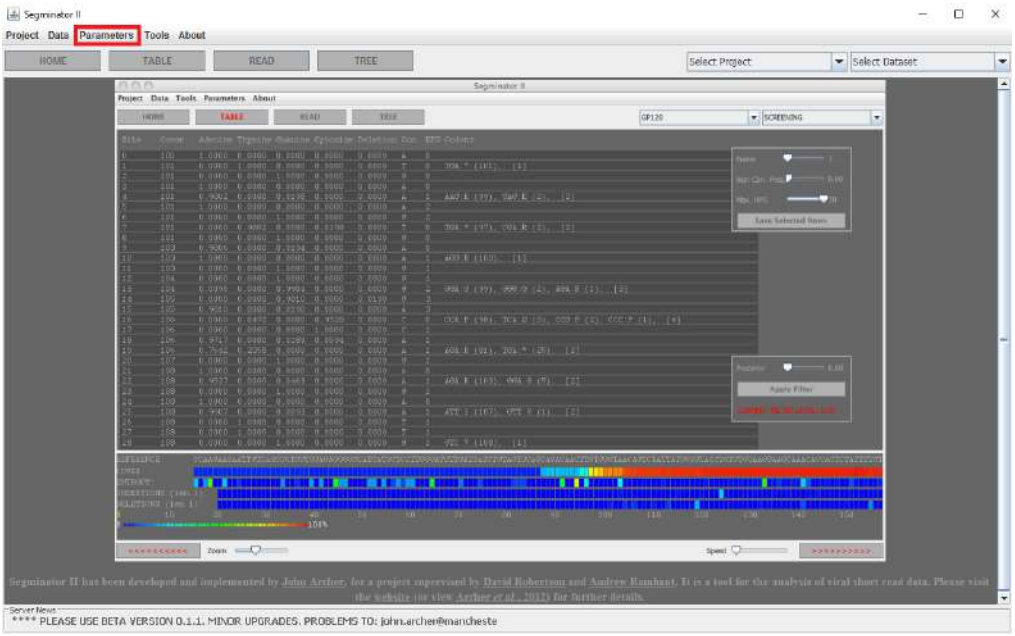


Inicialmente, o usuário tem a opção de atribuir um nome à sua análise na caixa de texto localizada no painel à esquerda. Esse nome será atribuído a todos os arquivos salvos pelo SIRA-HIV.

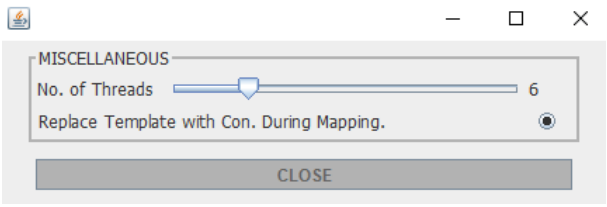
Na etapa 1, por meio do botão Iniciar, o usuário inicializa o programa Segminator II para realizar o mapeamento e alinhamento das sequências de nova geração.



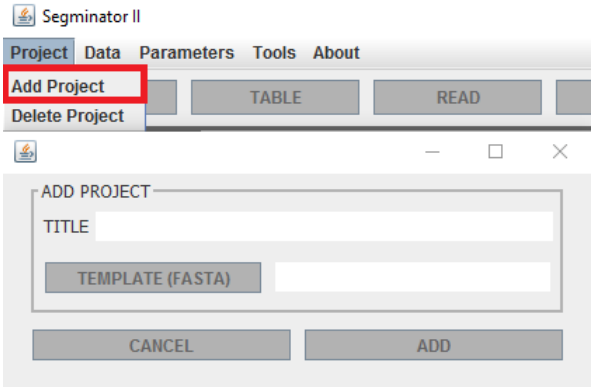
Antes da inserção do arquivo gerado pelo NGS, o ajuste de parâmetros de mapeamento pode ser realizado no Segminator II por meio do menu *Parameters*.



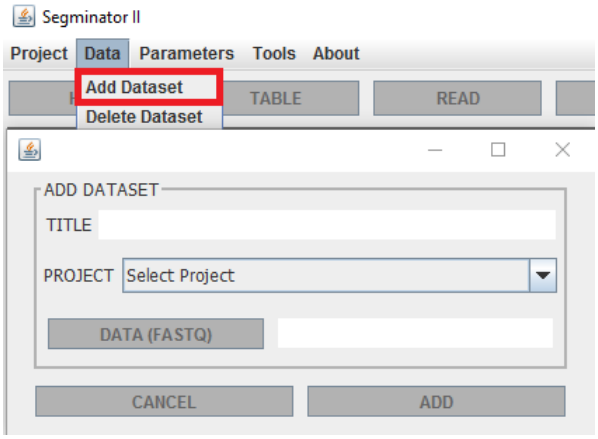
Recomenda-se a seleção da opção *Replace Template with Con. During Mapping*, em *Miscellaneous*. Essa função tem como objetivo remapear as leituras em relação a uma sequência consenso gerada no primeiro mapeamento, a fim de reduzir a diversidade entre elas.



Após as definições dos parâmetros, o usuário deve criar um projeto. A partir do menu *Project* → *Add project*, é necessário atribuir um nome (*Title*) ao projeto, que ficará salvo no Segminator II, e inserir a sequência de referência (*Template*) no formato FASTA.

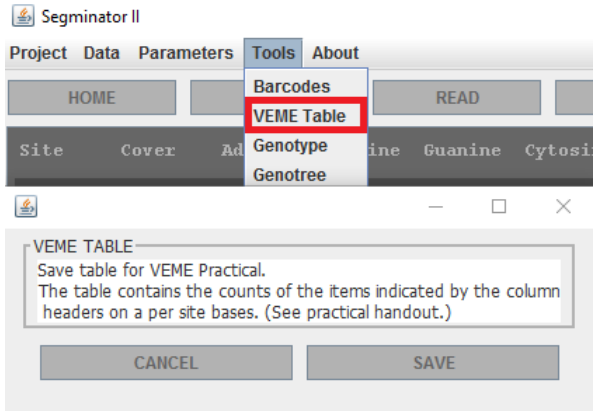


No menu *Data* → *Add dataset*, o usuário deve escolher um projeto já existente (*Project*), atribuir um nome (*Title*) ao conjunto de dados e inserir as leituras a serem mapeadas (*Data*) no formato FASTQ. O software realiza automaticamente o mapeamento dos dados.

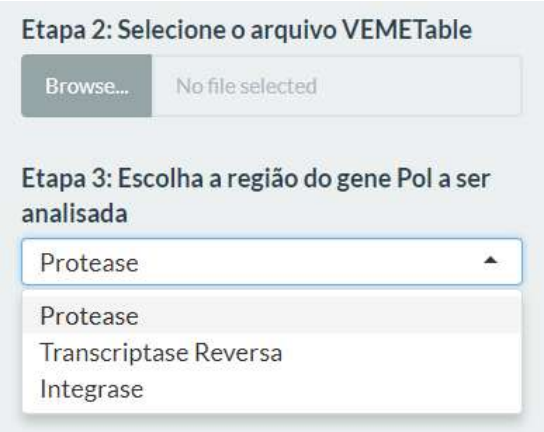


Posteriormente, o projeto criado e o conjunto de dados inserido devem ser selecionados nas opções *Select Project* e *Select Dataset* localizadas à direita do painel.

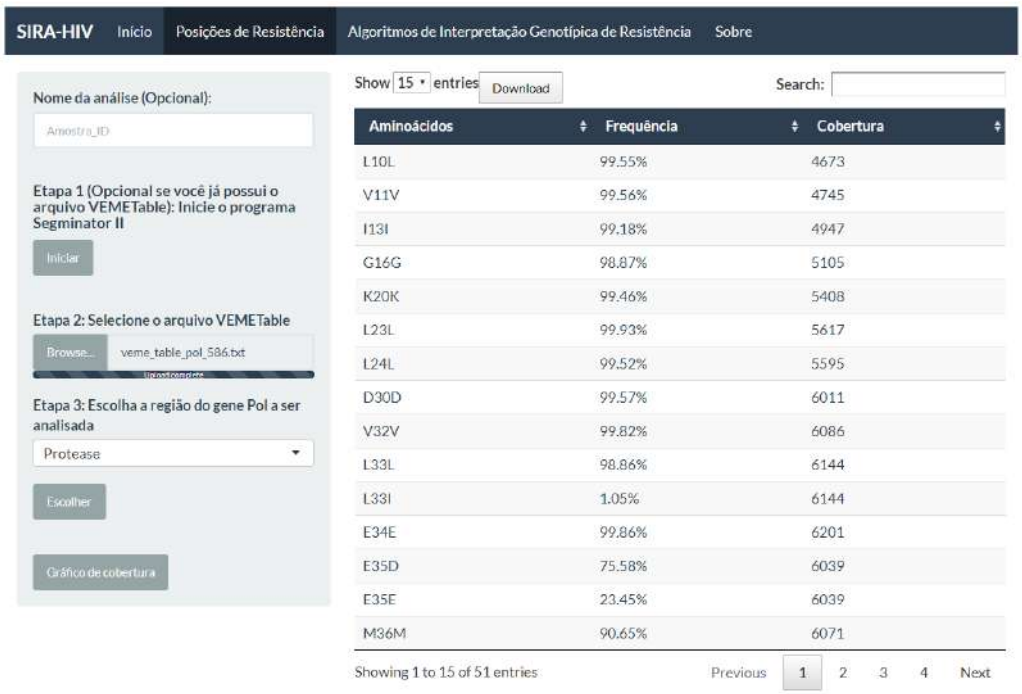
Os resultados são apresentados na tela do Segminator II e devem ser exportados por meio do arquivo VEMETable localizado no menu *Tools* → *VEME Table*. Caso o usuário já possua os arquivos VEMETable, a análise tem início na etapa 2.



Na etapa 2, o usuário deve inserir o arquivo VEMETable salvo e, na etapa 3, escolher a região do gene *pol* a ser analisada: PR, RT ou IN. Cada região é avaliada separadamente, de acordo com a opção selecionada.

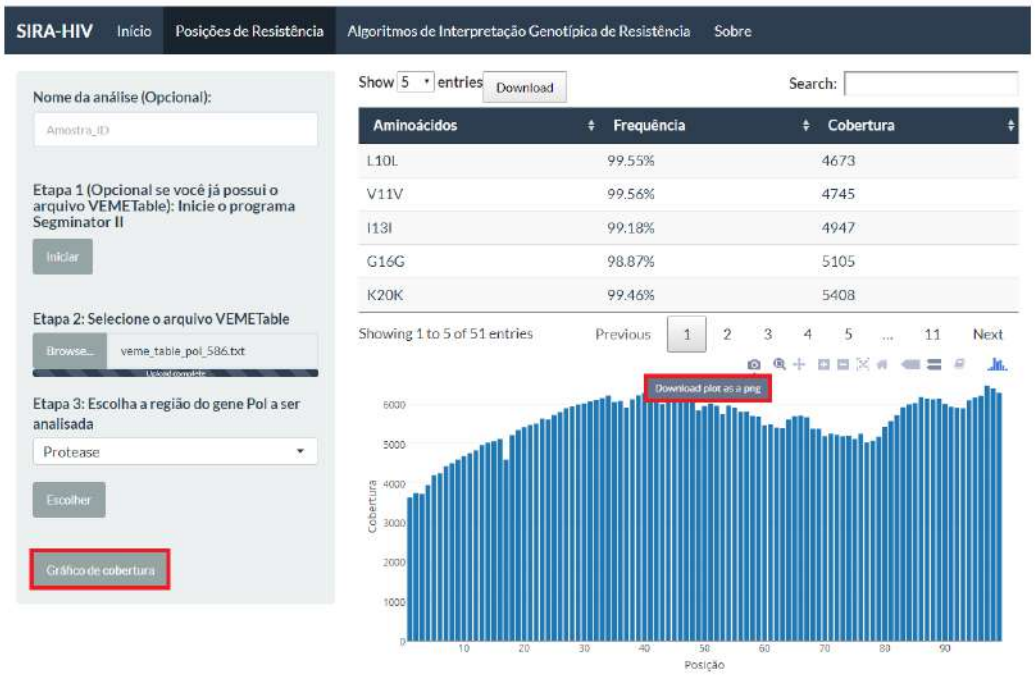


Ao clicar no botão Escolher, o sistema exibe na tela uma tabela contendo as posições de resistência com o aminoácido original e o presente nas sequências, as frequências de cada aminoácido e a cobertura de cada posição.



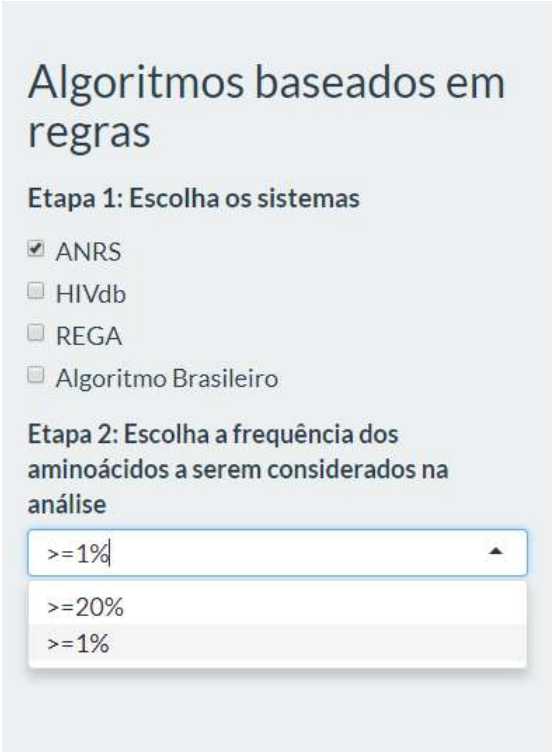
A tabela pode ser salva nos formatos .xls, .csv e .pdf. É importante que, antes de clicar no botão Download, a opção *Show* → *All* seja selecionada para que toda a tabela seja exibida na tela.

Por meio do botão Gráfico de cobertura, o usuário tem acesso à representação gráfica dos valores de cobertura por posição. O gráfico pode ser salvo no formato .png por meio do botão com formato de câmera fotográfica.



Na terceira tela, o usuário tem acesso ao módulo Algoritmos de Interpretação Genotípica.

Deve-se escolher os algoritmos (ANRS, HIVdb, Rega e/ou Algoritmo Brasileiro) e o nível de frequência dos aminoácidos ($\geq 20\%$ ou $\geq 1\%$) para classificar a amostra analisada.



Os níveis de suscetibilidade dos algoritmos selecionados são exibidos na tela em forma de tabela.

Os resultados da classificação de resistência também podem ser salvos nos formatos .xls, .csv e .pdf, por meio do botão Download.

Anexo 3

Este item apresenta os aminoácidos encontrados pelos programa DeepGen HIV e SIRA-HIV nas posições relacionadas à resistência para 10 indivíduos HIV+. Os valores sombreados de cinza correspondem às divergências encontradas entre os dois programas.

Posições da protease:

	Sistema	10L	11V	16G	20K	24L	30D
Indivíduo 1	DeepGen HIV	F (99,74%)	V (99,59%)	G (93,99%) E (5,61%)	R (99,35%)	L (99,98%)	D (99,79%)
	SIRA-HIV	F (99,39%)	V (99,59%)	G (93,73%) E (5,83%)	R (99,37%)	L (99,89%)	D (99,7%)
Indivíduo 2	DeepGen HIV	I (81,47%) L (17,97%)	V (99,71%)	G (98,74%) R (1,18%)	K (99,83%)	L (99,33%)	D (99,67%)
	SIRA-HIV	I (80,54%) L (18,6%)	V (99,53%)	G (98,96%)	K (99,86%)	L (99,57%)	D (99,59%)
Indivíduo 3	DeepGen HIV	L (60,4%) V (38,56%)	V (99,52%)	G (62,25%) E (37,25%)	K (62,12%) R (37,66%)	L (99,9%)	D (99,66%)
	SIRA-HIV	L (51,02%) V (47,71%)	V (99,29%)	G (57,48%) E (41,96%)	K (59,4%) R (40,19%)	L (99,87%)	D (99,59%)
Indivíduo 4	DeepGen HIV	I (99,63%)	V (98,74%)	G (90,07%) E (9,45%)	R (99,49%)	L (99,89%)	D (99,69%)
	SIRA-HIV	I (99,22%)	V (98,51%)	G (89,65%) E (9,74%)	R (99,44%)	L (99,83%)	D (99,66%)
Indivíduo 5	DeepGen HIV	F (99,64%)	V (97,5%) I (2,26%)	G (97,98%) R (1,96%)	T (99,65%)	L (99,78%)	D (99,53%)
	SIRA-HIV	F (99,4%)	V (97,47%) I (2,17%)	G (98,95%)	T (99,61%)	L (99,8%)	D (99,52%)
Indivíduo 6	DeepGen HIV	L (97,75%) I (1,67%)	V (99,79%)	G (98,93%) R (1,02%)	K (98,14%) I (1,64%)	L (99,9%)	D (99,74%)
	SIRA-HIV	L (97,7%) I (1,78%)	V (99,77%)	G (98,88%) R (1,05%)	K (98,03%) M (1,66%)	L (99,84%)	D (99,75%)
Indivíduo 7	DeepGen HIV	L (98,85%)	V (99,48%)	G (98,79%) R (1,19%)	K (99,57%)	L (99,96%)	D (99,6%)
	SIRA-HIV	L (99,13%)	V (99,37%)	G (98,66%) R (1,31%)	K (99,56%)	L (99,92%)	D (99,61%)
Indivíduo 8	DeepGen HIV	L (93,79%) P (5,81%)	V (99,86%)	G (99,46%)	K (99,69%)	L (99,96%)	D (99,76%)
	SIRA-HIV	L (98,93%)	V (99,67%)	G (99,64%)	K (99,58%)	L (99,86%)	D (99,74%)
Indivíduo 9	DeepGen HIV	L (97,05%) P (2,5%)	V (99,77%)	G (99,57%)	K (99,76%)	L (99,96%)	D (99,75%)
	SIRA-HIV	L (99,23%)	V (99,59%)	G (99,49%)	K (99,73%)	L (99,93%)	D (99,62%)
Indivíduo 10	DeepGen HIV	F (99,67%)	V (99,79%)	G (99,68%)	R (99,58%)	L (98,8%) I (1,1%)	D (99,3%)
	SIRA-HIV	F (99,52%)	V (99,67%)	G (99,61%)	R (99,6%)	L (98,8%) I (1,08%)	D (99,19%)

	Sistema	32V	33L	34E	36M	43K	46M
Indivíduo 1	DeepGen HIV	I (99,98%)	F (96,62%)	E (99,13%)	I (98,64%)	T (82,5%)	I (98,76%)
			L (3,29%)		L (1,21%)	K (16,03%)	L (1,01%)
						I (1,16%)	
	SIRA-HIV	I (99,91%)	F (97,76%)	E (99,32%)	I (98,67%)	T (84,39%)	I (98,46%)
			L (2,09%)		L (1,06%)	K (15,05%)	L (1,03%)
Indivíduo 2	DeepGen HIV	V (99,89%)	L (99,64%)	E (99,87%)	M (81,34%)	K (98,02%)	I (99,02%)
					V (17,99%)		
	SIRA-HIV	V (99,83%)	L (99,62%)	E (99,88%)	M (81,09%)	K (98,72%)	I (98,89%)
					V (17,79%)		
Indivíduo 3	DeepGen HIV	V (99,77%)	L (57,34%)	E (99,76%)	I (97,62%)	K (98,66%)	M (56,63%)
			F (42,39%)		K (1,99%)		I (42,73%)
	SIRA-HIV	V (99,8%)	L (59,37%)	E (99,74%)	I (99,01%)	K (99,21%)	M (57,12%)
			F (39,33%)				I (41,64%)
			S (1,18%)				
Indivíduo 4	DeepGen HIV	V (99,9%)	F (99,85%)	E (99,58%)	I (99,87%)	K (99,33%)	I (99,84%)
	SIRA-HIV	V (99,88%)	F (99,77%)	E (99,74%)	I (99,81%)	K (99,5%)	I (99,84%)
Indivíduo 5	DeepGen HIV	I (99,9%)	L (99,67%)	E (99,79%)	I (50,98%)	K (93,64%)	M (98,81%)
					L (34,11%)	N (5,18%)	
					M (14,43%)		
	SIRA-HIV	I (99,96%)	L (99,83%)	E (99,87%)	I (57,05%)	K (93,91%)	M (97,82%)
					L (33,03%)	N (5,53%)	
					M (9,41%)		
Indivíduo 6	DeepGen HIV	V (99,81%)	L (99,85%)	E (99,88%)	M (97,55%)	K (98,46%)	M (97,69%)
					I (2,26%)		I (2,1%)
	SIRA-HIV	V (99,85%)	L (99,94%)	E (99,84%)	M (97,65%)	K (98,93%)	M (97,48%)
					I (2,28%)		I (2,06%)
Indivíduo 7	DeepGen HIV	V (99,85%)	L (99,79%)	E (99,77%)	I (96,84%)	K (99,18%)	M (99,57%)
					K (2,44%)		
	SIRA-HIV	V (99,85%)	L (99,86%)	E (99,84%)	I (99%)	K (99,09%)	M (99,3%)
Indivíduo 8	DeepGen HIV	V (99,97%)	L (99,89%)	E (99,38%)	M (99,75%)	K (98,05%)	M (99,66%)
	SIRA-HIV	V (99,95%)	L (99,82%)	E (99,8%)	M (99,38%)	K (98,91%)	M (99%)
Indivíduo 9	DeepGen HIV	V (99,89%)	L (99,86%)	E (99,75%)	M (99,9%)	K (98,56%)	M (99,56%)
	SIRA-HIV	V (99,9%)	L (99,76%)	E (99,71%)	M (99,79%)	K (99,01%)	M (98,9%)
Indivíduo 10	DeepGen HIV	V (99,9%)	F (96,75%)	E (99,8%)	L (95,92%)	K (99,31%)	L (95,02%)
			L (3,06%)		I (3,41%)		I (4,74%)
	SIRA-HIV	V (99,85%)	F (96,99%)	E (99,72%)	L (96,19%)	K (99,48%)	L (95,32%)
			L (2,73%)		I (3,32%)		I (4,58%)

	Sistema	47I	48G	50I	53F	54I	58Q
Indivíduo 1	DeepGen HIV	V (98,34%)	G (98,29%)	I (99,62%)	F (99,8%)	L (99,55%)	Q (99,54%)
		G (1,42%)	R (1,47%)				
	SIRA-HIV	V (98,59%)	G (99,07%)	I (99,38%)	F (99,68%)	L (99,58%)	Q (99,83%)
Indivíduo 2	DeepGen HIV	I (97,87%)	G (98,41%)	I (99,31%)	F (99,52%)	V (99,48%)	Q (99,41%)
		S (1,06%)	R (1,46%)				
	SIRA-HIV	I (98,88%)	G (99,36%)	I (99,62%)	F (99,57%)	V (99,5%)	Q (99,66%)
Indivíduo 3	DeepGen HIV	I (98,79%)	G (99,85%)	I (62,43%)	F (99,29%)	I (99,84%)	Q (99,81%)
				V (34,8%)			
	SIRA-HIV	I (99,19%)	G (99,65%)	V (41,28%)	F (99,33%)	I (99,6%)	Q (99,8%)
				I (39,13%)			
				M (17,98%)			
Indivíduo 4	DeepGen HIV	I (98,64%)	G (98,59%)	V (99,68%)	L (94,45%)	I (99,87%)	Q (99,67%)
			R (1,36%)		F (5,44%)		
	SIRA-HIV	I (98,99%)	G (99,07%)	V (99,74%)	L (94,75%)	I (99,79%)	Q (99,84%)
					F (5,07%)		
Indivíduo 5	DeepGen HIV	V (81,15%)	G (99,58%)	I (99,02%)	F (98,29%)	I (99,68%)	E (99,41%)
		I (18,4%)			L (1,65%)		
	SIRA-HIV	V (81,78%)	G (99,55%)	I (99,36%)	F (98,62%)	I (99,23%)	E (99,6%)
		I (17,69%)			L (1,34%)		
Indivíduo 6	DeepGen HIV	I (99,4%)	G (99,86%)	I (99,24%)	F (99,91%)	I (99,94%)	Q (99,94%)
	SIRA-HIV	I (99,21%)	G (99,67%)	I (99,27%)	F (99,83%)	I (99,86%)	Q (99,57%)
Indivíduo 7	DeepGen HIV	I (99,45%)	G (99,73%)	I (99,07%)	F (99,75%)	I (99,86%)	Q (99,55%)
	SIRA-HIV	I (99,5%)	G (99,66%)	I (99,27%)	F (99,65%)	I (99,79%)	Q (99,73%)
Indivíduo 8	DeepGen HIV	I (99,32%)	G (99,7%)	I (99,38%)	F (99,59%)	I (99,6%)	Q (99,67%)
	SIRA-HIV	I (99,37%)	G (99,62%)	I (99,54%)	F (99,52%)	I (99,5%)	Q (99,81%)
Indivíduo 9	DeepGen HIV	I (99,39%)	G (99,76%)	I (99,45%)	F (99,57%)	I (99,76%)	Q (99,56%)
	SIRA-HIV	I (99,29%)	G (99,63%)	I (99,45%)	F (99,44%)	I (99,62%)	Q (99,75%)
Indivíduo 10	DeepGen HIV	I (99,87%)	G (99,68%)	V (99,25%)	F (99,21%)	I (96,55%)	Q (99,75%)
						V (3,17%)	
	SIRA-HIV	I (99,85%)	G (99,65%)	V (99,17%)	F (99,41%)	I (96,14%)	Q (99,76%)
						V (3,47%)	

	Sistema	60D	62I	63L	64I	69H	71A
Indivíduo 1	DeepGen HIV	D (99,31%)	V (99,43%)	T (87,67%)	I (97,25%)	H (98,38%)	V (99,39%)
				P (12,02%)	M (1,89%)	Y (1,2%)	
	SIRA-HIV	D (99,31%)	V (99,32%)	T (87,87%)	I (97,38%)	H (99,1%)	V (99,5%)
				P (11,92%)	M (1,66%)		
Indivíduo 2	DeepGen HIV	E (99,71%)	V (99,86%)	A (99,3%)	I (97,65%)	Q (99,05%)	A (99,59%)
					V (2,25%)		
	SIRA-HIV	E (99,66%)	V (99,83%)	A (99,26%)	I (97,68%)	Q (99,31%)	A (99,44%)
					V (2,18%)		
Indivíduo 3	DeepGen HIV	D (99,75%)	I (99,03%)	L (80,46%)	I (99,64%)	K (99,17%)	A (50,36%)
				V (18,88%)			V (48,68%)
	SIRA-HIV	D (99,68%)	I (98,95%)	L (81,16%)	I (99,58%)	K (99,16%)	A (52,6%)
				V (18,17%)			V (46,56%)
Indivíduo 4	DeepGen HIV	D (99,83%)	V (99,06%)	M (99,41%)	V (99,88%)	H (99,84%)	V (98,96%)
	SIRA-HIV	D (99,78%)	V (99,07%)	M (99,07%)	V (99,79%)	H (99,76%)	V (99%)
Indivíduo 5	DeepGen HIV	D (98,66%)	I (99,47%)	P (97,57%)	V (99,88%)	H (98,79%)	V (97,78%)
		E (1,1%)		L (2,32%)			A (1,73%)
	SIRA-HIV	D (98,47%)	I (99,38%)	P (97,65%)	V (99,83%)	H (99,58%)	V (97,85%)
		E (1,16%)		L (2,2%)			A (1,71%)
Indivíduo 6	DeepGen HIV	D (99,85%)	I (98,05%)	P (99,66%)	I (99,74%)	H (99,46%)	A (97,99%)
			V (1,8%)				V (1,74%)
	SIRA-HIV	D (99,73%)	I (97,87%)	P (99,68%)	I (99,58%)	H (99,62%)	A (97,82%)
			V (1,87%)				V (1,77%)
Indivíduo 7	DeepGen HIV	D (99,6%)	I (99,74%)	T (99,29%)	I (99,12%)	K (99,8%)	A (98,22%)
							T (1,62%)
	SIRA-HIV	D (99,49%)	I (99,55%)	T (99,2%)	I (99,13%)	K (99,71%)	A (98,37%)
							T (1,36%)
Indivíduo 8	DeepGen HIV	D (99,77%)	I (99,84%)	L (99,84%)	I (99,81%)	H (99,88%)	A (99,89%)
	SIRA-HIV	D (99,72%)	I (99,75%)	L (99,81%)	I (99,75%)	H (99,67%)	A (99,78%)
Indivíduo 9	DeepGen HIV	D (99,88%)	I (99,65%)	L (99,79%)	I (99,74%)	H (99,8%)	A (99,9%)
	SIRA-HIV	D (99,77%)	I (99,6%)	L (99,71%)	I (99,77%)	H (99,54%)	A (99,68%)
Indivíduo 10	DeepGen HIV	D (99,78%)	I (99,91%)	D (95,55%)	I (95,91%)	K (99,97%)	A (95,85%)
				E (3,7%)	L (3,94%)		V (3,96%)
	SIRA-HIV	D (99,68%)	I (99,83%)	D (95,49%)	I (95,82%)	K (99,97%)	A (95,48%)
				E (3,71%)	L (3,97%)		V (4,01%)

	Sistema	73G	74T	76L	77V	82V	83N
Indivíduo 1	DeepGen HIV	G (99,92%)	S (99,44%)	L (99,98%)	V (99,64%)	V (99,61%)	N (99,39%)
	SIRA-HIV	G (99,84%)	S (99,28%)	L (99,9%)	V (99,71%)	V (99,51%)	N (99,16%)
Indivíduo 2	DeepGen HIV	G (99,72%)	T (99,88%)	L (99,89%)	I (99,89%)	A (97,72%)	N (99,8%)
						V (2,21%)	
	SIRA-HIV	G (99,71%)	T (99,86%)	L (99,86%)	I (99,77%)	A (97,71%)	N (99,77%)
						V (2,18%)	
Indivíduo 3	DeepGen HIV	G (99,8%)	A (51,62%)	L (99,82%)	V (99,5%)	V (53,44%)	N (99,69%)
			S (34,03%)			A (46,14%)	
			T (14,28%)				
	SIRA-HIV	G (99,89%)	A (47,96%)	L (99,85%)	V (99,48%)	V (53,79%)	N (99,64%)
			S (37,16%)			A (45,83%)	
			T (14,69%)				
Indivíduo 4	DeepGen HIV	G (99,93%)	T (99,9%)	L (99,93%)	V (99,68%)	M (99,67%)	N (99,42%)
	SIRA-HIV	G (99,91%)	T (99,86%)	L (99,93%)	V (99,69%)	M (99,53%)	N (99,77%)
Indivíduo 5	DeepGen HIV	G (99,93%)	T (99,81%)	L (99,92%)	V (99,91%)	I (74,56%)	N (99,76%)
						L (25,27%)	
	SIRA-HIV	G (99,85%)	T (99,8%)	L (99,84%)	V (99,84%)	I (74,69%)	N (99,64%)
						L (25,1%)	
Indivíduo 6	DeepGen HIV	G (99,94%)	T (99,64%)	L (98,43%)	V (99,74%)	V (99,74%)	N (99,75%)
				V (1,53%)			
	SIRA-HIV	G (99,88%)	T (99,87%)	L (98,3%)	V (99,7%)	V (99,77%)	N (99,68%)
				V (1,65%)			
Indivíduo 7	DeepGen HIV	G (99,88%)	T (99,84%)	L (99,98%)	V (99,81%)	V (98,98%)	N (99,73%)
	SIRA-HIV	G (99,86%)	T (99,85%)	L (99,91%)	V (99,64%)	V (99,02%)	N (99,69%)
Indivíduo 8	DeepGen HIV	G (99,91%)	T (99,85%)	L (99,9%)	V (99,94%)	V (99,71%)	N (99,77%)
	SIRA-HIV	G (99,92%)	T (99,85%)	L (99,9%)	V (99,9%)	V (99,74%)	N (99,78%)
Indivíduo 9	DeepGen HIV	G (99,97%)	T (99,93%)	L (99,96%)	V (99,99%)	V (99,51%)	N (99,82%)
	SIRA-HIV	G (99,85%)	T (99,9%)	L (99,92%)	V (99,92%)	V (99,57%)	N (99,78%)
Indivíduo 10	DeepGen HIV	G (99,85%)	P (93,67%)	L (99,94%)	V (99,92%)	A (98,82%)	N (99,75%)
			T (4,66%)				
			A (1,57%)				
	SIRA-HIV	G (99,74%)	P (93,71%)	L (99,94%)	V (99,84%)	A (98,87%)	N (99,61%)
			T (4,69%)				
			A (1,51%)				

	Sistema	84I	85I	88N	89L	90L	93I
Indivíduo 1	DeepGen HIV	V (99,7%)	I (99,89%)	N (98,85%)	L (88,37%)	M (99,8%)	L (99,67%)
					V (11,29%)		
	SIRA-HIV	V (99,66%)	I (99,61%)	N (99,07%)	L (88,66%)	M (99,74%)	L (99,47%)
					V (11,05%)		
Indivíduo 2	DeepGen HIV	I (99,92%)	I (99,64%)	N (99,94%)	L (99,7%)	L (99,87%)	I (99,9%)
	SIRA-HIV	I (99,89%)	I (99,65%)	N (99,89%)	L (99,68%)	L (99,84%)	I (99,72%)
Indivíduo 3	DeepGen HIV	I (99,93%)	I (99,84%)	N (53,92%)	M (88,44%)	L (99,68%)	L (99,87%)
				G (45,61%)	L (11,35%)		
	SIRA-HIV	I (99,82%)	I (99,8%)	N (30,14%)	M (87,59%)	L (99,52%)	L (99,83%)
				S (24,79%)	L (11,34%)		
				D (24,67%)			
				G (20,29%)			
Indivíduo 4	DeepGen HIV	V (99,85%)	I (99,87%)	N (99,01%)	L (88,1%)	L (99,77%)	I (99,48%)
					F (11,26%)		
	SIRA-HIV	V (99,82%)	I (99,83%)	N (99,3%)	L (88,48%)	L (99,79%)	I (99,57%)
					F (10,83%)		
Indivíduo 5	DeepGen HIV	I (99,9%)	I (99,69%)	N (99,89%)	L (99,96%)	L (99,87%)	I (99,64%)
	SIRA-HIV	I (99,87%)	I (99,71%)	N (99,86%)	L (99,94%)	L (99,83%)	I (99,56%)
Indivíduo 6	DeepGen HIV	I (99,97%)	I (99,85%)	N (99,77%)	L (99,94%)	L (98,42%)	L (99,7%)
						M (1,42%)	
	SIRA-HIV	I (99,79%)	I (99,81%)	N (99,75%)	L (99,91%)	L (98,38%)	L (99,58%)
						M (1,44%)	
Indivíduo 7	DeepGen HIV	I (99,92%)	I (99,96%)	N (99,75%)	M (99,17%)	L (99,59%)	L (99,75%)
	SIRA-HIV	I (99,93%)	I (99,93%)	N (99,76%)	M (99,21%)	L (99,52%)	L (99,71%)
Indivíduo 8	DeepGen HIV	I (99,91%)	I (99,95%)	N (99,65%)	L (99,96%)	L (99,99%)	I (99,9%)
	SIRA-HIV	I (99,94%)	I (99,82%)	N (99,77%)	L (99,81%)	L (99,85%)	I (99,82%)
Indivíduo 9	DeepGen HIV	I (99,96%)	I (99,86%)	N (99,4%)	L (99,93%)	L (99,79%)	I (99,85%)
	SIRA-HIV	I (99,83%)	I (99,77%)	N (99,71%)	L (99,8%)	L (99,76%)	I (99,7%)
Indivíduo 10	DeepGen HIV	I (99,72%)	V (93,36%)	N (99,8%)	M (98,34%)	L (99,73%)	I (90,78%)
			I (6,62%)		I (1,07%)		L (9,15%)
	SIRA-HIV	I (99,47%)	V (93,46%)	N (99,69%)	M (98,4%)	L (99,69%)	I (90,94%)
			I (6,42%)		I (1,01%)		L (8,83%)

Posições da transcriptase reversa:

	Sistema	41M	62A	65K	67D	69T	70K
Indivíduo 1	DeepGen HIV	L (99,08%)	A (99,62%)	K (98,48%)	N (96,12%)	T (99,92%)	K (99,77%)
					D (3,54%)		
	SIRA-HIV	L (99,17%)	A (99,53%)	K (98,89%)	N (95,89%)	T (99,82%)	K (99,83%)
					D (3,74%)		
Indivíduo 2	DeepGen HIV	M (99,82%)	A (99,66%)	K (99,33%)	D (98,08%)	T (99,38%)	K (99,31%)
					N (1,66%)		
	SIRA-HIV	M (99,54%)	A (99,84%)	K (99,35%)	D (98,83%)	T (99,15%)	K (99,08%)
Indivíduo 3	DeepGen HIV	M (99,71%)	A (100%)	K (98,53%)	D (96,26%)	S (96,18%)	K (96,61%)
				R (1,47%)	N (3,48%)	N (3,56%)	R (3,39%)
	SIRA-HIV	M (99,48%)	A (100%)	K (98,02%)	D (96,6%)	S (92,75%)	K (96,46%)
				R (1,98%)	N (3,13%)	T (3,46%)	R (3,54%)
						Y (3,41%)	
Indivíduo 4	DeepGen HIV	L (98,62%)	A (99,77%)	K (99,54%)	N (99,28%)	T (99,75%)	K (99,76%)
	SIRA-HIV	L (99,02%)	A (99,35%)	K (99,55%)	N (99,54%)	T (99,54%)	K (99,32%)
Indivíduo 5	DeepGen HIV	M (99,51%)	A (99,8%)	K (98,86%)	D (98,44%)	T (99,71%)	K (99,71%)
					N (1,27%)		
	SIRA-HIV	M (99,42%)	A (99,03%)	K (99,35%)	D (97,92%)	T (99,53%)	K (99,07%)
Indivíduo 6	DeepGen HIV	M (97,22%)	A (99,48%)	K (99,32%)	D (98,09%)	T (99,8%)	K (99,68%)
		L (2,16%)			N (1,45%)		
	SIRA-HIV	M (97,44%)	A (99,58%)	K (99,46%)	D (98,51%)	T (99,81%)	K (99,51%)
		L (1,96%)					
Indivíduo 7	DeepGen HIV	M (99,35%)	A (99,79%)	R (96,96%)	D (99,67%)	T (100%)	K (99,89%)
				K (3,04%)			
	SIRA-HIV	M (99,45%)	A (100%)	R (98,79%)	D (99,36%)	T (99,89%)	K (99,9%)
				K (1,01%)			
Indivíduo 8	DeepGen HIV	M (99,61%)	A (99,65%)	K (99,5%)	D (97,62%)	T (99,2%)	K (99,79%)
					N (1,96%)		
	SIRA-HIV	M (99,68%)	A (99,54%)	K (99,72%)	D (98,11%)	T (99,15%)	K (99,41%)
					N (1,5%)		
Indivíduo 9	DeepGen HIV	M (99,75%)	A (99,54%)	K (99,49%)	D (96,8%)	T (98,69%)	K (99,92%)
					N (2,74%)		
	SIRA-HIV	M (99,57%)	A (99,44%)	K (99,79%)	D (97,19%)	T (98,78%)	K (99,79%)
					N (2,16%)		
Indivíduo 10	DeepGen HIV	M (99,72%)	A (72,78%)	K (99,25%)	G (98,14%)	T (99,76%)	E (98,24%)
			V (27,01%)		D (1,42%)		K (1,48%)
	SIRA-HIV	M (99,73%)	A (72,96%)	K (98,57%)	G (97,93%)	T (99,6%)	E (98,19%)
			V (26,6%)	R (1,4%)	D (1,97%)		K (1,58%)

	Sistema	74L	75V	77F	90V	98A	100L
Indivíduo 1	DeepGen HIV	L (99,56%)	V (99,4%)	F (99,95%)	V (99,86%)	A (99,58%)	L (98,47%)
	SIRA-HIV	L (99,73%)	V (99,43%)	F (99,74%)	V (99,76%)	A (99,41%)	L (98,72%)
Indivíduo 2	DeepGen HIV	L (94,75%)	V (99,91%)	F (99,48%)	V (99,69%)	A (97,84%)	L (95,51%)
		I (5,25%)				D (1,32%)	* (4,05%)
	SIRA-HIV	L (95,7%)	V (99,83%)	F (99,84%)	V (99,78%)	A (98,69%)	L (93,18%)
		I (4,11%)					* (5,01%)
Indivíduo 3	DeepGen HIV	L (100%)	V (100%)	F (100%)	V (98,89%)	A (95,67%)	L (99,18%)
						G (1,97%)	
						T (1,77%)	
	SIRA-HIV	L (100%)	V (100%)	F (100%)	V (99,15%)	A (95,61%)	L (96,23%)
						G (2,34%)	V (1,69%)
						T (1,61%)	F (1,49%)
Indivíduo 4	DeepGen HIV	L (99,75%)	M (100%)	F (100%)	V (99,79%)	G (99,62%)	L (95,13%)
							* (4,87%)
	SIRA-HIV	L (99,3%)	M (100%)	F (100%)	V (100%)	G (99,64%)	L (92,32%)
							* (5,46%)
							V (1,06%)
							F (1,03%)
Indivíduo 5	DeepGen HIV	L (99,9%)	V (99,8%)	F (99,9%)	V (99,92%)	A (99,59%)	L (98,64%)
	SIRA-HIV	L (99,72%)	V (99,72%)	F (99,91%)	V (99,85%)	A (99,28%)	L (98,97%)
Indivíduo 6	DeepGen HIV	L (99,94%)	V (99,75%)	F (99,82%)	V (99,84%)	A (99,85%)	L (99,65%)
	SIRA-HIV	L (99,94%)	V (99,82%)	F (99,59%)	V (99,59%)	A (99,61%)	L (99,29%)
Indivíduo 7	DeepGen HIV	I (73,62%)	V (100%)	F (99,89%)	V (100%)	A (98,3%)	L (99,89%)
		L (25%)					
		V (1,26%)					
	SIRA-HIV	I (70,07%)	V (99,57%)	F (99,89%)	V (99,8%)	A (98,18%)	L (99,59%)
		L (28,33%)					
		V (1,07%)					
Indivíduo 8	DeepGen HIV	L (99,73%)	V (99,95%)	F (99,89%)	V (99,85%)	A (99,29%)	L (98,87%)
	SIRA-HIV	L (99,65%)	V (99,75%)	F (99,8%)	V (99,82%)	A (99,31%)	L (96,72%)
							V (2,06%)
Indivíduo 9	DeepGen HIV	L (99,84%)	V (100%)	F (99,92%)	V (99,92%)	A (99,12%)	L (99,33%)
	SIRA-HIV	L (99,63%)	V (99,7%)	F (99,85%)	V (99,93%)	A (99,74%)	L (97,29%)
							V (1,74%)
Indivíduo 10	DeepGen HIV	L (99,79%)	V (99,78%)	F (99,72%)	V (99,76%)	A (96,98%)	L (97,9%)
						G (1,71%)	* (1,49%)
	SIRA-HIV	L (99,53%)	V (99,56%)	F (99,6%)	V (99,9%)	A (97,21%)	L (97,9%)
						G (1,66%)	* (1,46%)

*: códon de parada

	Sistema	101K	103K	106V	108V	115Y
Indivíduo 1	DeepGen HIV	K (99,98%)	K (99%)	V (99,94%)	V (99,89%)	Y (99,48%)
	SIRA-HIV	K (99,91%)	K (98,92%)	V (99,83%)	V (99,87%)	Y (99,24%)
Indivíduo 2	DeepGen HIV	K (99,93%)	K (99,29%)	V (99,93%)	I (57,64%)	Y (98,9%)
					V (42,3%)	
	SIRA-HIV	K (99,94%)	K (99,26%)	V (100%)	I (57,87%)	Y (97,66%)
					V (41,85%)	F (1,26%)
						* (1,07%)
Indivíduo 3	DeepGen HIV	K (93,8%)	K (98,34%)	V (100%)	V (99,81%)	Y (99,82%)
		G (6,2%)	E (1,45%)			
	SIRA-HIV	K (86,35%)	K (99,06%)	V (100%)	V (99,82%)	Y (98,6%)
		R (6,68%)				
		E (6,47%)				
Indivíduo 4	DeepGen HIV	K (99,81%)	K (98,67%)	V (99,81%)	V (99,81%)	Y (99,82%)
	SIRA-HIV	K (99,81%)	K (98,55%)	V (99,94%)	V (99,75%)	Y (99%)
			E (1,27%)			
Indivíduo 5	DeepGen HIV	K (99,94%)	K (98,6%)	V (100%)	V (99,74%)	Y (99,43%)
			E (1,2%)			
	SIRA-HIV	K (100%)	K (99,82%)	V (99,45%)	V (99,82%)	Y (99,47%)
Indivíduo 6	DeepGen HIV	K (100%)	N (94,48%)	V (99,8%)	V (99,86%)	Y (99,95%)
			K (4,49%)			
	SIRA-HIV	K (100%)	N (95,89%)	V (99,53%)	V (99,86%)	Y (99,86%)
Indivíduo 7	DeepGen HIV	E (98,62%)	K (98,08%)	V (99,78%)	V (99,79%)	F (97,8%)
		K (1,06%)	E (1,6%)			V (2,1%)
	SIRA-HIV	E (98,9%)	K (99,09%)	V (99,79%)	V (99,7%)	F (98,42%)
						V (1,3%)
Indivíduo 8	DeepGen HIV	K (99,14%)	K (98,6%)	V (99,81%)	V (99,91%)	Y (99,78%)
		N (0,86%)	E (1,17%)	I (0,14%)	A (0,09%)	L (0,18%)
	SIRA-HIV	K (99,33%)	K (98,74%)	V (99,96%)	V (99,82%)	Y (98,47%)
			E (1,01%)			
Indivíduo 9	DeepGen HIV	K (99,43%)	K (98,53%)	V (99,7%)	V (99,63%)	Y (99,71%)
			E (1,12%)			
	SIRA-HIV	K (99,43%)	K (98,73%)	V (99,87%)	V (99,8%)	Y (98,78%)
			E (1,21%)			
Indivíduo 10	DeepGen HIV	K (99,81%)	K (99,03%)	V (99,76%)	V (73,96%)	F (93,72%)
					I (25,94%)	Y (6,15%)
	SIRA-HIV	K (99,85%)	K (99,49%)	V (99,84%)	V (73,72%)	F (93,93%)
					I (26,08%)	Y (5,85%)

*: códon de parada

	Sistema	116F	138E	151Q	179V	181Y
Indivíduo 1	DeepGen HIV	F (99,91%)	E (99,52%)	Q (99,74%)	V (99,96%)	Y (99,59%)
	SIRA-HIV	F (99,98%)	E (99,5%)	Q (99,59%)	V (99,59%)	Y (99,59%)
Indivíduo 2	DeepGen HIV	F (99,94%)	E (62,61%)	Q (99,88%)	V (99,79%)	Y (99,93%)
			A (37,22%)			
	SIRA-HIV	F (100%)	E (62,92%)	Q (99,83%)	V (99,33%)	Y (100%)
			A (36,33%)			
Indivíduo 3	DeepGen HIV	F (100%)	E (100%)	Q (99,64%)	V (94,63%)	Y (99,51%)
					I (5,37%)	
	SIRA-HIV	F (100%)	E (100%)	Q (99,66%)	V (95,01%)	Y (98,86%)
					I (4,99%)	
Indivíduo 4	DeepGen HIV	F (100%)	E (100%)	Q (99,82%)	V (99,75%)	Y (100%)
	SIRA-HIV	F (99,83%)	E (100%)	Q (99,31%)	V (100%)	Y (100%)
Indivíduo 5	DeepGen HIV	F (100%)	E (99,15%)	Q (99,94%)	V (99,67%)	Y (99,92%)
	SIRA-HIV	F (100%)	E (99,2%)	Q (99,82%)	V (99,54%)	Y (100%)
Indivíduo 6	DeepGen HIV	F (100%)	E (99,69%)	Q (99,95%)	V (96,06%)	Y (99,95%)
					I (3,79%)	
	SIRA-HIV	F (99,91%)	E (99,54%)	Q (99,91%)	V (95,99%)	Y (99,95%)
Indivíduo 7	DeepGen HIV	F (99,9%)	E (99,34%)	Q (99,9%)	V (100%)	C (100%)
	SIRA-HIV	F (99,91%)	E (99,64%)	Q (99,91%)	V (100%)	C (99,9%)
Indivíduo 8	DeepGen HIV	F (99,96%)	E (99,76%)	Q (99,75%)	V (99,8%)	C (67,49%)
						Y (32,46%)
	SIRA-HIV	F (99,92%)	E (99,7%)	Q (99,73%)	V (99,51%)	C (67,98%)
						Y (31,88%)
Indivíduo 9	DeepGen HIV	F (99,93%)	E (99,47%)	Q (99,66%)	V (99,84%)	C (63,79%)
						Y (35,98%)
	SIRA-HIV	F (99,94%)	E (99,53%)	Q (99,64%)	V (99,79%)	C (64,51%)
						Y (35,21%)
Indivíduo 10	DeepGen HIV	Y (98,37%)	E (99,39%)	M (99,8%)	V (99,96%)	I (55,07%)
						C (35,31%)
	SIRA-HIV	Y (98,4%)	E (99,36%)	M (99,47%)	V (99,68%)	I (33,55%)
						F (23,99%)
						S (19,17%)
						C (13,69%)
						N (3,21%)
						V (2,37%)
						Y (2,3%)
						G (1,34%)

	Sistema	184M	188Y	190G	210L	215T
Indivíduo 1	DeepGen HIV	V (99,83%)	Y (99,87%)	G (98,91%)	W (99,8%)	Y (99,29%)
	SIRA-HIV	V (99,78%)	Y (99,66%)	G (99,04%)	W (99,76%)	Y (98,99%)
Indivíduo 2	DeepGen HIV	V (99,93%)	Y (99,79%)	G (99,79%)	L (95,44%)	T (99,48%)
					I (3,77%)	
	SIRA-HIV	V (99,87%)	Y (99,74%)	G (99,73%)	L (95,94%)	T (99,43%)
					I (3,49%)	
Indivíduo 3	DeepGen HIV	M (92,14%)	Y (99,77%)	G (99,54%)	L (98,72%)	T (99,78%)
		V (7,62%)			W (1,28%)	
	SIRA-HIV	M (92,24%)	Y (100%)	G (99,56%)	L (98,97%)	T (99,59%)
		V (7,31%)			W (1,03%)	
Indivíduo 4	DeepGen HIV	V (99,2%)	Y (100%)	G (98,85%)	W (100%)	Y (99,23%)
				E (1,15%)		
	SIRA-HIV	V (98,99%)	Y (100%)	G (98,34%)	W (100%)	Y (100%)
				E (1,1%)		
Indivíduo 5	DeepGen HIV	V (99,92%)	Y (99,91%)	G (99,92%)	L (99,91%)	T (99,9%)
	SIRA-HIV	V (99,69%)	Y (99,36%)	G (99,84%)	L (99,91%)	T (99,82%)
Indivíduo 6	DeepGen HIV	M (96,19%)	Y (99,71%)	G (99,71%)	L (96,49%)	T (96,5%)
		V (3,56%)			W (3,12%)	Y (2,59%)
	SIRA-HIV	M (96,1%)	Y (99,77%)	G (99,72%)	L (96,46%)	T (94,3%)
		V (3,71%)			W (3,1%)	S (2,46%)
						N (2,4%)
Indivíduo 7	DeepGen HIV	V (100%)	Y (99,89%)	S (99,56%)	L (100%)	T (99,76%)
	SIRA-HIV	V (100%)	Y (100%)	S (99,59%)	L (100%)	T (99,45%)
Indivíduo 8	DeepGen HIV	M (99,62%)	Y (99,81%)	G (99,95%)	L (99,95%)	T (99,85%)
	SIRA-HIV	M (99,6%)	Y (99,69%)	G (99,82%)	L (99,91%)	T (99,86%)
Indivíduo 9	DeepGen HIV	M (99,85%)	Y (99,92%)	G (99,92%)	L (100%)	T (99,7%)
	SIRA-HIV	M (99,79%)	Y (99,93%)	G (99,93%)	L (99,93%)	T (99,65%)
Indivíduo 10	DeepGen HIV	V (99,85%)	Y (99,96%)	G (68,22%)	L (99,68%)	T (99,76%)
				A (31,66%)		
	SIRA-HIV	V (99,82%)	Y (99,85%)	G (68,57%)	L (99,58%)	T (99,74%)
				A (31,21%)		

	Sistema	219K	221H	225P	227F	230M
Indivíduo 1	DeepGen HIV	K (99,58%)	H (87,93%) Y (10,17%) N (1,69%)	P (99,76%)	F (99,77%)	M (99,67%)
	SIRA-HIV	K (99,65%)	H (87,38%) Y (11,16%) N (1,21%)	P (99,49%)	F (99,82%)	M (99,67%)
Indivíduo 2	DeepGen HIV	K (99,63%)	H (99,92%)	P (99,74%)	F (97,76%) L (2,16%)	M (99,91%)
	SIRA-HIV	K (99,36%)	H (99,7%)	P (99,19%)	F (97,62%) L (2,21%)	M (99,49%)
Indivíduo 3	DeepGen HIV	K (92,62%) Q (5,74%) R (1,43%)	H (99,59%)	P (98,98%)	F (99,33%)	M (99,8%)
	SIRA-HIV	K (92,32%) Q (5,76%) R (1,53%)	H (99,8%)	P (98,47%) T (1,15%)	F (98,82%)	M (99,43%)
Indivíduo 4	DeepGen HIV	K (99,64%)	H (98,54%) N (1,09%)	P (100%)	F (100%)	M (99,64%)
	SIRA-HIV	K (99,66%)	H (99,31%)	P (99,66%)	F (99,67%)	M (98,96%)
Indivíduo 5	DeepGen HIV	K (99,63%)	H (98,19%) N (1,71%)	P (99,58%)	F (99,38%)	M (99,78%)
	SIRA-HIV	K (99,82%)	H (98,7%)	P (99,61%)	F (99,62%)	M (99,1%)
Indivíduo 6	DeepGen HIV	K (99,71%)	H (99,39%)	P (99,95%)	F (99,51%)	M (99,67%)
	SIRA-HIV	K (99,86%)	H (99,76%)	P (99,9%)	F (99,67%)	M (99,77%)
Indivíduo 7	DeepGen HIV	K (99,89%)	H (100%)	P (99,65%)	F (99,17%)	M (99,89%)
	SIRA-HIV	K (99,89%)	H (100%)	P (99,67%)	F (99,45%)	M (99,35%)
Indivíduo 8	DeepGen HIV	K (99,77%)	H (99,76%)	P (98,52%)	F (98,48%) I (1,37%)	M (99,27%)
	SIRA-HIV	K (99,69%)	H (99,28%)	P (98,29%)	F (99,02%)	M (99,73%)
Indivíduo 9	DeepGen HIV	K (99,5%)	H (99,7%)	P (98,79%)	F (98,8%)	M (99,78%)
	SIRA-HIV	K (99,73%)	H (99,17%)	P (98,05%) L (1,18%)	F (98,9%)	M (99,86%)
Indivíduo 10	DeepGen HIV	K (99,88%)	H (96,67%) Y (3,29%)	P (99,96%)	F (99,84%)	M (99,76%)
	SIRA-HIV	K (99,78%)	H (96,33%) Y (3,33%)	P (99,73%)	F (99,74%)	M (99,81%)

Posições da integrase:

	Sistema	51H	66T	74L	92E	97T
Indivíduo 1	DeepGen HIV	H (99,72%)	T (99,92%)	L (99,77%)	E (99,76%)	T (99,68%)
	SIRA-HIV	H (99,67%)	T (99,93%)	L (99,78%)	E (99,85%)	T (99,7%)
Indivíduo 2	DeepGen HIV	H (99,85%)	T (99,84%)	L (100%)	E (99,83%)	T (99,83%)
	SIRA-HIV	H (99,44%)	T (99,85%)	L (100%)	E (99,84%)	T (99,84%)
Indivíduo 3	DeepGen HIV	H (100%)	T (100%)	L (100%)	E (99,69%)	T (99,38%)
	SIRA-HIV	H (100%)	T (100%)	L (100%)	E (98,57%)	T (99,11%)
Indivíduo 4	DeepGen HIV	H (100%)	T (100%)	M (82,79%) L (17,21%)	E (100%)	T (100%)
	SIRA-HIV	H (100%)	T (100%)	M (83,94%) L (16,06%)	E (100%)	T (99,16%)
Indivíduo 5	DeepGen HIV	H (99,84%)	T (99,44%)	L (99,82%)	E (99,81%)	T (99,81%)
	SIRA-HIV	H (99,69%)	T (98,79%)	L (99,83%)	E (99,81%)	T (99,82%)
Indivíduo 6	DeepGen HIV	H (99,72%)	T (99,69%)	L (99,85%)	E (99,92%)	T (99,84%)
	SIRA-HIV	H (99,59%)	T (99,63%)	L (99,86%)	E (99,92%)	T (99,69%)
Indivíduo 7	DeepGen HIV	H (99,81%)	T (99,78%)	L (99,56%)	E (100%)	T (99,76%)
	SIRA-HIV	H (99,47%)	T (100%)	L (99,79%)	E (100%)	T (99,77%)
Indivíduo 8	DeepGen HIV	H (99,73%)	T (99,82%)	L (100%)	E (99,94%)	T (99,82%)
	SIRA-HIV	H (99,65%)	T (99,89%)	L (99,79%)	E (99,83%)	T (99,18%)
Indivíduo 9	DeepGen HIV	H (99,46%)	T (100%)	L (99,12%)	E (99,61%)	T (99,61%)
	SIRA-HIV	H (99,5%)	T (100%)	L (99,12%)	E (99,55%)	T (99,18%)
Indivíduo 10	DeepGen HIV	H (99,87%)	T (99,93%)	L (100%)	E (99,55%)	T (99,85%)
	SIRA-HIV	H (99,69%)	T (99,93%)	L (99,86%)	E (99,78%)	T (99,79%)

	Sistema	101L	121F	140G	143Y
Indivíduo 1	DeepGen HIV	L (99,46%)	F (99,92%)	G (100%)	Y (99,76%)
	SIRA-HIV	L (99,26%)	F (99,92%)	G (100%)	Y (99,46%)
Indivíduo 2	DeepGen HIV	L (54,65%)	F (100%)	G (99,83%)	Y (99,84%)
		I (43,52%)			
		F (1,66%)			
	SIRA-HIV	L (53,96%)	F (99,71%)	G (99,53%)	Y (99,69%)
		I (44,56%)			
		F (1,15%)			
Indivíduo 3	DeepGen HIV	I (100%)	F (99,71%)	G (100%)	Y (100%)
	SIRA-HIV	I (100%)	F (99,72%)	G (100%)	Y (99,68%)
Indivíduo 4	DeepGen HIV	I (100%)	F (99,25%)	A (88,06%)	Y (92,2%)
				G (8,96%)	C (7,8%)
				C (2,99%)	
	SIRA-HIV	I (100%)	F (99,32%)	A (84,74%)	Y (91,72%)
				G (11,71%)	C (8,28%)
				S (3,12%)	
Indivíduo 5	DeepGen HIV	L (98,87%)	F (100%)	G (99,81%)	Y (99,82%)
	SIRA-HIV	L (99,47%)	F (100%)	G (99,82%)	Y (99,65%)
Indivíduo 6	DeepGen HIV	L (99,6%)	F (99,91%)	G (100%)	Y (99,71%)
	SIRA-HIV	L (99,7%)	F (99,84%)	G (99,91%)	Y (99,35%)
Indivíduo 7	DeepGen HIV	I (100%)	F (100%)	G (100%)	Y (99,79%)
	SIRA-HIV	I (100%)	F (99,8%)	G (100%)	Y (99,8%)
Indivíduo 8	DeepGen HIV	I (61,64%)	F (99,82%)	G (99,93%)	Y (99,82%)
		L (33,31%)			
		F (4,37%)			
	SIRA-HIV	I (66,63%)	F (99,83%)	G (99,88%)	Y (99,65%)
		L (32,31%)			
Indivíduo 9	DeepGen HIV	I (62,52%)	F (99,91%)	G (99,9%)	Y (99,9%)
		L (31,17%)			
		F (4,97%)			
	SIRA-HIV	I (70,18%)	F (99,73%)	G (99,62%)	Y (99,52%)
		L (28,81%)			
Indivíduo 10	DeepGen HIV	I (99,93%)	F (99,93%)	G (99,92%)	R (96,6%)
					Y (2,19%)
	SIRA-HIV	I (99,65%)	F (100%)	G (99,92%)	R (94,86%)
					H (2,66%)
					C (2,1%)

	Sistema	147S	148Q	153S	155N	193G
Indivíduo 1	DeepGen HIV	S (99,92%)	Q (98,32%)	S (100%)	N (99,73%)	G (99,76%)
			H (1,36%)			
	SIRA-HIV	S (99,54%)	Q (98,32%)	S (99,92%)	N (99,47%)	G (99,82%)
			H (1,3%)			
Indivíduo 2	DeepGen HIV	S (99,84%)	Q (100%)	S (100%)	N (99,64%)	G (100%)
	SIRA-HIV	S (100%)	Q (99,56%)	S (99,84%)	N (99,49%)	G (99,69%)
Indivíduo 3	DeepGen HIV	S (99,66%)	Q (100%)	S (100%)	N (100%)	G (100%)
	SIRA-HIV	S (99,36%)	Q (99,68%)	S (100%)	N (100%)	G (100%)
Indivíduo 4	DeepGen HIV	S (100%)	R (91,39%)	S (100%)	N (93,62%)	G (98,78%)
			Q (7,95%)		H (5,67%)	E (1,22%)
	SIRA-HIV	S (98,71%)	R (92,36%)	S (100%)	N (93,24%)	G (97,83%)
			Q (7,59%)		H (6,08%)	E (2,15%)
Indivíduo 5	DeepGen HIV	S (100%)	Q (99,82%)	S (99,8%)	N (99,38%)	G (100%)
	SIRA-HIV	S (99,64%)	Q (99,82%)	S (99,81%)	N (99,6%)	G (99,53%)
Indivíduo 6	DeepGen HIV	S (99,41%)	Q (100%)	S (99,89%)	N (99,33%)	G (100%)
	SIRA-HIV	S (99,81%)	Q (100%)	S (99,79%)	N (99,89%)	G (99,12%)
Indivíduo 7	DeepGen HIV	S (99,8%)	Q (99,79%)	S (99,78%)	N (99,77%)	G (100%)
	SIRA-HIV	S (99,81%)	Q (99,81%)	S (100%)	N (100%)	G (99,53%)
Indivíduo 8	DeepGen HIV	S (99,94%)	Q (99,94%)	S (100%)	N (99,81%)	G (98,06%)
						E (1,55%)
	SIRA-HIV	S (99,89%)	Q (99,89%)	S (99,7%)	N (99,64%)	G (98,54%)
						E (1,19%)
Indivíduo 9	DeepGen HIV	S (99,91%)	Q (99,91%)	S (100%)	N (99,71%)	G (99,39%)
	SIRA-HIV	S (99,39%)	Q (99,74%)	S (99,82%)	N (99,73%)	G (98,76%)
						E (1,03%)
Indivíduo 10	DeepGen HIV	S (99,92%)	Q (99,92%)	S (99,91%)	N (94,81%)	G (99,8%)
					H (4,82%)	
	SIRA-HIV	S (99,69%)	Q (99,69%)	S (99,91%)	N (95,28%)	G (99,49%)
					H (4,18%)	