AVALIAÇÃO DE PARÂMETROS DE DADOS PARA CONSTRUÇÃO DE MODELO CLASSIFICADOR DE TROPISMO DE HIV-1

José Fernando dos Anjos Rodrigues

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Biomédica, COPPE, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do grau de Mestre em Ciências em Engenharia Biomédica.

Orientadores: Flavio Fonseca Nobre

Leticia Martins Raposo

Rio de Janeiro

Dezembro de 2020

AVALIAÇÃO DE PARÂMETROS DE DADOS PARA CONSTRUÇÃO DE MODELO CLASSIFICADOR DE TROPISMO DE HIV-1

José Fernando dos Anjos Rodrigues

DISSERTAÇÃO 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 MESTRE EM CIÊNCIAS EM ENGENHARIA BIOMÉDICA.

Examinada por:

a) _____
Prof. Flavio Fonseca Nobre, Ph.D.

b) _____
Prof. Letícia Martins Raposo, D.Sc.

c) _____
Prof. Carlos Julio Tierra Criollo, D.Sc.

d) _____
Prof. Oswaldo Gonçalves Cruz, D.Sc.

RIO DE JANEIRO, RJ BRASIL
DEZEMBRO DE 2020

Rodrigues, José Fernando dos Anjos.

Avaliação de Parâmetros de Dados para Construção de Modelo classificador de Tropismo de HIV-1/ José Fernando dos Anjos Rodrigues. – Rio de Janeiro: UFRJ/COPPE, 2020.

xxii, 66 f.: il.; 29,7 cm.

Orientadores: Flavio Fonseca Nobre, Leticia Martins Raposo.

Dissertação (mestrado) – UFRJ/ COPPE/ Engenharia Biomédica, 2020.

Referências Bibliográficas: p. 49-56.

1. Machine Learning. 2. HIV-1. 3. Preparação de dados. I. Nobre, Flavio Fonseca et al.. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Biomédica. III. Título.

Agradecimentos

Agradeço a FAPERJ, CAPES e CNPq pelo apoio financeiro a mim e ao meu programa, sem o qual este trabalho não poderia ter sido realizado.

Agradeço a minha família, Raquel, Fernando e Ana, e meus amigos, pelo apoio tão necessário durante a realização deste trabalho.

Agradeço aos meus orientadores, Prof. Flavio e Profª Letícia, pela paciência, pelos ensinamentos e pelas discussões que geraram ideias concretizadas nesta dissertação.

Agradeço a Yuri, pelas horas despendidas, nas dependências do LESS ou fora delas, pela presença, física ou virtual, tão necessária nos momentos da realização de um trabalho tão solitário quanto a programação.

Agradeço a toda a comunidade do StackOverflow, pelas dúvidas sanadas durante a escrita do código que gerou este trabalho.

E, finalmente, agradeço a Alexandra Elbakyan, por auxiliar a mim, e ao mundo todo, ao abrir portas antes impenetráveis para buscar informações científicas.

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

AVALIAÇÃO DE PARÂMETROS DE DADOS PARA CONSTRUÇÃO DE MODELO CLASSIFICADOR DE TROPISMO DE HIV-1

José Fernando dos Anjos Rodrigues

Dezembro/2020

Orientadores: Flavio Fonseca Nobre
Leticia Martins Raposo

Programa: Engenharia Biomédica

A AIDS é uma doença de importância mundial, causada pelo vírus HIV-1. Dos vários subtipos existentes, os mais prevalentes são os subtipos B e C. Apesar de não ter cura, diversos medicamentos foram desenvolvidos ao longo do tempo para reduzir sua disseminação no organismo. Por exemplo, a administração de Maraviroque® exige determinar que o vírus possua tropismo pelo receptor CCR5. Existem testes fenotípicos altamente precisos, mas muito caros e pouco ágeis para uso na rotina clínica. Como alternativa, foram desenvolvidos modelos de inteligência artificial para determinar o tropismo ao observar a sequência de 35 aminoácidos da região V3 da gp120 do vírus. Os modelos enfrentam dificuldades em classificar corretamente vírus de tropismo não-R5. Neste trabalho, avaliamos etapas de seleção automática de variáveis e balanceamento de dados para o desempenho dos classificadores. Utilizou-se o algoritmo de *random forest* para desenvolver modelos treinados separadamente com 1.622 sequências do subtipo B e 560 sequências de subtipo C. Os modelos foram comparados com os classificadores geno2pheno e T-CUP 2.0, já estabelecidos. Para o subtipo B, a AUC de todos os modelos apresentou valores próximos a 0.95 e apresentaram paridade de desempenho com os preditores consagrados. Para o subtipo C, os modelos apresentaram AUC variantes, porém de desempenho superior aos classificadores consagrados. Os modelos apresentaram posições que, apesar de possuir pouca variabilidade, mostraram-se muito importantes para o modelo. Concluiu-se que o balanceamento de dados não trouxe melhoras e a seleção de variáveis é uma etapa desejável, porém deve ser realizada considerando informações anteriores obtidas empiricamente.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

EVALUATION OF DATA PARAMETERS FOR THE CONSTRUCTION OF A CLASSIFYING MODEL FOR TROPISM OF HIV-1

José Fernando dos Anjos Rodrigues

Dezembro/2020

Advisors: Flavio Fonseca Nobre
Leticia Martins Raposo

Department: Biomedical Engineering

AIDS is a disease of worldwide importance, caused by HIV-1. Of the many existent subtypes, the most prevalent are subtypes B and C. Though it has no cure, several medicines were developed over time in order to reduce its spreading in organism. For instance, Maraviroc® prescription demands to appoint whether the virus is CCR5-tropic. There are phenotypical tests which are highly accurate, but too expensive and no agile enough for clinical usage. As alternative, models of artificial intelligence were developed toward the definition of viral tropism by analyzing the 35 amino-acid sequence of the V3 loop of HIV-1's gp120. These models struggle in classifying correctly the non-R5 tropic viruses. In this work, we evaluate steps of automatic variable selection and data balancing for the classifiers performance. We utilized random forest algorithm to build up models separately trained with 1.622 subtype B sequences and 560 subtype C sequences. The models were compared with two existing classifiers: geno2pheno and T-CUP 2.0. For the subtype B, all models' AUC produced values close to 0.95 and performance parity with existing classifiers. For the subtype C, models produced fluctuating AUC values and superior performance, however. Our models presented variables that, despite of low variability, they turned up very importante for the model. We concluded that data balancing produced no gain in performance, and variable selection is a desirable step, however it must be executed regarding prior information obtained empirically.

Lista de abreviaturas e siglas

AIDS – Síndrome da Imunodeficiência Adquirida

AUC – Área Sob Curva

AZT – Azidotimidina

CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CCR5 – Receptor de Quimiocina C-C 5

CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico

CRF – Forma Recombinante Circulante

CXCR4 – Receptor de Quimiocina C-X-C 4

FAPERJ – Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro

HIV – Vírus da Imunodeficiência Humana

MCC – Coeficiente de Correlação de Matthews

NRTI – Inibidor Nucleosídico de Transcriptase Reversa

OOB – *Out-of-bag*

RF – *Random Forest*

ROC – Reciever Operation Characteristic

SVM – Máquina de Vetores de Suporte

TARV – Terapia Antirretroviral

T-CUP – Two-level Coreceptor Usage Prediction

URF – Forma Recombinante Única

“Saber programar é saber procurar

o que outros já programaram.”

-Fox, L.

“Em Deus, confiamos;

todos os outros devem trazer dados.”

-Deming, W.E.

Sumário

1 Introdução

O vírus da imunodeficiência humana tipo 1 (*Human Immunodeficiency Virus type 1*, HIV-1) é considerado o agente etiológico da Síndrome da Imunodeficiência Adquirida (*Acquired ImmunoDeficiency Syndrome*, Aids), doença documentada pela primeira vez no início da década de 1980, como resultado de diversos casos reportados de homens jovens vindo a falecer acometidos por infecções raras ou de fácil resolução.

Segundo a Organização das Nações Unidas, em 2018, 38 milhões de pessoas portavam o HIV, tendo 1,7 milhão de novos casos e 770 mil mortes relacionadas às complicações da Aids (UNAIDS, 2019). No Brasil, segundo o Ministério da Saúde, desde 1980, foram registrados 966 mil casos de infecção pelo HIV, com média anual de 39 mil novas detecções nos último cinco anos (BRASIL, 2019).

O vírus possui diversos subtipos, sendo o subtipo C o mais frequente no mundo, com incidência próxima a 50% dos casos de infecção. Na América do Norte e na Europa, a incidência é predominante do subtipo B (BBOSA, KALEEBU, *et al.*, 2019). No Brasil, apesar de o subtipo B ser associado à maioria das detecções, o subtipo C tem mostrado aumento importante, principalmente na região Sul do Brasil (ARRUDA, BOULLOSA, *et al.*, 2018).

O HIV-1 infecta os linfócitos T-CD4, célula fundamental para a coordenação do sistema imune. Para entrar nestas células, o vírus se liga ao receptor CD4 e a um segundo receptor para estabilizar a ligação. A escolha do correceptor depende da afinidade química entre as proteínas de vírus e o receptor. Esta propriedade é conhecida como tropismo viral (CANN, 2012, p. 115). Com o avanço das pesquisas sobre o HIV-1, descobriu-se que o tropismo do vírus depende de uma região hiper variante da proteína gp120, chamada alça V3, e que esta região pode mostrar afinidade entre dois receptores: CCR5 e CXCR4 (BERGER, DOMS, *et al.*, 1998).

A Aids não tem cura, porém há uma gama de medicamentos usados para reduzir e controlar a carga viral no paciente. Muitas vezes, o tratamento gera efeitos colaterais que levam o paciente ao abandono. Por isso, a busca por medicamentos tão efetivos quanto os anteriores, e com efeitos colaterais mais leves, perdura até hoje.

Um medicamento usado para controlar a proliferação do HIV-1 no organismo do paciente é um bloqueador do receptor CCR5, chamado Maraviroque®. Ele possui boa

efetividade e poucos efeitos adversos, mas é necessário que o vírus que está no organismo do indivíduo seja de tropismo para CCR5. Para determinar o tropismo, o teste fenotípico Trofile® foi desenvolvido. Entretanto, o teste possui diversos desafios logísticos e não pode ser utilizado clinicamente. Como alternativa, testes genotípicos começaram a ser desenvolvidos e, com o auxílio das tecnologias de sequenciamento e técnicas de modelagem estatística, vários grupos de pesquisa desenvolveram classificadores automáticos que analisam a composição da alça V3 (CARDOZO, KIMURA, *et al.*, 2007).

Diversos classificadores foram desenvolvidos, como o geno2pheno, o mais utilizado atualmente, e o T-CUP 2.0. Entretanto, pouco foi discutido na literatura se aspectos inerentes aos modelos, como perfil de balanceamento, tamanho da amostra ou quantidade de variáveis explicativas seriam efetivos. Também chama atenção o fato de que a maioria dos classificadores é treinada com conjuntos de sequências de vírus de vários subtipos, mas de maioria subtipo B.

Neste trabalho, avaliamos estes aspectos aplicados em modelos estatísticos, comparando com outros classificadores de tropismo já consagrados na literatura.

1.1 Objetivo

Avaliar parâmetros de tratamento de dados na aplicação de modelos de predição do tropismo de HIV-1 com base na sequência da alça V3.

1.1.1 Objetivos específicos

- Construir modelos de predição para o tropismo do HIV-1;
- Avaliar formas de amostragem mais adequadas para o conjunto de treinamento, em relação tanto ao balanceamento de observações quanto a remoção de variáveis;
- Avaliar a importância de posições para a predição do tropismo;
- Comparar modelos de predição de tropismo com classificadores já conhecidos.

2 Fundamentação teórica

2.1 HIV

2.1.1 Taxonomia

O HIV é um membro do gênero *Lentivirus* pertencente à família *Retroviridae* e subfamília *Orthoretroviridae*. Infecções causadas por lentivírus demonstram longo período de incubação, e extenso período de latência, antes da manifestação de seus principais sintomas (FANALES-BELASIO, RAIMONDO, *et al.*, 2010).

O HIV é dividido em dois tipos: HIV-1 e HIV-2, sendo o primeiro o mais disseminado, causador da maior parte dos casos de aids no mundo. Já o segundo é restrito a regiões da África Central e Ocidental (FANALES-BELASIO, RAIMONDO, *et al.*, 2010). O tipo 1 é dividido em 4 grupos: M, N, O e P. O grupo M é composto por nove subtipos (A, B, C, D, F, G, H, J e K), além de diversas formas recombinantes circulantes (*Circulating Recombinant Forms*, CRFs) e formas recombinantes únicas (*Unique Recombinant Forms*, URFs). Não há subgrupos descritos para os grupos N, O e P, dado o baixo número de casos de infecção por estes vírus (GAC BLOOD, 2016).

2.1.2 Estrutura

O genoma do HIV-1 é composto por duas cópias idênticas de fita única de RNA de aproximadamente 9,7 mil nucleotídeos divididos em nove genes que codificam dezenove proteínas. Segundo a classificação de Baltimore, o HIV se encaixa no Grupo VI: Vírus de RNA que usa DNA como intermediário no ciclo replicativo (BALTIMORE, 1971, FOLEY, KORBER, *et al.*, 2018).

Dos nove genes, os genes *gag, pol* e *env* são os principais responsáveis pela produção de proteínas estruturais. A Figura 2.1 esquematiza as principais estruturas do HIV-1: O gene *gag* codifica as proteínas do capsídeo, do nucleocapsídeo e a proteína matricial; o gene *pol* codifica as enzimas transcripstase reversa, protease e integrase; e o gene *env* produz duas glicoproteínas que, junto com a membrana lipídica, formam o

envelope utilizado para se ligar à célula: gp120 (de superfície) e gp41 (transmembrana) (FRANKEL, YOUNG, 1998, GAC BLOOD, 2016).

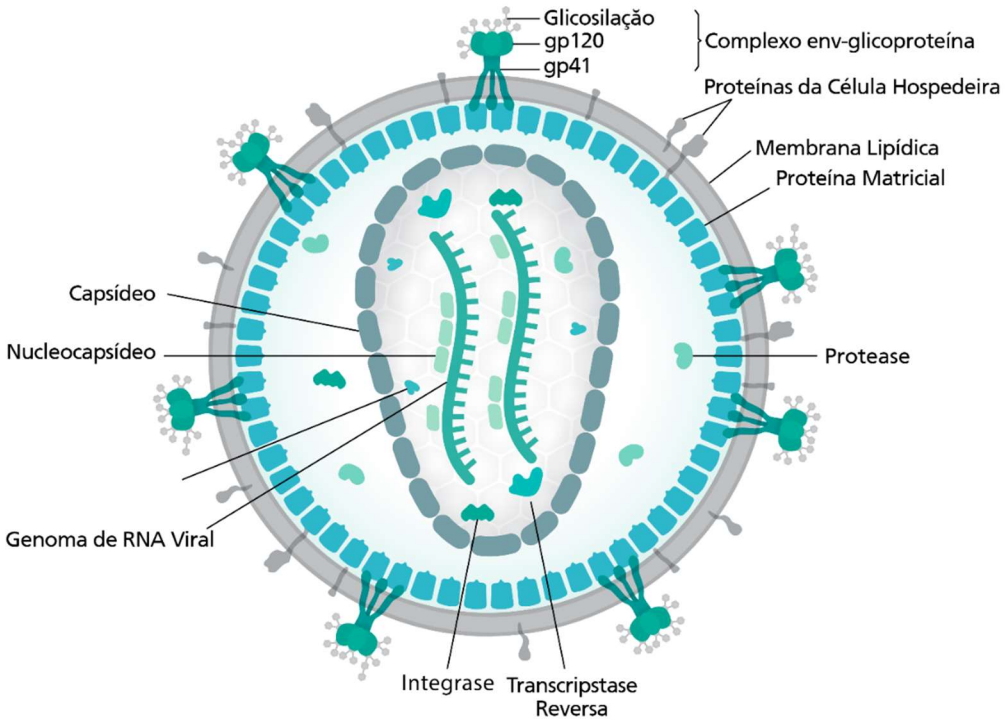

Figura 2.1: Esquema da estrutura do HIV-1 (Wikimedia, adaptado, acesso em 02/20).

2.1.3 Epidemiologia do HIV

Estima-se que, em todo o mundo, 38 milhões de pessoas portem o HIV. Em 2018, cerca de 1,7 milhão de novos casos e 770 mil óbitos relacionadas à AIDS foram reportados. A maior concentração de casos encontra-se na África, com cerca de 26 milhões de casos, seguidos da Ásia (5,9 milhões) e da América Latina (1,9 milhão) (UNAIDS, 2019). No Brasil, registrou-se cerca de 960 mil casos de AIDS desde 1980. Em 2018, a taxa de infecção foi de 38 mil novos casos anuais, com tendência de queda nos últimos cinco anos (BRASIL, 2019).

O subtipo mais frequente no mundo é o subtipo C, que abrange o sul e o leste da África, além da Índia, com mais de 50% dos casos registrados. Apesar disso, o registro de sequências genéticas pelo mundo tem predominância do vírus de subtipo B, mais presente na América do Norte e na Europa Ocidental (BBOSA, KALEEBU, *et al.*, 2019).

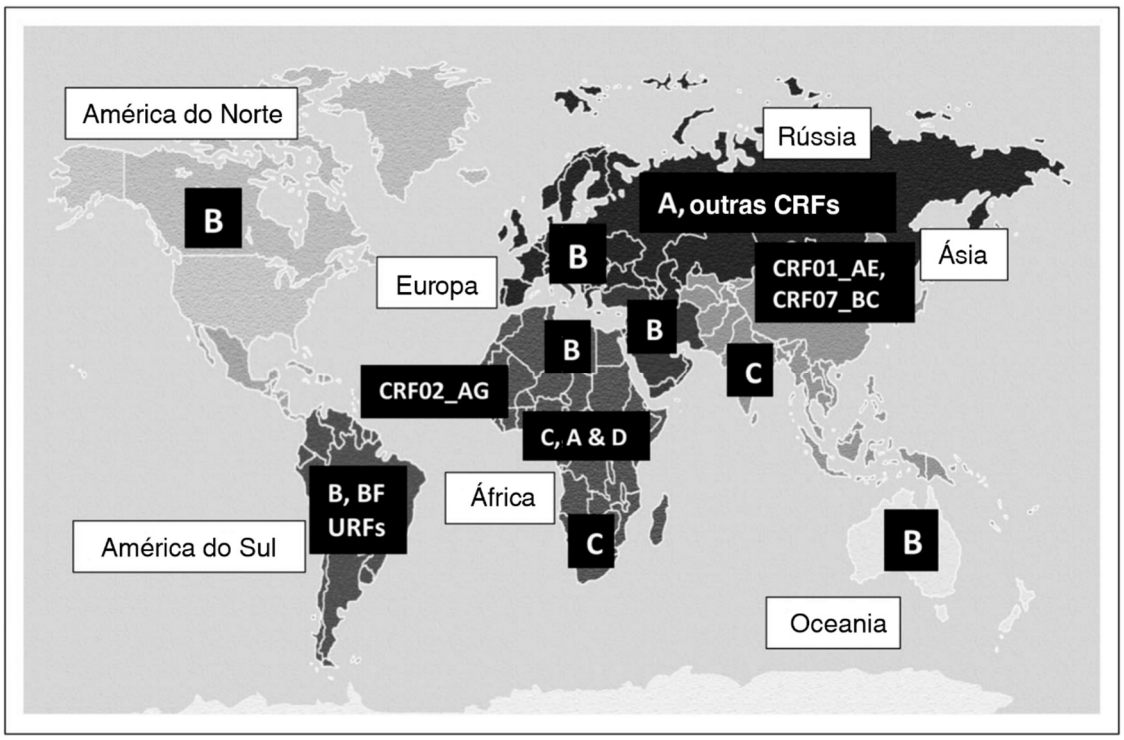

Figura 2.2 - Mapa de distribuição dos subtipos de HIV no mundo (BBOSA, KALEEBU, et al., 2019, adaptado. Uso de imagem autorizado pelo autor).

Na América do Sul, o subtipo B é o mais comum em circulação, seguido do subtipo F e uma recombinação dos dois subtipos, BF. No Brasil, os subtipos mais frequentes são os B e C, sendo a incidência do subtipo C maior na região Sul do país. Também é importante ressaltar que estudos recentes mostraram grandes proporções de casos envolvendo CRFs, entre 19 e 40% da prevalência no Brasil (ALVES, SIQUEIRA, *et al.*, 2019, ARRUDA, BOULLOSA, *et al.*, 2018, BBOSA, KALEEBU, *et al.*, 2019).

2.1.4 Infecção

A transmissão do HIV-1 se dá pelo contato sexual (contato com sêmen ou fluidos vaginais), transmissão vertical (da mãe para o filho, durante a gestação, parto ou amamentação), ou por via parenteral (transfusão de sangue, compartilhamento de agulhas ou por infusão de hemoderivados em hemofílicos) (MUSHAHWAR, 2006).

Em estágios iniciais da infecção, o HIV-1 pode infectar macrófagos e células dendríticas, além de linfócitos primários. Entretanto, sua célula-alvo são linfócitos T-CD4

(também chamados de linfócitos T-auxiliar), responsáveis pela coordenação do sistema imunológico em situações de infecção aguda (ABBAS, LICHTMAN, *et al.*, 2018, p. 478, LIN, KURITZKES, 2009).

Na Figura 2.3, encontra-se um esquema geral da infecção pelo HIV-1. O vírus infecta suas células-hospedeiro ligando-se pela gp120 com receptores celulares. No linfócito T-CD4, há a ligação com o receptor CD4 e um correceptor, que estabiliza a adsorção da gp41 e permite a fusão das membranas, dando prosseguimento à entrada da partícula viral na célula (FOUCHIER, GROENINK, *et al.*, 1992, IVANOFF, DUBAY, *et al.*, 1992, SATTENTAU, Quentin J, MOORE, 1993, YOON, FRIDKIS-HARELI, *et al.*, 2010).

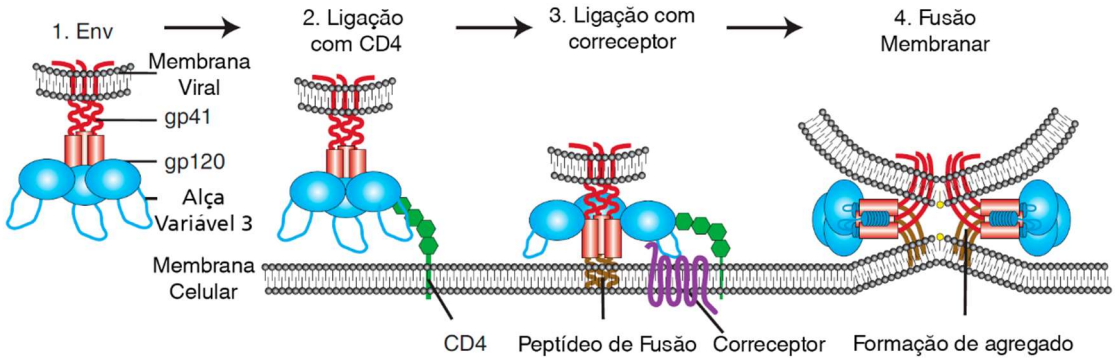

Figura 2.3: Esquema de interação entre as glicoproteínas do envelope do HIV-1 e os receptores do linfócito T-CD4 (WILEN, TILTON, et al., 2012, adaptado. Uso de imagem autorizado pelo autor).

2.1.5 Tropismo

Em virologia, o termo "tropismo" refere-se à tendência de um vírus a infectar uma célula ou outra para seguir seu ciclo replicativo. Na maioria dos casos, o tropismo é determinado pela presença de receptores específicos na célula-alvo (CANN, 2012, p. 115). O tropismo se determina pela composição das proteínas que compõem tanto o vírus quanto o receptor, expondo regiões de aminoácidos com propriedades diversas, como carga e hidrofobicidade, que se combinam em uma interação (CLAPHAM, MCKNIGHT, 2002).

Sabe-se que o tropismo do HIV está intimamente relacionado com uma região hipervariante da gp120, conhecida como alça variável 3 (alça V3). Esta região, composta por cerca de 35 aminoácidos, pode ter afinidade pelos receptores CCR5 ou CXCR4. Em

teoria, outros receptores também podem ser usados, mas apenas esses dois foram capazes de contribuir para a infecção do HIV-1 em sistemas *in vivo*. Portanto, classifica-se o tropismo do HIV-1 em três categorias: (i) R5, para uso do receptor CCR5, (ii) X4 para uso de CXCR4, e (iii) R5X4, quando o vírus é capaz de usar qualquer um dos receptores (IVANOFF, DUBAY, *et al.*, 1992, DRAGIC, LITWIN, *et al.*, 1996, SATTENTAU, Quentin J, MOORE, 1993, MOORE, TRKOLA, *et al.*, 1997, BERGER, DOMS, *et al.*, 1998, CLAPHAM, MCKNIGHT, 2001).

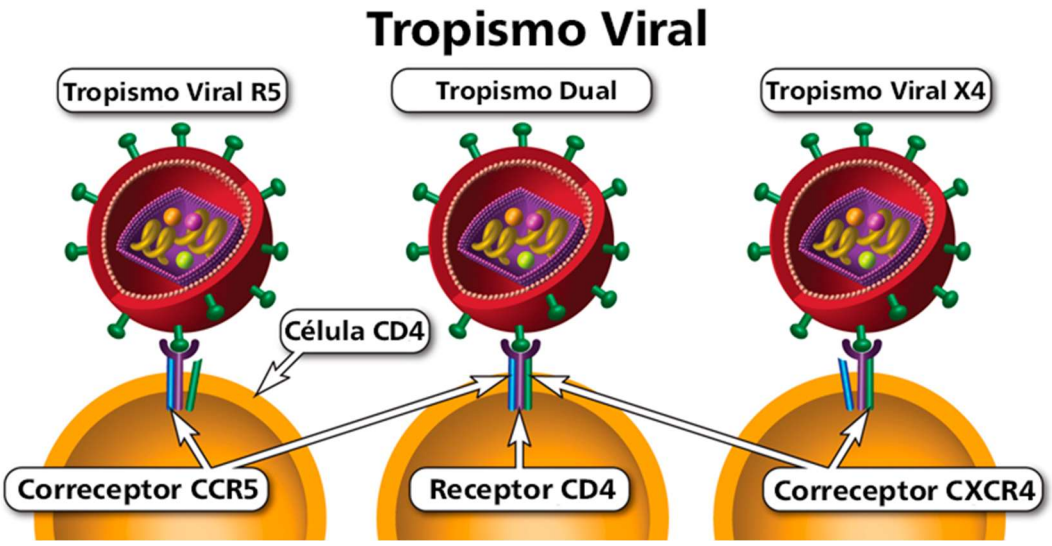

Figura 2.4: Ilustração do Tropismo Viral do HIV (Disponível em https://bit.ly/3dE5Imd*, acesso em 03/20, adaptado).*

2.1.6 Tratamento

Atualmente, não existe cura para a Aids. Entretanto, há a terapia antirretroviral (TARV), que visa o controle da replicação do vírus e a redução da carga viral no indivíduo. A terapia consiste na combinação de medicamentos antirretrovirais (ARV) que inibem a replicação do HIV-1 em diversas etapas, como na construção de proteínas, polimerização do código genético e montagem de partículas virais. Os ARVs podem ser administrados individualmente ou em combinações específicas, ditas “coquetéis antirretrovirais”.

A TARV é dividida em dois grupos: medicamentos de primeira e segunda linha. A terapia de primeira linha consiste em fármacos recomendados para a fase inicial da doença. Geralmente, estes são mais seguros, eficazes e convenientes para pacientes que nunca tomaram ARVs antes. Já os medicamentos de segunda linha (conhecidos também

como terapia de resgate) são introduzidos na TARV uma vez que os medicamentos de primeira linha percam sua eficácia em controlar a carga viral, ou que o vírus tenha adquirido resistência ao medicamento (HEALTH, SERVICES, *et al.*, 2009).

O primeiro fármaco utilizado para o combate do HIV foi a zidovudina (ou azidotimidina, AZT), um inibidor nucleosídico da transcriptase reversa (*Nucleoside Reverse Transcriptase Inhibitor*, NRTI). Substâncias desta classe visam interromper a transcrição reversa do código genético (Figura 2.5). Apesar da eficácia comprovada desde a década de 1980, diversos problemas envolvendo a alta toxicidade do medicamento vieram interferindo na condução do tratamento, inclusive levando ao seu abandono por parte do paciente (CHIU, DUESBERG, 1995, RICHMAN, FISCHL, *et al.*, 1987, STAMBUK, YOULE, *et al.*, 1989).

Atualmente, existem 22 medicamentos e 22 coquetéis aprovados pela agência de controle sanitário dos Estados Unidos (*Food and Drug Aministration*, FDA). Os medicamentos são divididos em seis classes: (1) inibidores nucleosídicos da transcriptase reversa, (2) inibidores não-nucleosídicos da transcripstase reversa, (3) inibidores de protease, (4) inibidores de fusão, (5) inibidores de integrase, (6) inibidores de entrada (HEALTH, SERVICES, *et al.*, 2009).

O principal representante da última categoria chama-se Maraviroque®. Este medicamento é capaz de alterar a estrutura física do receptor CCR5, impedindo a ligação da alça V3 à célula-alvo, sem causar grandes danos colaterais ao indivíduo (KANMOGNE, WOOLLARD, 2015, LIEBERMAN-BLUM, FUNG, *et al.*, 2008).

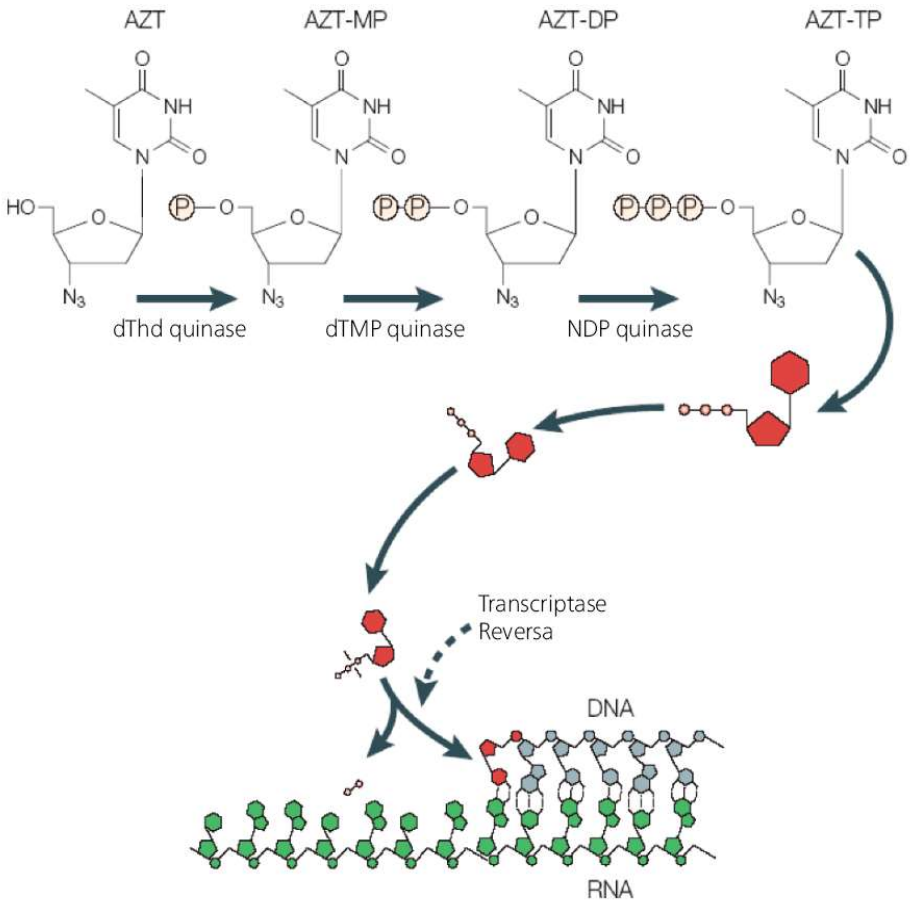

Figura 2.5 Esquema representando o mecanismo de ação da AZT (DE CLERCQ, NEYTS, 2009, adaptado. Uso de imagem autorizado pelo autor).

2.2 Métodos de Classificação de Tropismo do HIV-1

Para que o Maraviroque® seja introduzido no coquetel antirretroviral, é necessária a confirmação de que o indivíduo porte vírus que usem CCR5 como correceptor (BISWAS, TAMBUSSI, *et al.*, 2007). Existem duas formas principais de determinar o tropismo do HIV-1: ensaios fenotípicos ou genotípicos.

2.2.1 Trofile®

O teste fenotípico Trofile® é considerado o padrão-ouro para determinação do tropismo de HIV-1, sendo usado para triagem de pacientes soropositivos que participaram de estudos de novos medicamentos, tamanha é sua confiança que as principais bases de

dados do HIV-1 usam suas informações para catalogar as sequências (WHITCOMB, HUANG, *et al.*, 2007).

O ensaio consiste em replicar partículas virais manipuladas a partir do material genético do vírus extraído de um paciente. Em seguida, duas culturas celulares são expostas a estes vírus: uma delas é preparada para superexpressar o correceptor CCR5 e a outra o CXCR4. A cultura que apresentar atividade viral demonstrará qual o tropismo do vírus (LOW, MCGOVERN, *et al.*, 2009).

Apesar de sua eficácia, o Trofile® não pode ser implementado na rotina clínica por razões de natureza logística: o ensaio é caro (em 2010, cada teste custava cerca de 1.500 USD), o resultado é demorado (mais de duas semanas após o recebimento da amostra) e não tem sua tecnologia distribuída (somente um laboratório possui a estrutura necessária para realizá-lo) (POVEDA, ALCAMÍ, *et al.*, 2010).

2.2.2 Geno2pheno[coreceptor]

O geno2pheno[coreceptor] é um sistema genotípico, lançado como uma página *web*, e que visa predizer o tropismo do HIV-1 a partir da região V3 da proteína gp120 do envelope do HIV. Para a construção do sistema, foram utilizadas cerca de 1100 sequências para treinamento e validação. Não é totalmente clara a origem de todos os dados de treinamento deste classificador, porém sabe-se que parte das sequências são originárias da base de dados Los Alamos (LENGAUER, SANDER, *et al.*, 2007).

O sistema geno2pheno[coreceptor] baseia-se no algoritmo de máquina de vetores de suporte (*Support Vector Machine*, SVM), um método de aprendizado estatístico utilizado em problemas de regressão e classificação, mas principalmente neste último cenário. A princípio, na classificação, o SVM utiliza os dados de treinamento para reconhecer padrões entre as observações e, para novas observações, atribui classes, sem atribuir probabilidades que pertença a esta ou aquela classe. Ou seja, trata-se de um classificador linear não-probabilístico.

A ideia principal do algoritmo é encontrar uma fronteira que melhor separa as duas classes em estudo. Dado um conjunto de treinamento rotulado, a SVM utiliza dados numéricos para criar dois vetores de margem, os chamados vetores de suporte, (Figura *2.6*, em tracejado), construídos pelas observações de uma classe mais próximas da outra.

Com estes vetores definidos, calcula-se um hiperplano (Figura *2.6*, em vermelho) que definirá a separação entre as classes de forma a colocar o maior número de pontos da mesma classe do mesmo lado, enquanto maximiza a distância de cada classe a esse hiperplano. Essa distância entre o hiperplano e o primeiro ponto de cada classe recebe o nome de margem (Figura *2.6*, área em amarelo). O hiperplano possui o mesmo número de dimensões que o número de variáveis explicativas do conjunto de dados (CORTES, VAPNIK, 1995).

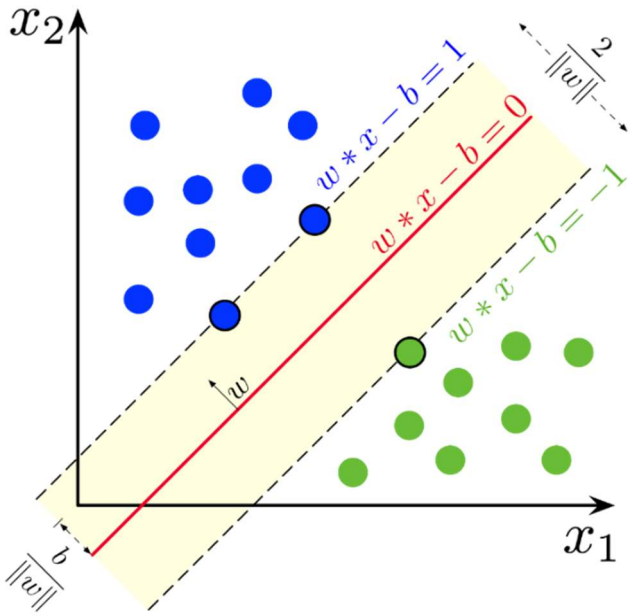

Figura 2.6*: Representação gráfica dos vetores de suporte (linhas tracejadas) e do hiperplano (linha vermelha), separando as observações (círculos azuis e verdes) em duas classes (Wikimedia).*

O geno2pheno[coreceptor] é o teste genotípico mais utilizado atualmente, principalmente como uma ferramenta de triagem de portadores de HIV-1 que possam se candidatar a estudos de medicamentos bloqueadores da ligação com o correceptor (KAGAN, JOHNSON, *et al.*, 2014).

2.2.3 Two-level Coreceptor Usage Prediction (T-CUP) 2.0

O T-CUP 2.0 é um classificador de tropismo de HIV lançado como pacote para a linguagem de programação R. O modelo se baseia no algoritmo de *Random Forest* (RF)

(seção 2.3) e foi treinado com um conjunto de 1351 sequências de aminoácidos da alça V3 (1151 de tropismo R5, 166 X4 e 34 R5X4) obtidos da base de dados de Los Alamos. Deste conjunto, cerca de 52% são de subtipo B, 17% de subtipo C e 9% de subtipo D. Os outros 22% são compostos por outros subtipos e CRFs (HEIDER, DYBOWSKI, *et al.*, 2014).

Para transformar os aminoácidos em valores numéricos, foi utilizado o pacote `Interpol`, utilizando os critérios de hidrofobicidade e descrição da carga líquida dos aminoácidos (HEIDER, 2012, R CORE TEAM, 2018).

Este modelo utiliza informações estruturais da alça V3 analisando as coordenadas do carbono alfa de cada aminoácido da sequência em relação ao uma sequência modelo. Dessa maneira, é possível driblar problemas relacionados com diferenças de aminoácidos em cada posição (HEIDER, DYBOWSKI, *et al.*, 2014).

2.3 Random Forest

O algoritmo de *Random Forest* (RF) é um método supervisionado de aprendizado de máquina que utiliza a agregação de preditores do tipo árvore combinada com técnicas de randomização. Cada árvore de decisão construída atribui um valor à observação, valendo como um voto (para árvores de classificação) ou um valor numérico (para árvores de regressão). Nos problemas de classificação, o valor dado pela maioria das árvores (voto majoritário) é atribuído pela RF. Já nas árvores de regressão, o valor final é dado pela média (ou mediana) dos valores encontrados pelas árvores (BREIMAN, 2017). Originalmente, o algoritmo de árvore de decisão utilizado é o CART (*Classification and Regression Trees*) (BREIMAN, 2017).

Árvores de decisão (Figura 2.7) são estruturas hierárquicas organizadas em nós, onde se faz a divisão de um conjunto de dados com base numa análise booleana de suas variáveis explicativas. As árvores são iniciadas em um nó raiz, compreendendo todo o conjunto de treinamento. Em seguida, busca-se recursivamente a variável explicativa que fornecerá a melhor divisão binária dos dados de forma a minimizar uma medida da impureza do nó (critério de divisão). O conjunto de dados é dividido de acordo com os valores (numéricos ou categóricos) em dois nós descendentes: que atendem (lado esquerdo) e não atendem a condição prevista pelo nó (lado direito). Ao final, a árvore

define os grupos que passaram por aquele circuito de critérios de separação (folhas). É possível que cada nó se divida em diversos ramos, mas aplica-se apenas a separação binária, uma vez que a separação em muitos grupos pode fragmentar excessivamente o conjunto de dados, impossibilitando novas separações nos nós seguintes. Esse processo se repete até atingir um critério de parada. Dada uma nova observação, esta atravessa a árvore da raiz até a folha (nó final), passando pelos nós de decisão (HASTIE, TIBSHIRANI, *et al.*, 2009).

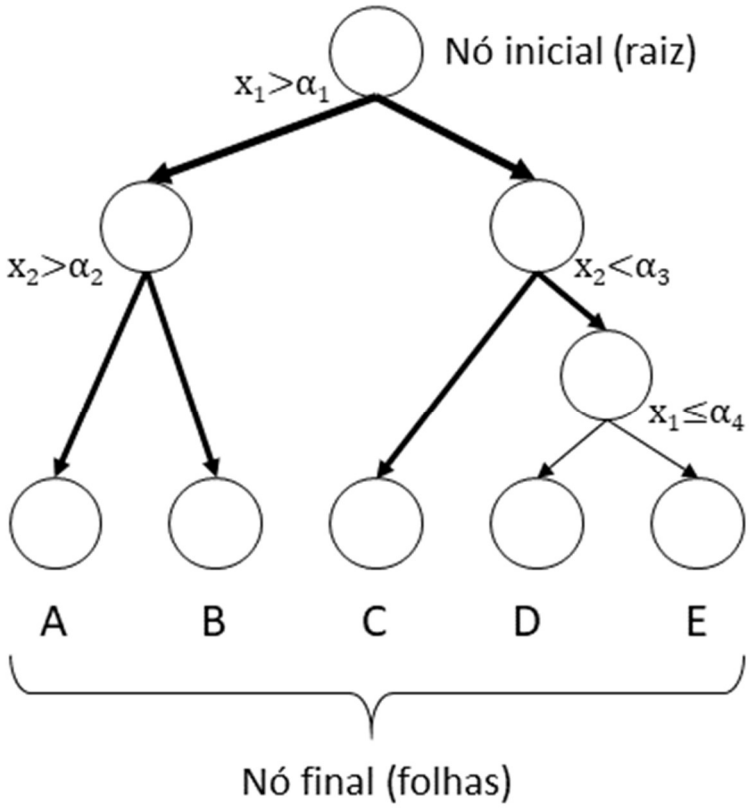

Figura 2.7: Representação de uma árvore de decisão (Do autor).

As árvores de decisão são modelos muito bem vistos pela sua facilidade de compreensão e velocidade de execução. Entretanto, a árvore de decisão pode sofrer problemas quando o conjunto de dados usado em seu treinamento possui uma grande quantidade de variáveis. O uso de muitos preditores pode levar a um superajuste do modelo, de tal forma que se torna inviável sua utilização para predizer corretamente um novo conjunto de dados. Esta condição é chamada *overfitting* (DIETTERICH, 1995, TIN KAM HO, 1995).

2.3.1 Bagging

Em uma RF, as árvores são construídas seguindo o mecanismo de *bagging* (do acrônimo ***b**ootstrap **agg**regati**ng***, agregação por *bootstrap*) (Figura 2.8). Cada árvore é construída a partir de um subconjunto obtido por amostragem aleatória e com repetição do conjunto de treinamento. Cada subamostra possui o mesmo tamanho do conjunto original, implicando na possibilidade de observações do conjunto de treinamento aparecerem diversas vezes em um subconjunto. Estima-se que cada subamostra possui, em média, dois terços das observações do conjunto original. Este mecanismo permite a construção de RFs com menor variância e mais resistentes ao problema do *overfitting* (BREIMAN, 1996, EFRON, TIBSHIRANI, 1993, HASTIE, TIBSHIRANI, *et al.*, 2009, JAMES, WITTEN, *et al.*, 2013).

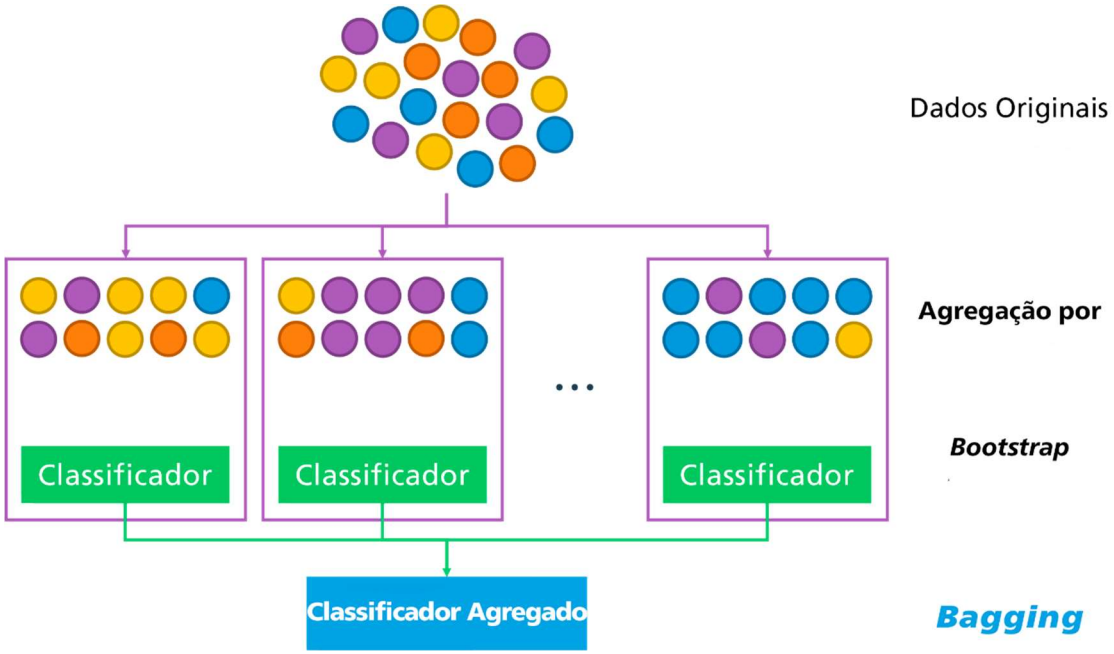

Figura 2.8: Ilustração do conceito de bagging (Wikimedia, adaptado).

As observações deixadas de fora de cada subamostra, ditas "fora-da-bolsa" (*out-of-bag,* OOB), são utilizadas para testar a preditibilidade de cada árvore, avaliando o desempenho de cada uma destas. O índice de erro destas árvores é chamado de *erro OOB* e, em uma RF com o número de árvores suficientemente grande, o erro OOB equivale a

uma de validação cruzada *leave-one-out*. Inclusive, esta avaliação exime a necessidade de um conjunto de validação para a RF, tornando o algoritmo muito conveniente (JAMES, WITTEN, *et al.*, 2013).

2.3.2 Randomização

Outro mecanismo usado para randomizar as árvores é a amostragem de variáveis preditivas. Em um conjunto de n variáveis, o algoritmo seleciona aleatoriamente k preditores e, destes, seleciona a melhor variável para estabelecer uma divisão em dois nós descendentes. Caso a RF seja usada para classificação, recomenda-se que o número de variáveis por árvore seja determinado pela função $k = \lfloor \sqrt{n} \rfloor$, onde k é o número de variáveis usada por árvore, e n é o total de variáveis do conjunto. Porém, se for usado para regressão, recomenda-se o uso da fórmula $k = \left\lfloor \frac{n}{3} \right\rfloor$, com um tamanho mínimo de cinco nós por árvore (BREIMAN, 2001).

2.3.3 Importância das variáveis

Além das vantagens já apresentadas em relação à RF, o algoritmo também fornece medidas para avaliar a importância das variáveis preditivas, a fim de identificar quais preditores são mais importantes para explicar o problema em questão. Isso porque as variáveis de um conjunto de dados podem ter capacidades distintas de separar corretamente as observações entre as classes (para o caso de um modelo de classificação) ou de estimar o valor médio (para o caso de um modelo de regressão).

Uma das estratégias utilizadas para medir a importância das variáveis em problemas de classificação é a permutação. A ideia principal é que, se a variável não for importante para o problema, reorganizar os valores dessa variável não interferirá na acurácia da predição. Supõe-se uma RF com *N* árvores de decisão e *J* variáveis no seu conjunto de treinamento. Na construção da n-ésima árvore, a acurácia é obtida ao testá-la com o conjunto OOB correspondente. Em seguida, os valores da *j*-ésima variável na amostra OOB são permutados aleatoriamente, preservando todas as outras variáveis. Esses dados OOB modificados são transmitidos à árvore e novos valores preditos são obtidos. A nova acurácia obtida com os dados permutados também é registrada e a

diferença entre a acurácia com e sem permutação é calculada. Repete-se o mesmo procedimento em todas as árvores. Ao final, calcula-se a média das diferenças entre a acurácia com e sem permutação em todas as árvores, fornecendo a importância da *j*-ésima variável. Este procedimento é repetido para todas as variáveis de interesse. Quanto maior o valor da importância da permutação de uma variável, mais importante é a variável para a acurácia geral da predição (HASTIE, TIBSHIRANI, *et al.*, 2009).

3 Revisão da literatura

Desde a descoberta de Gallo et al (1983) e de Barré-Sinoussi et al (1983) que a Aids é causada pelo vírus HIV, houve uma corrida para se entender as vias que o vírus usa para infectar o ser humano. Maddon et al (1986) e McDougal et al (1986) foram os primeiros a demonstrar que o receptor CD4 era fundamental para a interação vírus-célula. Levy (1986) mostrou que as proteínas presentes no envelope viral possuem alta variabilidade e classificou o HIV como um retrovírus.

O primeiro trabalho relacionando a infecção por HIV e a alça V3, publicado por Sattentau e Moore (1991), demonstrou que, bloqueando esta região com anticorpos, a infecção era reduzida drasticamente. De Jong (1992) e Ivanoff (1992) sugeriram que as mutações ocorridas na alça V3 estariam diretamente relacionadas com a infectividade do vírus. Fouchier et al (1992) atestaram que a mudança específica do 11º e 28º aminoácidos da alça V3 determinaria a afinidade do HIV por determinadas células, enquanto Milich et al (1993) adicionaram o 25º aminoácido da sequência peptídica como um fator determinante.

Deng et al (1996) e Trkola et al (1996) descobriram separadamente que o receptor CCR5 agia como cofator para a entrada do HIV nas células-alvo. Moore et al (1997) mostraram que, além do CCR5, o receptor CXCR4 também poderia auxiliar na entrada do vírus. Por isso, Berger (1998) sugeriu classificar o fenótipo dos vírus encontrados também por seu tropismo pelos correceptores: fenótipo R5 para os afins por CCR5; X4 para os afins por CXCR4, e R5X4 para os que possuíam tropismo pelos dois receptores.

O primeiro trabalho que sugeriu algum tipo de relação direta entre os aminoácidos na sequência V3 e o tropismo viral por um receptor especificamente foi de Xiao et al (1998). Nele, o grupo utilizou vinte e quatro isolados de HIV-1, representando diversos subtipos virais. Estes isolados foram colocados em cultura de linfócitos carentes do receptor CCR5. Parte das linhagens não conseguiu infectar as células, evidenciando que o tropismo destas partículas virais era de fato por CCR5. A alça V3 destas foi analisada para entender o que haveria de comum. Foi constatado que a posição 11 da sequência da alça havia conservação dos resíduos de aminoácidos de carga negativa (especialmente serina e glicina). Além disso, a posição 25 também mostrava resíduos de aminoácidos negativamente carregados, como glutamato e aspartato. Quando confrontado com as

linhagens virais que não possuíam tropismo por CCR5, viu-se que o resíduo da posição 25 era substituído por arginina ou glutamina, aminoácidos de carga positiva. Isso levou a estipular o consenso de que, se os resíduos das posições 11 e 25 fossem negativos, o vírus teria tropismo por CCR5, caso contrário, CXCR4. Esse entendimento ficou posteriormente conhecido como "regra 11/25".

Resch et al (2001) confirmaram a hipótese de Xiao, ao construir redes neurais *feedforward*. O grupo obteve 778 sequências V3 ao combinar diversos conjuntos de dados disponíveis. As sequências foram alinhadas de modo a conter 35 resíduos cada, permitindo *gaps*. A partir de análises estatísticas e empíricas, foram selecionadas 15 posições (5, 7, 8, 10, 11, 13, 18, 19, 20, 21, 22, 24, 25, 27, 32) mais a carga global da sequência peptídica, totalizando 16 variáveis de entrada. O trabalho de Resch, além de criar um classificador baseado em rede neural artificial, também testou os métodos existentes à época para prever o tropismo do HIV-1. De fato, o método para a classificação do tropismo com melhor desempenho foi a regra 11/25, mesmo com o poder de predizer o fenótipo X4 por volta de 50%. Então o novo classificador foi comparado a esta regra e obteve desempenho mais elevado, tanto para predizer o fenótipo R5 quanto X4.

Diversos estudos, como o de Stanfield et al (2006), revelaram que a alça V3 possuía uma estrutura muito bem conservada, mesmo com a variabilidade que a sequência apresenta. Porém, em outra linha de pesquisa, Sharon et al (2003) já mostrava que a alça das partículas virais com fenótipo R5 possuía homologia estrutural com três ligantes do receptor, RANTES, MIP-1α e MIP-1β, enquanto o vírus de fenótipo X4 era homólogo ao ligante SDF-1. Assim, haveria uma diferença significativa que poderia indicar a seletividade entre as estruturas da alça de espécimes com fenótipos R5 e X4. Com isso, Cardozo et al (2007) investigou, utilizando uma base de 240 sequências peptídicas, como a interação entre os aminoácidos poderiam determinar o tropismo do HIV-1. Inicialmente, ficou claro que, pelas dobras da alça, além dos aminoácidos 11 e 25, o resíduo da posição 24 também poderia influenciar na determinação do tropismo. Os dados deste estudo sugeriram uma nova regra, chamada de "regra 11/24/25", que enuncia: "um aminoácido positivamente carregado nas posições 11, 24 ou 25 define X4; senão, R5".

Trouplin et al (2001) desenvolveram uma técnica de recombinação genética para determinar o tropismo viral. Nela, uma etapa de subclonagem é adicionada à cultura do

HIV-1, na qual produz-se pseudovírus (partículas virais capazes de infectar as células de uma cultura, mas incapazes de se replicar) com o gene da gp120 V3 retirado de uma amostra sanguínea de um portador do HIV-1, junto com um gene-repórter de β-galactosidase. Para testar o tropismo, faz-se a infecção de duas culturas celulares, uma expressando receptores CCR5, outra CXCR4. A partir da detecção colorimétrica, determina-se qual o tropismo daquela amostra.

Whitcomb et al (2007) modificaram a técnica, com alterações como o uso de todo o gene *env* derivado do paciente, além de trocar a β-galactosidase por luciferase. O ensaio aperfeiçoado foi capaz de detectar populações minoritárias com até 5% de presença entre as partículas virais colhidas do portador. Assim surgiu o Trofile®, um teste altamente preciso e automatizado para determinação do tropismo do HIV-1.

A predição a partir do sequenciamento da alça V3 não ficou apenas nas regras enunciadas por Xiao e Cardozo. Sander et al (2007), com um conjunto de 363 sequências peptídicas únicas, mostrou que, não apenas a avaliação das características dos aminoácidos, mas também adicionando informações estruturais da alça pode melhorar a predição do uso do correceptor. O grupo propôs um classificador baseado em um algoritmo de SVM que combinaria distâncias de átomos dos aminoácidos, características físico-químicas das moléculas as quais pertenciam, e os próprios aminoácidos, transformados em valores numéricos usando algum tipo de indicador. Os resultados sugeriram que alguns sítios de ligação de hidrogênio e algumas cadeias alifáticas foram particularmente relevantes para a predição do tropismo. Este trabalho foi o primeiro a combinar a análise estrutural da alça V3 e métodos de aprendizado estatístico. De acordo com Lengauer et al (2007), este método desenvolvido foi incorporado ao Geno2Pheno, uma ferramenta *online* desenvolvida por Beerenwinkel et al (2003) que, através da análise genética do HIV, inicialmente, preveria a resistência à terapia antirretroviral.

Xu et al (2007) aplicaram pela primeira vez o algoritmo de *Random Forest* para predizer o tropismo do HIV-1 a partir da alça V3. Este estudo também foi o primeiro a considerar a polaridade da alça como uma variável explicativa válida, além da carga líquida e a distribuição dos aminoácidos nas 35 posições da sequência peptídica, totalizando 37 variáveis. O estudo, a partir de um conjunto de 651 sequências peptídicas, resulto em um modelo que obteve alto desempenho, com acurácia de 95.1% e pontuação do Critério de Correlação de Matthews (*Matthews's Correlation Criteria*, MCC) de 0.87.

Das variáveis usadas para a construção do modelo preditor de tropismo, sete foram destacadas como as mais importantes: Carga líquida, polaridade e cinco posições da sequência (22, 25, 11, 12, e 13).

Lieberman-Blum et al (2008) apresentaram o Maraviroque®, fármaco inibidor de CCR5 que logo foi adicionado à lista de medicamentos que poderiam ser usados no tratamento contra o HIV-1. Porém, para que o medicamento seja bem administrado, é importante que o tropismo viral seja determinado corretamente. Com isto em mente, Dybowski et al (2010) desenvolveram o classificador T-CUP, um método baseado em *Random Forest* em dois níveis: no primeiro, o modelo considera, separadamente, a sequência de 35 aminoácidos, e duas características físico-químicas (Hidrofobicidade, usando a escala Kyte-Doolittle, e força eletrostática na superfície do peptídeo, resolvendo equações de Poisson-Boltzmann para calcular os valores). No segundo nível, essas abordagens separadas são combinadas, chegando ao resultado final. O modelo mostrou elevada robustez para predizer o uso dos correceptores, com área sob a curva ROC de 0.937. Desta vez, as posições da sequência que obtiveram maiores níveis de importância foram 11, 22, 23, 24 e 27. Supreendentemente, a posição 25 teve pouca importância na construção do modelo. Comparado a outros classificadores como Geno2Pheno e as regras 11/25 e 11/24/25, o T-CUP se mostrou com melhor desempenho nos quesitos de sensibilidade, especificidade e acurácia.

Heider et al (2014) aprimorou o modelo de Dybowski e relançou como T-CUP 2.0. O princípio do classificador foi mantido: uma combinação em dois níveis de hidrofobicidade, força eletrostática e a sequência dos aminoácidos. Porém, para reduzir a demanda computacional que a resolução das equações de Poisson-Boltzmann pedia, um modelo simplificado dos aminoácidos foi desenvolvido de tal maneira que os átomos da cadeia lateral de cada resíduo foram considerados "agrupados" na posição espacial do carbono-α do aminoácido correspondente. Dessa forma, as equações se tornaram muito mais simples para ser resolvidas, permitindo a execução dos modelos com menor exigência computacional por parte do usuário. Além desta melhoria da eficiência, T-CUP 2.0 também se mostrou ainda mais robusto, com área sob a curva ROC de 0,968. Neste modelo as posições mais importantes foram 7, 8, 12, 14, 20, 21, 24, 27, 29 e 35. Apesar de não aparecer as duas posições da regra 11/25, concluiu-se que o modelo concorda com a regra, uma vez que o alinhamento das sequências realizado e a posição espacial estimada

na preparação dos dados poderiam acarretar pequenas diferenças no cálculo da posição do aminoácido na sequência.

Também motivado pela determinação correta do tropismo para o direcionamento do tratamento contra o vírus, Chen et al (2019) desenvolveram dois classificadores, um baseado no algoritmo de gradiente de *boosting* (XGBpred) e outro baseado no modelo oculto de Markov (HMMpred). Os dois classificadores foram testados contra o Geno2pheno e atingiram níveis de sensibilidade superiores ao classificador já consagrado (72,56% para o XGBpred, 72,09% para HMMpred, e 61,6% para o Geno2pheno, nivelando os modelos pela especificidade). Ademais, o modelo XGBpred demonstrou altía robustez (área sob a curva ROC de 0.949) ao ser testado em um conjunto de sequências com maior proporção de amostras de tropismo dual (R5X4). Demonstrou alta capacidade de predizer as amostras R5X4 como X4, o que é desejável para este tipo de classificador binário. Por fim, ficou determinado pelos modelos que as posições 5, 11, 13, 18, 22, 24 e 25 foram correlacionadas com o tropismo do HIV-1.

É fato que houve uma grande evolução nos algoritmos que predizem o tropismo do HIV ao longo do tempo. Ainda assim, nenhum dos estudos buscou esclarecer se o tratamento dos dados iniciais poderia melhorar o desempenho dos modelos. Neste sentido, nosso trabalho inicia esta discussão ao comparar métodos de amostragem dos dados e seleção automática das variáveis.

4 Materiais e Métodos

Todos as operações foram realizadas a partir da linguagem de programação R (R CORE TEAM, 2018), versão 3.6.3, utilizando a interface RStudio, versão 1.12. Dentro desta linguagem, foram utilizados diversos pacotes de funções, mencionados ao longo do texto.

4.1 Dados

O conjunto de dados utilizado neste trabalho foi obtido em agosto de 2018 a partir da base de dados de HIV do Laboratório Nacional de Los Alamos (Los Alamos, NM, EUA, https://www.hiv.lanl.gov/). 8.536 sequências correspondentes à região da alça V3 da gp41 do vírus HIV-1 foram obtidas no total, sendo 6.820 R5-trópicas, 590 X4-trópicas e 1126 dual trópicas (R5X4). De acordo com o banco de dados de HIV, as informações de tropismo do correceptor de cada sequência anotada foram baseadas em ensaios fenotípicos.

4.1.1 Filtragem e preparação das sequências

Os dados coletados foram lidos a partir de um arquivo de formato FASTA com auxílio do pacote `seqinr` (CHARIF, LOBRY, 2007) e, a partir daí, foram filtrados conforme mostra o esquema da Figura 4.1: Inicialmente, foram removidas 5.099 sequências duplicadas da base original. Em seguida, foram removidas 449 sequências que não começavam e terminavam com cisteína (C), uma vez que, de acordo com Shpaer et al (1994), os aminoácidos da primeira e última posições são altamente conservados. Depois, foram removidas 10 sequências que possuem entre 31 e 39 aminoácidos. A seguir, com uso dos pacotes `stringi` (GAGOLEWSKI, 2019) e `stringr` (WICKHAM, 2019), foram removidas 261 sequencias que contivessem símbolos de *frameshifts* (#) e *stop codons* ($).

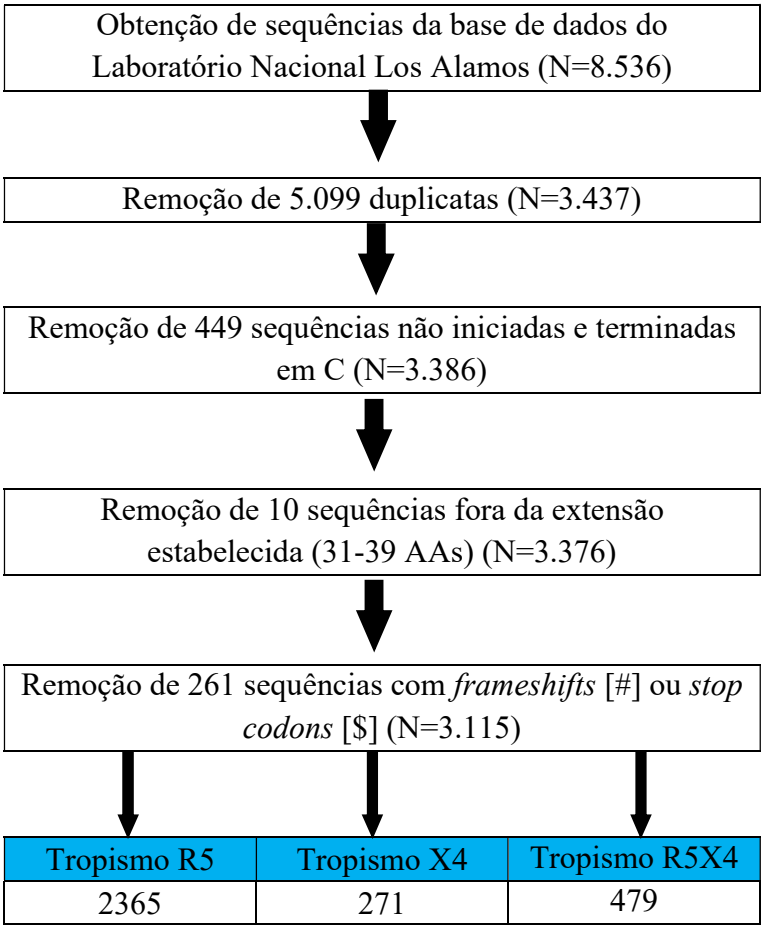

Figura 4.1: Esquema de filtragem das sequências da base de Los Alamos.

As sequências foram alinhadas a uma sequência consenso de 35 aminoácidos (CTRPNNNTRKSIRIGPGQAFYATGDIIGDIRQAHC), obtida a partir da base de dados Los Alamos, por meio do algoritmo Needleman-Wunsch com parâmetros *default* do pacote `Biostrings` (PAGÈS, ABOYOUN, *et al.*, 2018). Eventuais *gaps* foram completados com um pseudoaminoácido B, como sugerido por Shen et al (2016), e, mais tarde, substituídos pelo aminoácido mais frequente da posição. Ao final, foi obtido um conjunto de dados de 3.115 sequências com o mesmo comprimento (35 aminoácidos), sendo 2.365 R5-trópicas, 271 X4-trópicas e 479 dual-trópicas (R5X4).

4.1.2 Conversão

Para serem consistente com todos os outros métodos com os quais nossos resultados foram comparados e devido à escassez de variantes estritamente X4 no conjunto de dados original, as variantes X4 e R5X4 foram agrupadas em uma única classe denominada NR5 (não-R5).

Para este estudo, apenas as sequências de subtipo B e C foram utilizadas. O subtipo B (n = 1.622) apresentou 1.284 sequências R5 e 338 NR5, enquanto o subtipo C (n = 560) apresentou 482 sequências R5 e 78 NR5, como esquematizado na Tabela 4.1.

Tabela 4.1 Distribuição das sequência segundo o subtipo e tropismo.

		Sequências	
		Subtipo B	Subtipo C
Tropismo	R5	1284	482
	NR5	338	78
	Total	1622	560

Para a representação dos dados, os aminoácidos foram codificados usando escores de hidrofobicidade de acordo com a escala de Engelman (ENGELMAN, STEITZ, *et al.*, 1986) (Tabela 4.2) por meio do pacote `Peptides` (OSORIO, RONDON-VILLARREAL, *et al.*, 2015). Esta escala foi escolhida por apresentar valores únicos para cada aminoácido.

Tabela 4.2: Aminoácidos e seus valores correspondentes na escala de Engelman.

Aminoácido	Símbolo	Escala de Engelman
Alanina	A	-1,6
Cisteína	C	-2,0
Aspartato	D	9,2
Glutamato	E	8,2
Fenilalanina	F	-3,7
Glicina	G	-1,0
Histidina	H	3,0
Isoleucina	I	-3,1
Lisina	K	8,8
Leucina	L	-2,8
Metionina	M	-3,4
Asparagina	N	4,8
Prolina	P	0,2
Glutamina	Q	4,1
Arginina	R	12,3
Serina	S	-0,6
Treonina	T	-1,2
Valina	V	-2,6
Triptofano	W	-1,9
Tirosina	Y	0,7

4.2 Construção dos modelos

4.2.1 Conjuntos para o subtipo B

Para o subtipo B, os dados foram divididos aleatoriamente, mantendo as proporções originais de tropismo R5 e NR5, em dois subconjuntos mutuamente exclusivos, por meio do pacote `caret` (KUHN, WESTON, *et al.*, 2019), na proporção 7:3. O conjunto de treinamento dispunha de 1.136 sequências (899 R5 e 237 NR5) e o conjunto de teste, 486 sequências (385 R5 e 101 NR5). O conjunto de teste foi utilizado apenas nas etapas de avaliação dos modelos propostos e dos métodos genotípicos já consolidados.

O tropismo das sequências do conjunto de treinamento estava distribuído na proporção 4:1 (79% para R5 e 21% para NR5). Para conferir se este desequilíbrio de fato afetaria o desempenho dos modelos, novos conjuntos de treinamento foram criados

seguindo formas de balancear o número de sequências: (i) balanceamento por sobreamostragem, em que as sequências NR5 foram reamostradas aleatoriamente de maneira que o número de representantes desta classe minoritária fosse igual ao de sequências R5 e (ii) balanceamento por subamostragem, em que a classe majoritária (R5) foi subamostrada aleatoriamente. Para realizar a reamostragem, foi usada a função `sample`, pertencente às funções básicas do R.

Como mostrado na Tabela 4.3, os conjuntos tomaram as seguintes quantidades: 1.136 sequências (899 R5 e 237 NR5) para o conjunto sem balanceamento; 1.798 sequências (899 R5 e 899 NR5) para o conjunto balanceado por sobreamostragem; e 474 sequências (237 R5 e 237 NR5) para subamostragem.

Tabela 4.3 Distribuição das sequências de subtipo B, de acordo com o tipo de amostragem.

		Tropismo		
Amostragem		R5	NR5	Total
	Sem balanceamento	899	237	1.136
	Sobreamostragem	899	899	1.798
	Subamostragem	237	237	474

Com o intuito de verificar se a exclusão de variáveis menos variantes poderia impactar no desempenho do classificador, também foi feita uma filtragem nas variáveis a serem utilizadas pelo modelo. Foi utilizada a função `nearZeroVar` do pacote `caret` para determinar quais variáveis seriam removidas do conjunto. A função, em sua configuração-padrão, detecta quais colunas de um conjunto de dados se encaixam em dois critérios: (i) possuem razão entre o valor mais frequente de uma variável e o segundo mais frequente superior à frequência de corte de 95/5; e (ii) possuem razão entre o número de valores únicos em cada variável e o total de observações inferior a 0.1.

Das 35 posições das sequências peptídicas, 17 delas (1, 3, 4, 6, 7, 8, 9, 15, 16, 17, 23, 24, 28, 30, 31, 33, e 35) foram removidas.

Ao final, seis conjuntos foram criados a partir das sequências de subtipo B: três conjuntos com todas as posições (chamados de conjuntos completos) que diferenciam pela forma de amostragem (sem balanceamento, sobreamostragem e subamostragem) e

mais três conjuntos sem as variáveis apontadas pela função `nearZeroVar` (chamados conjuntos reduzidos).

4.2.2 Conjuntos para o subtipo C

Para os dados do subtipo C, o desequilíbrio entre as observações foi ainda mais discrepante que o visto para o subtipo B: 6:1 (86% R5 e 14% NR5), e foram utilizadas as mesmas técnicas de balanceamento supracitadas, como mostrada na Tabela 4.4.

Tabela 4.4: Distribuição das sequências de subtipo C, de acordo com o tipo de amostragem.

		Tropismo		
Amostragem		R5	NR5	Total
	Sem balanceamento	482	78	560
	Sobreamostragem	482	482	964
	Subamostragem	78	78	156

Também foi utilizada a função `nearZeroVar` para indicar as posições com menor variância. As posições indicadas como pouco variantes para o subtipo C foram as mesmas para o subtipo B (1, 3, 4, 6, 7, 8, 9, 15, 16, 17, 23, 24, 28, 30, 31, 33, e 35).

Devido ao reduzido número de observações, não houve divisão em subconjuntos mutuamente exclusivos de treinamento e teste. Foi utilizada a técnica de validação cruzada com 5 *folds*, por meio do pacote `caret`, em que cinco conjuntos foram construídos aleatoriamente a partir de todas as observações dos conjuntos para subtipo C; um conjunto foi usado para teste e os demais foram usados para treinamento. Esse processo foi repetido cinco vezes, de modo que cada conjunto de teste foi usado uma vez para o teste. O desempenho final foi dado pela média de desempenho dos cinco subconjuntos de teste.

4.2.3 Parâmetros dos modelos

Foram construídos modelos de *Random Forest* por meio da função `randomForest`, do pacote de mesmo nome (LIAW, WIENER, 2002) com os parâmetros *default*. Dessa forma, os modelos foram construídos com 500 árvores e cada nó teve o sorteio de 5 variáveis, conforme explicação no item 2.3.2.

Os valores de importância das variáveis em cada modelo (calculados conforme a explicação do item 2.3.3) foram mostrados em um gráfico gerado a partir da função `varImpPlot`, também do pacote `randomForest`.

Foram verificados, ao longo da construção da RF, os índices de erro OOB dos modelos, e seus valores foram representados em gráficos com o uso de funções dos pacotes `ggplot2` (WICKHAM, 2016), `ggpubr` (KASSAMBARA, 2020) e `reshape2` (WICKHAM, 2007). Para o subtipo C, uma vez que foi utilizada a validação cruzada, foi calculada a média do erro OOB ao longo da construção das cinco florestas.

4.3 Avaliação dos modelos

Para determinar o desempenho dos modelos, foram construídas, para cada modelo, curvas ROC (*Receiver-Operator Characteristic*) a fim de ilustrar a relação entre especificidade e sensibilidade com base na capacidade geral de discriminar entre as sequências R5 e NR5. A partir das curvas ROC foram determinados a área sob a curva (*Area Under Curve*, AUC) e os melhores pontos de corte para cada modelo. A AUC é uma medida capaz de avaliar a capacidade do modelo predizer corretamente o desfecho: uma área de 0,5 indica uma classificação aleatória, enquanto uma área igual a 1 indica um classificador perfeito. Foi utilizado o pacote `pROC` (ROBIN, TURCK, *et al.*, 2011) para a construção das curvas ROC e cálculo das AUCs. Os valores das AUCs foram representados em um gráfico *dotplot*, utilizando funções dos pacotes `ggplot2`, `ggpubr` (KASSAMBARA, 2020) e `broman` (BROMAN, BROMAN, 2019).

A partir dos pontos ótimos de cada curva, matrizes de confusão (Figura 4.2) foram determinadas pela função `confusionMatrix` do pacote `caret`, a fim de confrontar as predições dos modelos com os dados reais. Como as matrizes de confusão geralmente

denotam valores como “Positivo” e “Negativo”, foi determinado que as amostras NR5 seriam tratadas como Positivas neste estudo.

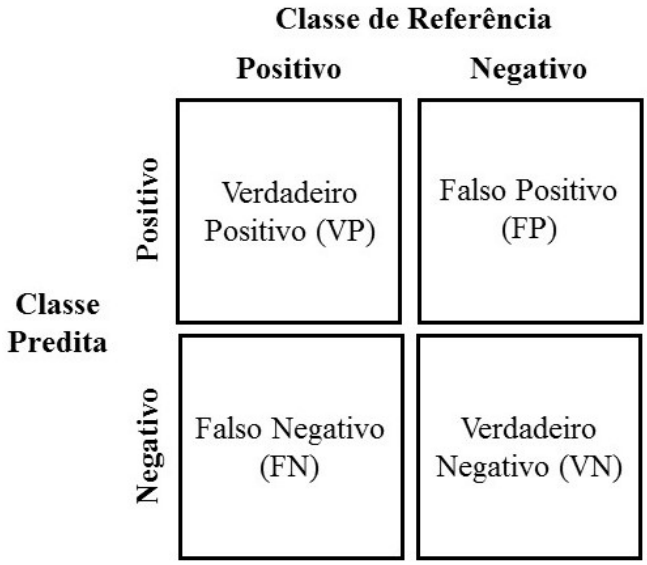

Figura 4.2: Matriz de confusão (do autor).

A partir desta matriz, foram obtidas as seguintes medidas de desempenho para avaliar os modelos: acurácia (Acc), sensibilidade (S), especificidade (E) e coeficiente de correlação de Matthews (*Matthews' Correlation Coefficient*, MCC). Este último foi escolhido devido à sua robustez ao lidar com dados desbalanceados. Valor de MCC igual a 1 corresponde a uma predição perfeita, enquanto o valor 0 aponta para uma predição completamente aleatória. Essas medidas podem ser calculadas a partir das seguintes equações:

$$Acc = \frac{VP + VN}{VP + FP + VN + FN} \quad (1)$$

$$S = \frac{VP}{VP + FN} \quad (2)$$

$$E = \frac{VN}{VN + FP} \quad (3)$$

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP)(VP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

nas quais VP é o número de verdadeiros positivos (X4 predito como X4), VN é o número de verdadeiros negativos (R5 predito como R5), FP é o número de falsos positivos (R5 predito como X4) e FN é o número de falsos negativos (X4 predito com R5).

4.4 Comparação com classificadores estabelecidos

Para dar suporte ao desempenho alcançado no presente trabalho, avaliamos dois outros métodos genotípicos com o mesmo conjunto de teste utilizado no diagnóstico dos modelos propostos: (i) geno2pheno$_{[coreceptor]}$, e (ii) T-CUP 2.0.

O T-CUP 2.0 foi executado a partir do pacote `TCUP2` (HEIDER, DYBOWSKI, 2013) com suas configurações padrão e para o geno2pheno$_{[coreceptor]}$ foram utilizados três pontos de corte para a taxa de falso-positivo: 2,5%, 5% e 10%.

As medidas de sensibilidade, especificidade, acurácia e MCC foram utilizadas como critério de avaliação e comparação com os modelos propostos no presente trabalho. Os mesmos conjuntos de teste utilizados para a avalição dos modelos propostos foram utilizados na avaliação dos métodos genotípicos já consolidados.

5 Resultados

5.1 Subtipo B

5.1.1 Características dos modelos

Foram criados seis modelos de classificação treinados com 1.136 sequências peptídicas da alça V3 do vírus HIV-1, subtipo B.

A AUC dos modelos completos (com todas as posições das sequências) e seus intervalos de confiança estão demonstrados na Figura 5.1-A. A AUC para o modelo treinado com o conjunto de dados subamostrado teve valor de 0.956, o modelo sobreamostrado teve valor de 0.955 e, para o modelo desbalanceado, 0.960. Os três IC 95% apresentaram interseção entre seus valores, indicando não haver diferença estatisticamente significativa entre os modelos.

Para os modelos reduzidos (apenas com variáveis selecionadas automaticamente), na Figura 5.1-B, o valor da AUC foi de 0.944 para o modelo subamostrado, 0.960 para o modelo sobreamostrado e 0.962 para o modelo desbalanceado. Assim como para os modelos completos, os três IC 95% apresentaram interseção entre seus valores, indicando não haver diferença estatisticamente significativa entre os modelos.

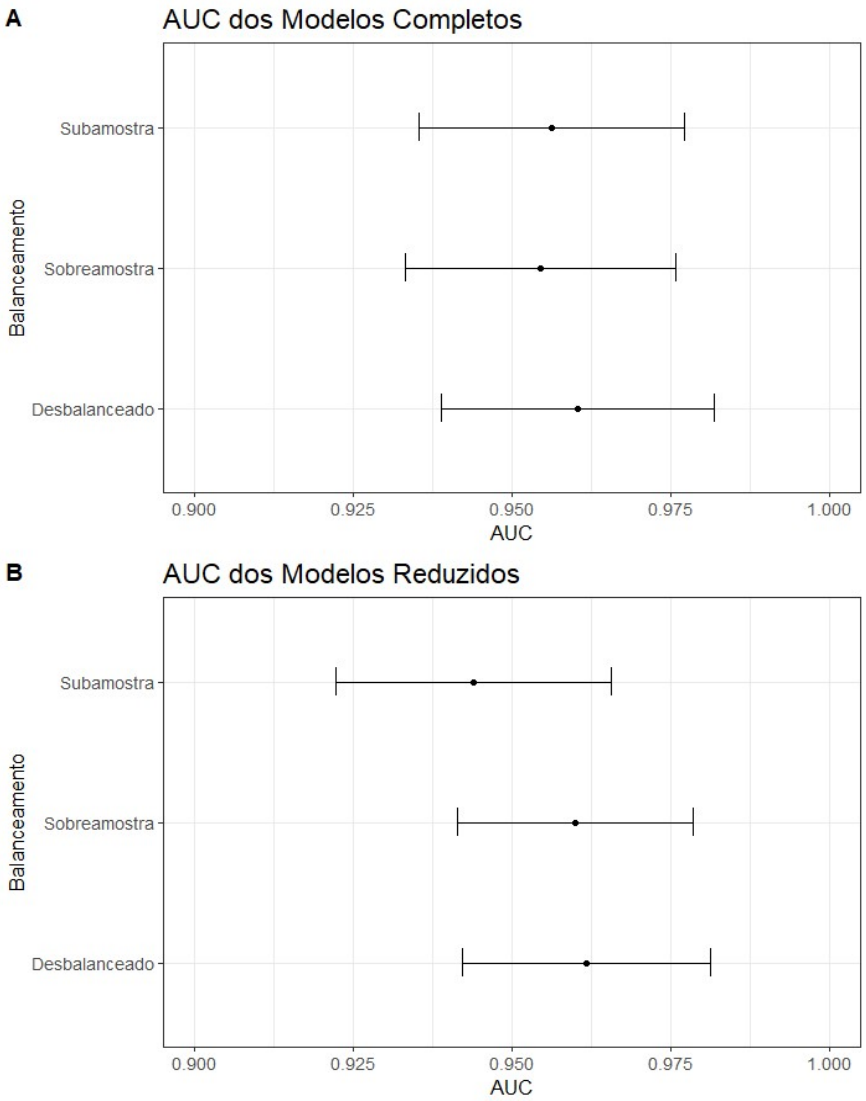

Figura 5.1: Área sob a curva ROC (AUC) dos modelos treinados com sequências do subtipo B. O intervalo de confiança de 95% foi calculado a partir de 2000 amostras bootstraps do conjunto de teste.

A Figura 5.2, que mostra a taxa de erro dos modelos durante a construção das árvores, mostra que os modelos sobreamostrados (completos ou reduzidos) obtiveram as menores taxas de erro, atingindo valores próximos a 5%. Os modelos desbalanceados se mantiveram com erro próximo a 10%, e os modelos subamostrados, próximo a 15%. Destaca-se que, com exceção dos modelos subamostrados, os modelos atingiram estabilidade no índice de erro por volta das 200 árvores. O modelo reduzido subamostrado continua diminuindo sua taxa de erro até o número máximo de árvores.

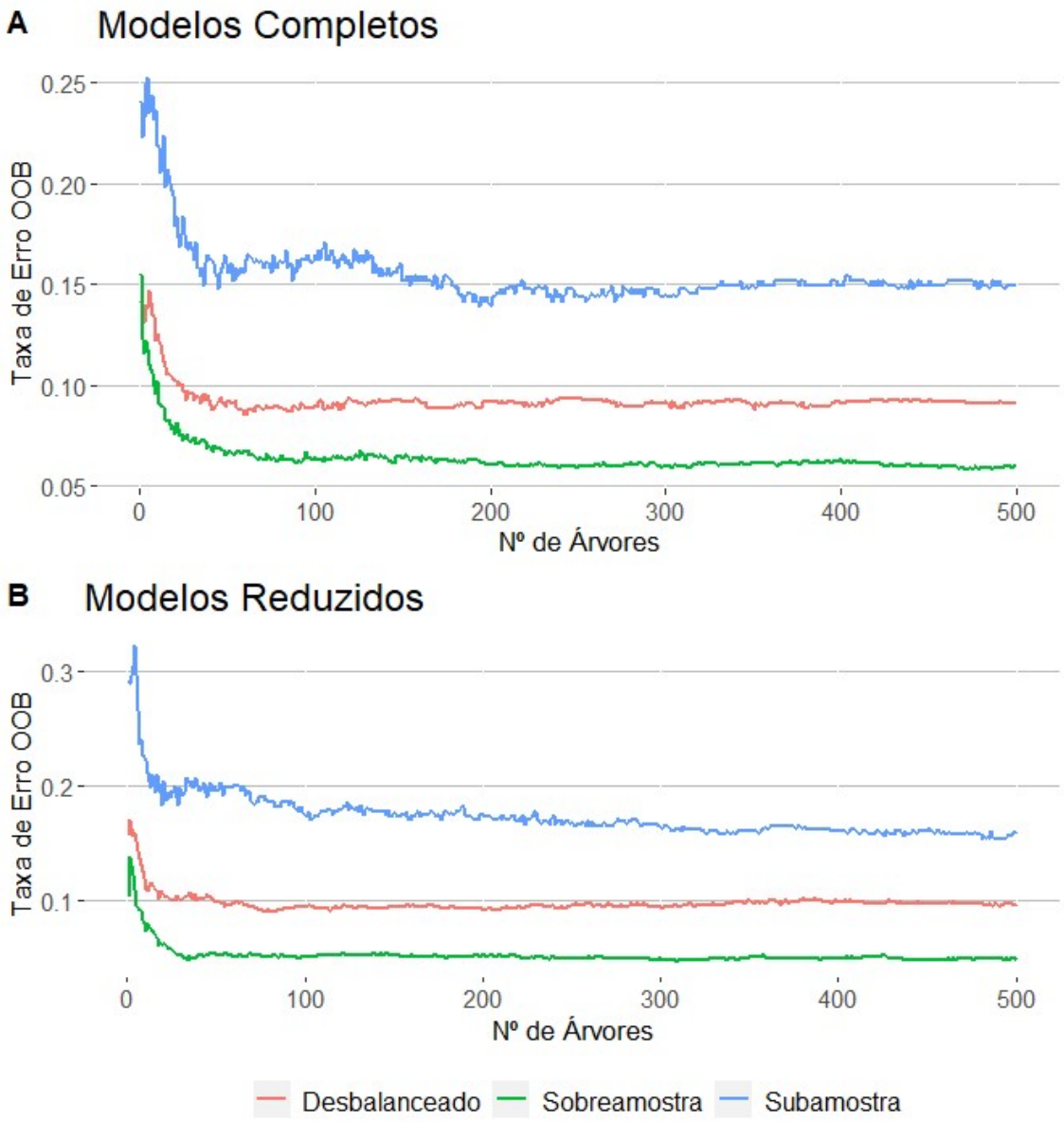

Figura 5.2 Taxa de erro para os conjuntos out-of-bag (OOB) durante a construção das árvores de cada modelo para o subtipo B.

5.1.2 Importância das variáveis

Nos modelos completos, conforme a Figura 5.3, observa-se que as posições 2, 7, 11, 13, 18, 24, 25 e 32 apareceram nos três conjuntos entre as cinco variáveis mais importantes dos modelos. Destas, as variáveis 11 e 18 se apresentaram em todos os três modelos. As variáveis 13 e 24 aparecem em dois modelos e as variáveis 2, 7, 22, 25 e 32 em um modelo.

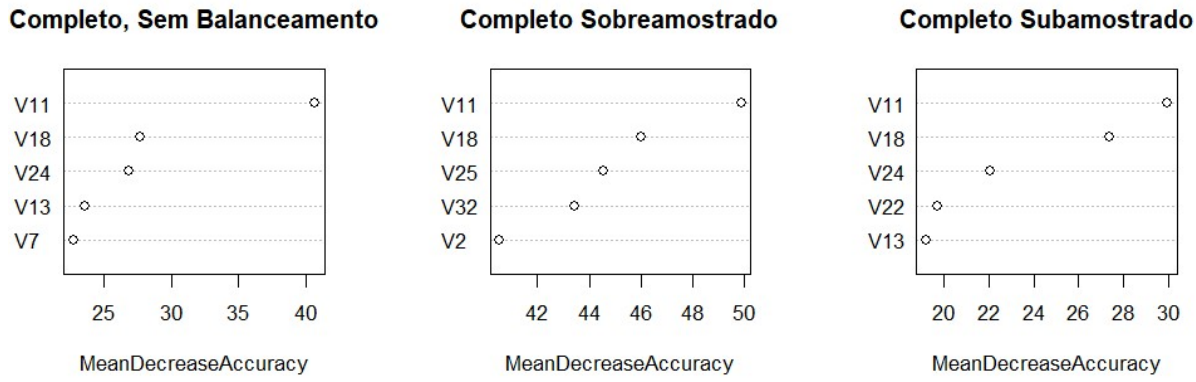

Figura 5.3 Pontuação de importância das cinco primeiras variáveis dos modelos completos.

Nos modelos reduzidos, como mostra a Figura 5.4, as variáveis 11, 13, 18, 22, 25 e 32 figuraram entre as cinco mais importantes. Destas, as variáveis 11, 13 e 18 apareceram em todos os três modelos, as variáveis 22 e 25 em dois modelos, e a variável 32 em um modelo.

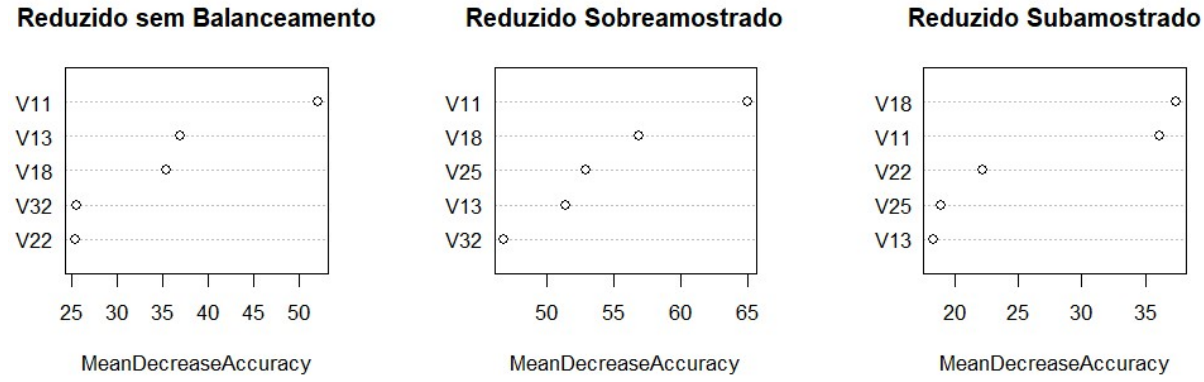

Figura 5.4 Pontuação de importância das cinco primeiras variáveis dos modelos reduzidos.

5.1.3 Desempenho e comparação dos modelos

De acordo com a Tabela 5.1, observa-se que o modelo reduzido e sobreamostrado (R-sobre) atingiu os maiores valores de acurácia (93,21%) e MCC (0,803). O modelo completo e desbalanceado (C-desb) obteve a maior sensibilidade (90,1%), e o classificador geno2pheno, com ponto de corte em 2,5% de falsos-positivo (g2p 2.5%), obteve o maior valor de especificidade (98,18%).

Tabela 5.1 Desempenho dos modelos e classificadores estabelecidos a partir do conjunto de teste do subtipo B. O maior valor de cada medida de desempenho foi marcado em negrito.

Classificador	Acurácia (%)	Especificidade (%)	Sensibilidade (%)	MCC
C-desb	91,15	91,43	**90,1**	0,759
C-sobre	91,98	94,03	84,16	0,763
C-sub	91,56	92,47	88,12	0,763
R-desb	92,18	93,51	87,13	0,774
R-sobre	**93,21**	94,29	89,11	**0,803**
R-sub	91,56	94,29	81,19	0,747
g2p 2.5%	93,00	**98,18**	73,27	0,778
g2p 5%	90,74	93,77	79,21	0,722
g2p 10%	87,45	87,79	86,14	0,671
T-CUP 2.0	90,95	93,25	82,18	0,734

5.2 Subtipo C

5.2.1 Características dos modelos

Foram criados seis modelos de classificação treinados com 560 sequências peptídicas da alça V3 do vírus HIV-1, subtipo C.

Para o subtipo C, os modelos completos apresentaram AUC média de 0,922 para o conjunto desbalanceado, 1 para o conjunto sobreamostrado, e 0,925 para o conjunto subamostrado. Quanto aos modelos reduzidos, a AUC média foi de 0,932 para o conjunto desbalanceado, 0,999 para o conjunto sobreamostrado, e 0,928 para o conjunto subamostrado. Devido ao reduzido número de observações para o subtipo, a avaliação foi realizada a partir da validação cruzada 5-fold. Observa-se, na Figura 5.5, a grande variabilidade da amplitude dos IC 95% dos diferentes *folds*.

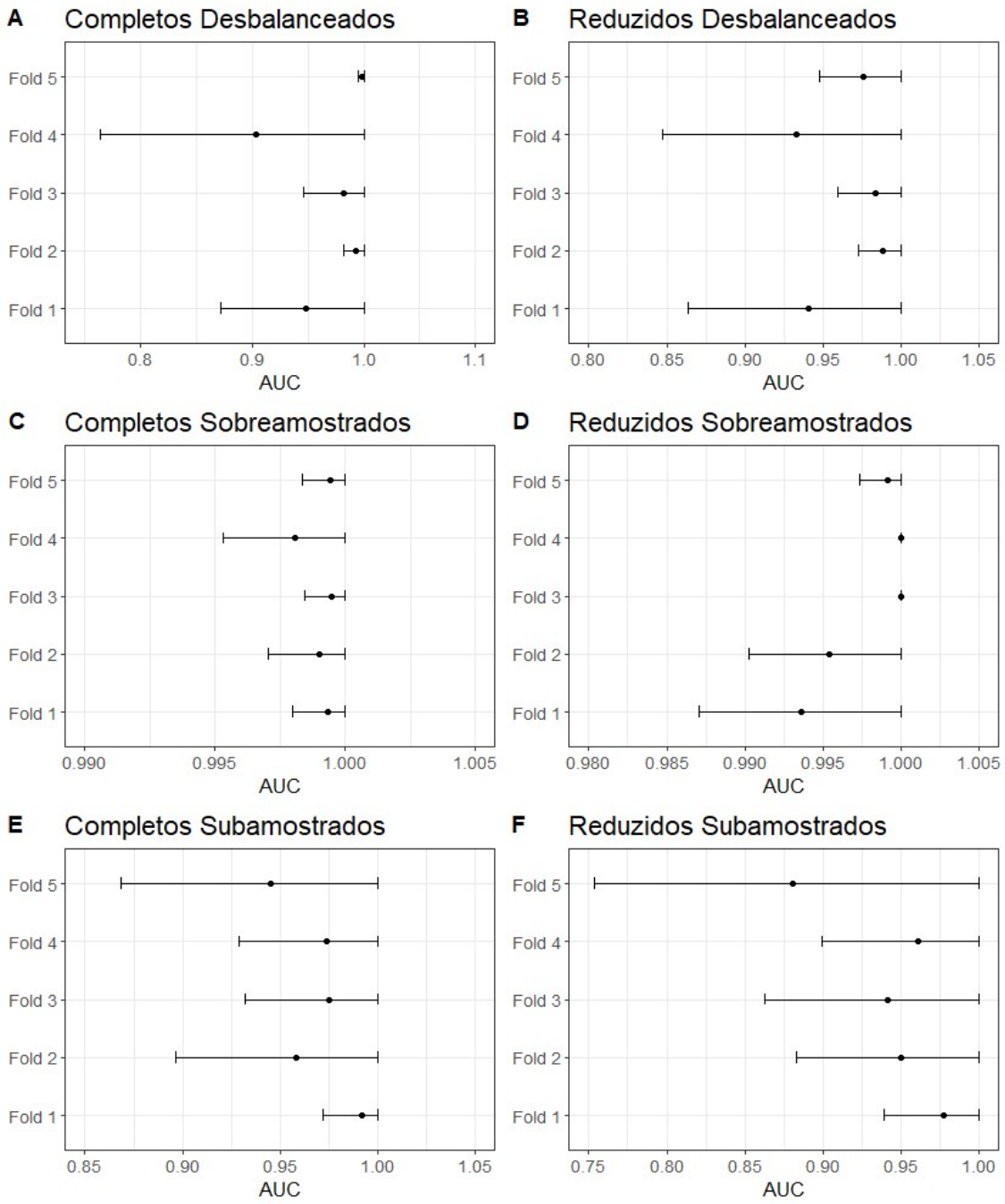

Figura 5.5: Área sob a curva ROC (AUC) dos modelos treinados com sequências do subtipo C e testados por validação cruzada 5-fold. O intervalo de confiança de 95% foi calculado a partir de 2000 amostras bootstraps.

A Figura 5.6 mostra a taxa de erro dos modelos durante a construção das árvores. É possível observar que os modelos sobreamostrados (completos ou reduzidos) obtiveram as menores taxas de erro, atingindo valores abaixo de 5%. Os modelos desbalanceados se

mantiveram com erro próximo a 5%, e os modelos subamostrados ficaram com suas taxas de erro próximas de 15%. Destaca-se que todos os modelos atingiram a estabilidade de seus índices de erro a partir de cerca de 150 árvores.

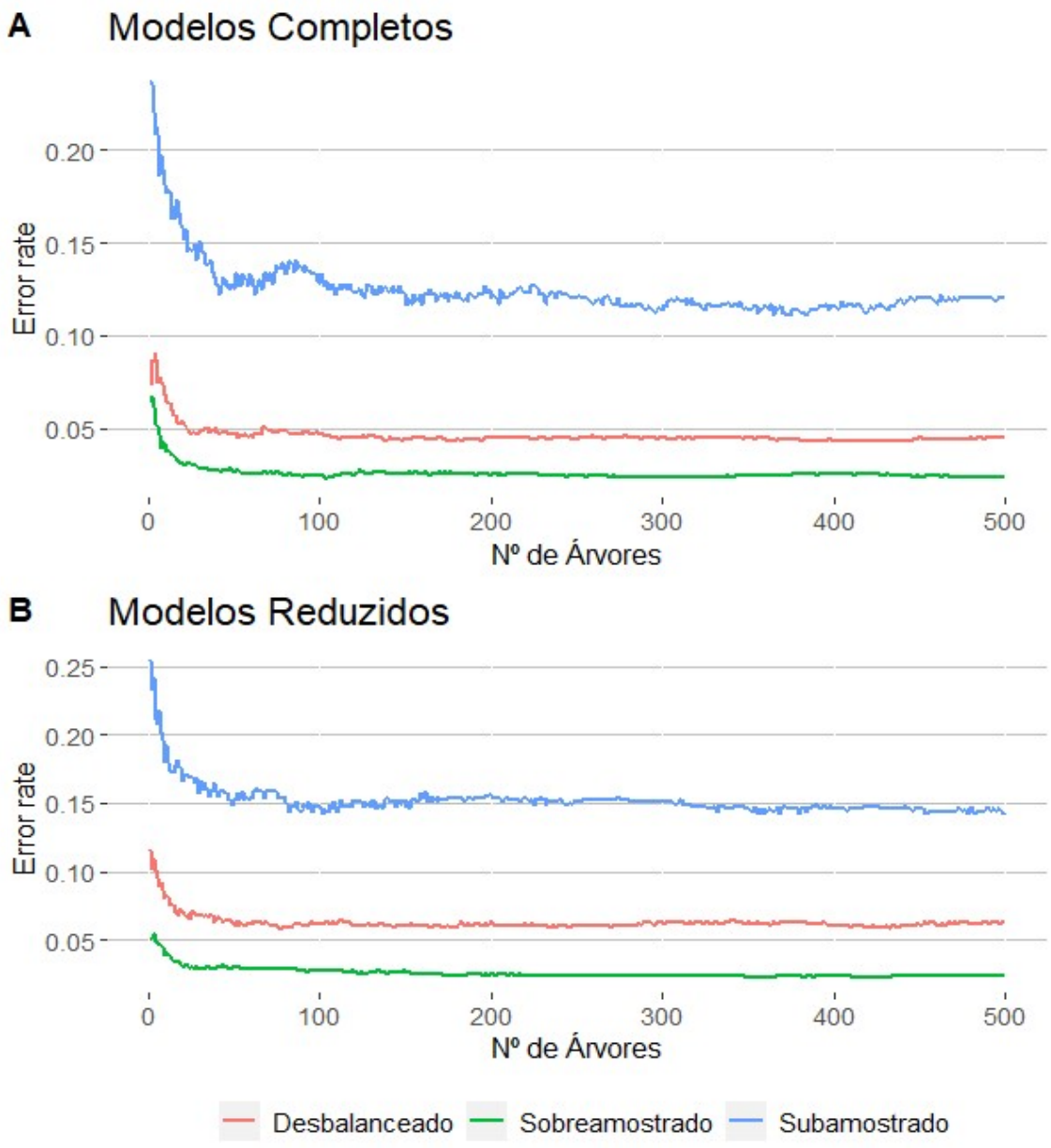

Figura 5.6 Média da taxa de erro para os conjuntos out-of-bag (OOB) ao longo da construção das árvores para os modelos do subtipo C.

5.2.2 Importância das variáveis

A Figura 5.7 mostra a frequência com qual cada posição foi destacada entre as cinco mais importantes nos *folds* para os modelos completos. As posições 5, 11, 13, 16, 19 e 25

apareceram, pelo menos uma vez, em todos os modelos. Destas posições, as posições 11 e 13 aparecem nos cinco *folds* de cada modelo. No modelo sem balanceamento, também apareceram as posições 5 e 25 (cinco vezes), 18 (duas vezes) e 19 (três vezes). Para o modelo sobreamostrado, aparecem as variáveis 25 (cinco vezes), 5 e 16 (quatro vezes) e 19 (duas vezes). No modelo subamostrado, além das variáveis 11 e 13, as variáveis 5 e 16 (cinco vezes), 25 (duas vezes), e as variáveis 9, 12 e 19 (uma vez cada).

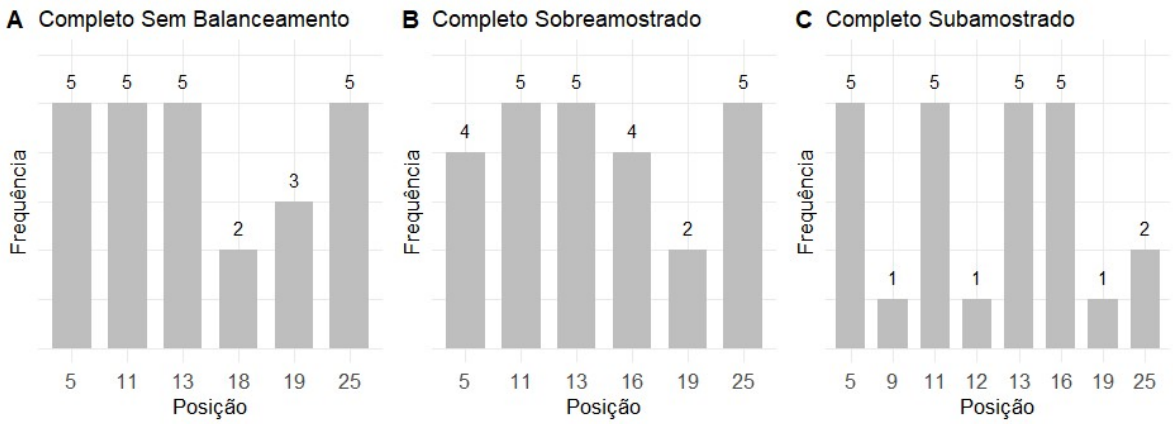

Figura 5.7 Distribuição das posições mais importantes de cada modelo completo, de acordo com o número de vezes que figuraram entre as cinco de maior importância de cada modelo construído.

Para os modelos reduzidos, de acordo com a Figura 5.8, as posições 11, 13, e 25 foram destacadas nos 5 *folds* como entre as mais importantes em todos os modelos. Para o conjunto desbalanceado, também se destacaram as posições 18 e 19, ambas aparecendo cinco vezes nos *folds*. No conjunto sobreamostrado, as posições 5 e 19 se juntam entre as mais importantes (cinco vezes cada), e no conjunto subamostrado, aparecem as posições 5 (cinco vezes), 19 (três vezes) e 18 (duas vezes) entre as mais importantes em algum dos *folds* treinados.

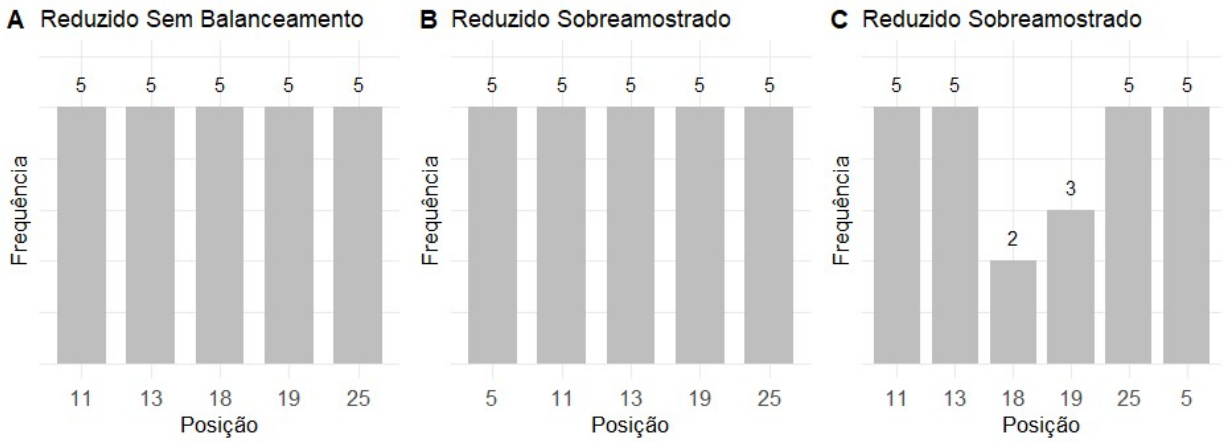

Figura 5.8 Distribuição das posições mais importantes de cada modelo, de acordo com o número de vezes que figuraram entre as cinco de maior importância de cada modelo construído.

5.2.3 Desempenho e comparação

De acordo com a Tabela 5.2, o modelo reduzido sobreamostrado (R-sobre) obteve as melhores medidas de desempenho para acurácia (99,27%), sensibilidade (99,60%) e MCC (0,989). O modelo completo e sobreamostrado (C-sobre) obteve a maior especificidade entre os modelos (99,38%).

Tabela 5.2: Desempenho dos modelos e classificadores estabelecidos a partir do cinco conjuntos de validação cruzada do subtipo C. O maior valor de cada medida de desempenho foi marcado em negrito.

Classificador	Acurácia (dp)(%)	Especificidade (dp) (%)	Sensibilidade (dp) (%)	MCC (dp)
C-desb	96,28 (2,48)	96,39 (3,00)	94,33 (6,87)	0,863 (0,077)
C-sobre	99,07 (0,43)	**99,38 (0,56)**	98,79 (0,83)	0,971 (0,009)
C-sub	92,92 (2,73)	90,62 (5,67)	94,16 (6,82)	0,859 (0,054)
R-desb	95,54 (1,67	96,49 (2,44)	89,54 (6,71)	0,824 (0,075)
R-sobre	**99,27 (0,59)**	99,02 (1,19)	**99,60 (1,18)**	**0,989 (0,011)**
R-sub	89,74 (3,5)	92,47 (8,0)	87,53 (9,1)	0,803 (0,065)
G2p 2.5%	93,75	98,13	66,67	0,720
G2p 5%	93,57	96,47	75,64	0,729
G2p 10%	90,00	90,46	87,18	0,668
T-CUP 2.0	92,68	93,98	84,62	0,725

dp: desvio-padrão

6 Discussão

Neste trabalho, desenvolvemos, no total, 12 modelos de predição baseado no algoritmo de *Random Forest* a fim de determinar o tropismo do HIV-1. Foram utilizadas 2.182 sequências peptídicas da alça V3 do vírus HIV-1 de subtipos B e C, com aminoácidos codificados numericamente pela escala de hidrofobicidade de Engelman.

É importante avaliar corretamente o desempenho de modelos, principalmente à luz do que pode acarretar erros demais por parte do classificador: um erro tipo I (predizer NR5 quando o tropismo é R5) significaria que uma medicação diferente deveria ser usada para controlar a carga viral em vez do Maraviroque®, já consagrado. A inefetividade poderia levar ao abandono do tratamento por parte do paciente que, em momento futuro, terá uma infecção disseminada e uma futura terapia com Maraviroque® não seria efetiva. Por outro lado, um erro tipo II (predizer R5 quando o tropismo é NR5) significaria que, além de o tratamento não funcionar, a abundância de partículas virais com tropismo X4 ou R5X4 significaria estágios avançados da infecção do HIV-1, uma vez que há uma "troca" de tropismo ao curso da doença (ALMEIDA, ZAPAROLI, *et al.*, 2014, SARMATI, PARISI, *et al.*, 2010). A construção dos modelos se diferenciou em três aspectos: (1) pelo subtipo viral (B ou C), (2) pela forma de reamostragem como técnica de balanceamento de dados (sem balanceamento, balanceamento por sobreamostragem ou subamostragem), e (3) pela remoção ou não de variáveis pouco significativas. Cada modelo apresentou um ponto de corte ótimo definido por meio das curvas ROC.

O presente trabalho também avaliou a importância das variáveis na construção dos modelos. Neste sentido, a importância de uma variável indica a capacidade de uma variável sozinha classificar corretamente as observações, obtendo o maior valor possível de acurácia.

Por fim, o desempenho dos modelos construídos foi comparado a um classificador já amplamente utilizado (geno2pheno$_{[coreceptor]}$) e a outro classificador estabelecido, que utiliza o mesmo algoritmo de *Random Forest* (T-CUP 2.0). A primeira comparação serviu para mostrar se nosso trabalho poderia competir com o estado da arte, observando se o algoritmo tem boa aplicabilidade para a questão. A segunda comparação mostrou que, mesmo dentro do algoritmo de *Random Forest*, é possível modificar as condições de uso e melhorar o desempenho dos classificadores.

Tanto o algoritmo de SVM (usado no geno2pheno$_{[coreceptor]}$) quanto o algoritmo de *Random Forest* são conceitualmente adequados para basear classificadores de tropismo do HIV-1. Ambos são algoritmos de aprendizado supervisionado que buscam soluções matemáticas para classificação binária das observações treinadas. Ainda assim, o mecanismo da *Random Forest* é, no geral, mais facilmente compreensível (uma separação encadeada e lógica dos grupos até que cheguem à observação final) que o da máquina de vetores de suporte (uma função com diversas variáveis independentes que, em certo ponto, é impossível de ser representada graficamente, o que dificulta o entendimento e possíveis ajustes finos do modelo).

As formas de amostragem testadas aqui foram as mais simples, porém, adequadas aos dados utilizados. Técnicas mais avançadas de balanceamento, como a SMOTE (*Synthetic Minority Over-sampling Technique,* Técnica de Sobreamostragem de Minoria Sintética) (CHAWLA, BOWYER, *et al.*, 2002), apesar de sua popularidade, podem não ser adequadas à proposta do modelo. A abordagem funciona criando observações para o conjunto desbalanceado, calculando novos valores para cada variável. Neste caso, introduzir novos valores significaria a introdução de novos aminoácidos. Apesar de a técnica fazer sentido matematicamente, ela faria com que o modelo destoasse do meio natural.

A remoção de variáveis pouco informativas é um passo geralmente recomendado para a construção de modelos que utilizam algum algoritmo preditivo. Em modelos que utilizam dados com grande quantidade de variáveis, esta etapa é fundamental para reduzir o ruído de variáveis redundantes ou irrelevantes para a predição, o que "diminui o tempo de computação, melhora a acurácia do aprendizado e facilita uma melhor compreensão para o modelo de aprendizado ou para os dados" (CAI, LUO, *et al.*, 2018). No caso da *Random Forest*, quando há esta limpeza dos preditores, as árvores são construídas com uma quantidade menor de variáveis, reduzindo o ruído causado por nós pouco relevantes para a divisão dos grupos. Em geral, isso torna o modelo mais parcimonioso, melhorando-o em aspectos como exigência computacional, além de destacas os pontos de corte de cada variável com maior clareza.

Uma diferença importante entre os modelos construídos e os classificadores já estabelecidos foi a maneira de converter os aminoácidos em dados numéricos. Os classificadores já estabelecidos usam escala de hidrofobicidade de Kyte-Dollittle (1982),

em vez da escala de Engelman. Ao contrário da escala escolhida para este trabalho, a escala Kyte-Dollittle apresenta valores iguais para alguns aminoácidos, o que pode mascarar possíveis correspondências de aminoácidos com o tropismo viral. Avaliar se a mudança de escalas de fato altera o desempenho dos modelos configura uma possibilidade para eventuais estudos.

Com exceção dos modelos subamostrados, todos os modelos construídos mostraram estabilidade no índice de erro a partir da ducentésima árvore, aproximadamente. Esta informação pode indicar que um classificador baseado em *Random Forest* teria desempenho parecido, tendo 200 ou 500 árvores. É um aspecto interessante a ser ressaltado, uma vez que a construção de uma *Random Forest* com o mínimo de árvores pode reduzir a demanda computacional para executar o classificador.

Em relação à importância das variáveis, os modelos construídos concordaram com a literatura estabelecida: as posições 11 e 13 da sequência figuraram com grande destaque, para ambos os subtipos, aparecendo entre as cinco mais importantes em todos os modelos (e *folds*) construídos. As posições 18 e 25 também receberam destaque importante, aparecendo em muitos dos modelos.

Esse destaque concorda, de uma forma ou de outra, com todos os classificadores revisados neste trabalho, mostrando a consistência dos modelos para avaliar os dados. Xiao et al (1998) mostraram que as células CCR5-positivas eram infectadas por partículas virais que possuíam, na posição 11, serina ou glicina (aminoácidos neutros), e, na posição 25, ácido aspártico ou ácido glutâmico (aminoácidos negativos).

Segundo Monno et al (2011) e Shen et al (2016), caso a posição 18 estivesse ocupada pela arginina (aminoácido positivamente carregado), o vírus teria tropismo NR5 e, se ocupado pela glutamina (aminoácido neutro) ou pela lisina (aminoácido positivamente carregado), o tropismo seria R5. Apesar de não ter maiores explicações, fora relatado que a arginina poderia possuir um papel específico na interação entre a alça V3 e o receptor CXCR4. Shen et al (2016) também mostra que a posição 13 é ocupada pelo aminoácido histidina quando a sequência corresponde a um vírus de tropismo NR5.

De maneira peculiar, a posição 24 (para o subtipo B) e a variável 16 (para o subtipo C) mostraram-se como as mais importantes para a construção dos modelos treinados com a sequência completa, mas foram excluídas durante a remoção automática.

Cardozo et al (2007) relataram que o aminoácido da posição 25 somente ocuparia o espaço da interação da alça com o receptor se o aminoácido da posição 24 fosse a glicina, o único aminoácido sem cadeia lateral. Caso contrário, o aminoácido da posição 24 seria o fator decisório para a determinação do tropismo do HIV-1. Daí a importância que a posição possui, mesmo que não sofra muitas variações e seja excluído dos modelos reduzidos.

Quanto à posição 16, Bagnarelli et al (2003) relataram, sem maiores detalhes, que variações na posição 16 levam a mudanças no tropismo viral. O estudo informa que, alterando o aminoácido de glicina (aminoácido da sequência de consenso) para leucina, é possível mudar o tropismo de R5 para X4.

Uma variável que se mostrou importante para os subtipos C, mas não recebera destaque na literatura foi a variável 19. As posições 5 e 12 também foram destacadas no modelo completo e subamostrado, mas pode estar relacionado à maior taxa de erro OOB e maior ruído que este modelo apresentou (Figura 5.6-A).

Quanto ao desempenho, os modelos construídos neste trabalho, treinados com sequências de um único subtipo, obtiveram desempenho tão bom ou superior em relação aos modelos já estabelecidos, tanto para o subtipo B quanto para o subtipo C. Mesmo assim, a melhoria nas medidas talvez não justifique a separação das sequências, visto que um novo classificador deve abranger o maior número de subtipos possível.

Os modelos propostos foram tão robustos quanto os modelos já experimentados por Dybowski et al (2010), Heider et al (2014) e Chen et al (2019). As AUCs dos modelos de subtipo B não mostraram grandes diferenças, e todas obtiveram valores próximos a 0,95, mostrando um bom desempenho preditivo.

Houve pouca diferença entre o desempenho e a robustez dos modelos com todas as posições de sequências e dos modelos com variáveis selecionadas. Essa informação, contudo, dá mais crédito aos modelos reduzidos, uma vez que o modelo com menor número de variáveis pode trabalhar com menor demanda computacional. Em larga escala, esta diferença é muito importante, evitando consumir toda a capacidade de memória do computador do usuário.

Entre os modelos comparados durante o teste de subtipo B, os desempenhos foram muito parelhos entre si. Contudo, os modelos que obtiveram pior desempenho geral foram

os modelos geno2pheno$_{[coreceptor]}$. Embora tenham se sobressaído em relação à especificidade, a sensibilidade foi muito inferior em relação aos outros modelos.

O número de sequências usadas para treinar os modelos também pode ser dado como um diferencial para comparar os modelos testados. Enquanto os modelos construídos foram treinados com 1.136 sequências únicas (somente de subtipo B), o geno2pheno foi treinado com 1.100 sequências de vários subtipos e o T-CUP 2.0, com 1.351 sequências de diversos subtipos (sendo cerca de 650 sequências de subtipo B). É possível que este aspecto seja o suficiente para a melhora do desempenho do classificador vista na Tabela 4.3.

Além disso, uma diferença importante entre os modelos construídos e os classificadores estabelecidos foi a ausência de uma variável que verse sobre a carga líquida da sequência peptídica, presente em trabalhos como os de Xu et al (2007) e Heider et al (2014). Mesmo assim, o desempenho foi bastante elevado, colocando em xeque a necessidade da carga líquida da sequência como variável relevante.

Os modelos treinados com o subtipo C receberam uma quantidade muito menor de sequências que os com o subtipo B. A base de dados de Los Alamos possui uma quantidade muito menor deste subtipo, uma vez que ele não é abundante na Europa ou na América do Norte. Como explica Bbosa et al (2019), isso é um problema sério, uma vez que o subtipo C abrange mais da metade dos casos e pouco se conhece sobre a sequência peptídica da alça V3 desta linhagem viral.

As curvas ROC construídas nos modelos com subtipo C obtiveram valores muito divergentes entre si. Os modelos cujo conjunto de treinamento passaram por subamostragem (tanto o modelo reduzido quanto o completo) obtiveram os valores de AUC mais variantes entre todos, seguidos dos desbalanceados e dos sobreamostrados. Essa diferença pode ser explicada pela baixa quantidade de sequências de nosso banco de dados.

Mesmo com uma quantidade baixa de dados, os modelos obtiveram desempenho muito alto, se comparado com os classificadores estabelecidos. Destaca-se que o classificador T-CUP 2.0 teve um desempenho parecido com os modelos construídos somente com as sequências de subtipo C.

A exemplo do subtipo B, o classificador com desempenho inferior foi o geno2pheno. O desempenho observado coincide com o que já era conhecido na literatura.

Riemenschneider et al (2016), revisando os principais classificadores genotípicos de tropismo de HIV-1 para o subtipo C, encontrou valores próximos de sensibilidade do geno2pheno para predizer o tropismo NR5 corretamente.

O desempenho e a robustez dos modelos sobreamostrados para o subtipo C, como mostrados na Figura 5.5-C e D e na Tabela 5.2, foi surpreendentemente alto, próximo da perfeição. Porém, ao analisar a construção destas *random forests*, pode-se estar diante de um problema conhecido como *data leakage*, quando dados que não deveriam ser usados no conjunto de treinamento são usados para treinar o modelo e, geralmente, levam ao aumento indevido do desempenho dele (BROWNLEE, 2016). No caso deste trabalho, o evento que pode ter desencadeado o *data leakage* foi a sobreamostragem do conjunto de sequências de subtipo C. Como o desbalanço dos dados seguia a proporção de 6:1, a reamostragem criou pelo menos 5 cópias de cada sequência e, durante o processamento dos modelos por validação cruzada, é possível que sequências iguais tenham sido usadas tanto no conjunto de treinamento do modelo quanto no conjunto de teste. Assim, o altíssimo desempenho destes modelos pode ser fruto da leitura de observações já existentes no modelo, o que, decerto, não aconteceria em uma situação real.

Outro aspecto que pode ter enviesado a análise de desempenho do subtipo C é a pequena quantidade de sequências designada para o conjunto de teste para cada *fold*. Com tão poucas observações no conjunto, cada erro ou acerto tem um peso muito grande no valor final das medidas de desempenho e, por isso, resultando em grandes variações entre os *folds*.

7 Conclusão

Entende-se com este trabalho que, embora existam formas automatizadas de qualificar variáveis de um conjunto de dados para treinamento de um modelo, é muito importante verificarmos empiricamente seu valor. Como foi mostrado, mesmo uma variável que tenha predominância de um único valor, ainda pode possuir relevância para o modelo. No caso do balanceamento de observações, não foi observado ganho significativo com a adição da etapa, seja através da sobreamostragem ou subamostragem.

A observação de modelos construídos com maior desempenho que os classificadores já conhecidos da literatura mostra que há espaço para melhorias nos classificadores de tropismo de HIV-1, especialmente no que tange à detecção do tropismo X4. De fato, com a contínua coleta de amostras para o sequenciamento do HIV, mais dados para alimentar os classificadores poderiam levar ao aumento das medidas de desempenho.

É importante ressaltar que há um viés de seleção para o subtipo das partículas virais sequenciadas na base de dados utilizada neste trabalho. Sendo assim, mesmo com a grande prevalência de infecções do subtipo C, são poucas as amostras deste tipo. É necessário observar por maior esforço para angariar estas sequências e trazer novas informações aos classificadores.

8 Referências bibliográficas

ABBAS, A. K., LICHTMAN, A. H., PILLAI, S., *et al.* **Cellular and molecular immunology**. Ninth edition ed. Philadelphia, PA, Elsevier, 2018.

ALMEIDA, F. J., ZAPAROLI, M. S., MOREIRA, D. H., *et al.* "Association of X4 tropism with disease progression in antiretroviral-treated children and adolescents living with HIV/AIDS in São Paulo, Brazil", **The Brazilian Journal of Infectious Diseases**, v. 18, n. 3, p. 300–307, maio 2014. DOI: 10.1016/j.bjid.2013.10.002. .

ALVES, B. M., SIQUEIRA, J. D., PRELLWITZ, I. M., *et al.* "Estimating HIV-1 Genetic Diversity in Brazil Through Next-Generation Sequencing", **Frontiers in Microbiology**, v. 10, p. 749, 9 abr. 2019. DOI: 10.3389/fmicb.2019.00749. .

ARRUDA, M. B., BOULLOSA, L. T., CARDOSO, C. C., *et al.* "Brazilian network for HIV Drug Resistance Surveillance (HIV-BresNet): a survey of treatment-naive individuals", **Journal of the International AIDS Society**, v. 21, n. 3, p. e25032, mar. 2018. DOI: 10.1002/jia2.25032. .

BAGNARELLI, P., FIORELLI, L., VECCHI, M., *et al.* "Analysis of the functional relationship between V3 loop and gp120 context with regard to human immunodeficiency virus coreceptor usage using naturally selected sequences and different viral backbones", **Virology**, v. 307, n. 2, p. 328–340, mar. 2003. DOI: 10.1016/S0042-6822(02)00077-6. .

BALTIMORE, D. "Expression of animal virus genomes", **Bacteriological Reviews**, v. 35, n. 3, p. 235–241, set. 1971. .

BARRE-SINOUSSI, F., CHERMANN, J., REY, F., *et al.* "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)", **Science**, v. 220, n. 4599, p. 868–871, 20 maio 1983. DOI: 10.1126/science.6189183. .

BBOSA, N., KALEEBU, P., SSEMWANGA, D. "HIV subtype diversity worldwide", **Current Opinion in HIV and AIDS**, v. 14, n. 3, p. 153–160, maio 2019. DOI: 10.1097/COH.0000000000000534. .

BEERENWINKEL, N. "Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes", **Nucleic Acids Research**, v. 31, n. 13, p. 3850–3855, 1 jul. 2003. DOI: 10.1093/nar/gkg575. .

BERGER, E. A., DOMS, R. W., FENYÖ, E.-M., *et al.* "A new classification for HIV-1", **Nature**, v. 391, n. 6664, p. 240–240, jan. 1998. DOI: 10.1038/34571. .

BISWAS, P., TAMBUSSI, G., LAZZARIN, A. "Access denied? The status of co-receptor inhibition to counter HIV entry", **Expert Opinion on Pharmacotherapy**, v. 8, n. 7, p. 923–933, maio 2007. DOI: 10.1517/14656566.8.7.923. .

BRASIL, M. da S. do. **Boletim Epidemiológico de HIV/Aids 2019**. **Departamento de Doenças de Condições Crônicas e Infecções Sexualmente Transmissíveis**. Brasília, Departamento de Doenças de Condições Crônicas e Infecções Sexualmente

Transmissíveis, 28 nov. 2019. Disponível em: http://www.aids.gov.br/pt-br/pub/2019/boletim-epidemiologico-de-hivaids-2019.

BREIMAN, L. "Bagging predictors", **Machine Learning**, v. 24, n. 2, p. 123–140, ago. 1996. DOI: 10.1007/BF00058655. .

BREIMAN, L. **Classification and regression trees**. [S.l: s.n.], 2017. Disponível em: http://www.myilibrary.com?id=1043565. Acesso em: 6 maio 2020.

BREIMAN, L. "Random Forests", **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. DOI: 10.1023/A:1010933404324. .

BROMAN, K. W., BROMAN, A. T. **broman: Karl Broman's R Code**. [S.l: s.n.], 2019. Disponível em: https://CRAN.R-project.org/package=broman.

BROWNLEE, J. **Data Leakage in Machine Learning**. **Machine Learning Mastery**. [S.l: s.n.]. Disponível em: https://machinelearningmastery.com/data-leakage-machine-learning/. Acesso em: 13 set. 2020. , 1 ago. 2016

CAI, J., LUO, J., WANG, S., *et al.* "Feature selection in machine learning: A new perspective", **Neurocomputing**, v. 300, p. 70–79, 2018. DOI: https://doi.org/10.1016/j.neucom.2017.11.077. .

CANN, A. **Principles of Molecular Virology**. Fifth ed. Waltham, USA, Elsevier Ltd, 2012.

CARDOZO, T., KIMURA, T., PHILPOTT, S., *et al.* "Structural Basis for Coreceptor Selectivity by The HIV Type 1 V3 Loop", **AIDS Research and Human Retroviruses**, v. 23, n. 3, p. 415–426, mar. 2007. DOI: 10.1089/aid.2006.0130. .

CHARIF, D., LOBRY, J. R., "SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis.". In: BASTOLLA, U., PORTO, M., ROMAN, H. E., *et al.* (Org.), **Structural approaches to sequence evolution: Molecules, networks, populations**, Biological and Medical Physics, Biomedical Engineering. New York, Springer Verlag, 2007. p. 207–232.

CHAWLA, N. V., BOWYER, K. W., HALL, L. O., *et al.* "SMOTE: Synthetic Minority Over-sampling Technique", **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 1 jun. 2002. DOI: 10.1613/jair.953. .

CHEN, X., WANG, Z.-X., PAN, X.-M. "HIV-1 tropism prediction by the XGboost and HMM methods", **Scientific Reports**, v. 9, n. 1, p. 9997, dez. 2019. DOI: 10.1038/s41598-019-46420-4. .

CHIU, D. T., DUESBERG, P. H. "The toxicity of azidothymidine (AZT) on human and animal cells in culture at concentrations used for antiviral therapy", **Genetica**, v. 95, p. 7, 1995. .

CLAPHAM, P. R., MCKNIGHT, Á. "Cell surface receptors, virus entry and tropism of primate lentiviruses", **Journal of General Virology**, v. 83, n. 8, p. 1809–1829, 1 ago. 2002. DOI: 10.1099/0022-1317-83-8-1809. .

CLAPHAM, P. R., MCKNIGHT, Á. "HIV-1 receptors and cell tropism", **British Medical Bulletin**, v. 58, n. 1, p. 43–59, 1 set. 2001. DOI: 10.1093/bmb/58.1.43. .

CORTES, C., VAPNIK, V. "Support-vector networks", **Machine Learning**, v. 20, n. 3, p. 273–297, set. 1995. DOI: 10.1007/BF00994018. .

DE CLERCQ, E., NEYTS, J., "Antiviral Agents Acting as DNA or RNA Chain Terminators". In: KRÄUSSLICH, H.-G., BARTENSCHLAGER, R. (Org.), **Antiviral Strategies**, Handbook of Experimental Pharmacology. Berlin, Heidelberg, Springer Berlin Heidelberg, 2009. v. 189. p. 53–84. DOI: 10.1007/978-3-540-79086-0_3. Disponível em: http://link.springer.com/10.1007/978-3-540-79086-0_3. Acesso em: 14 mar. 2020.

DE JONG, J. J., DE RONDE, A., KEULEN, W., *et al.* "Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution.", **Journal of Virology**, v. 66, n. 11, p. 6777–6780, 1992. DOI: 10.1128/JVI.66.11.6777-6780.1992. .

DENG, H., LIU, R., ELLMEIER, W., *et al.* "Identification of a major co-receptor for primary isolates of HIV-1", **Nature**, v. 381, n. 6584, p. 661–666, 20 jun. 1996. DOI: 10.1038/381661a0. .

DIETTERICH, T. "Overfitting and undercomputing in machine learning", **ACM Computing Surveys (CSUR)**, v. 27, n. 3, p. 326–327, set. 1995. DOI: 10.1145/212094.212114. .

DRAGIC, T., LITWIN, V., ALLAWAY, G. P., *et al.* "HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5", **Nature**, v. 381, n. 6584, p. 667–673, 20 jun. 1996. DOI: 10.1038/381667a0. .

DYBOWSKI, J. N., HEIDER, D., HOFFMANN, D. "Prediction of Co-Receptor Usage of HIV-1 from Genotype", **PLoS Computational Biology**, v. 6, n. 4, p. e1000743, 15 abr. 2010. DOI: 10.1371/journal.pcbi.1000743. .

EFRON, B., TIBSHIRANI, R. **An introduction to the bootstrap**. New York, Chapman & Hall, 1993. (Monographs on statistics and applied probability, 57).

ENGELMAN, D. M., STEITZ, A., GOLDMAN, A. "Identifying Nonpolar Transbilayer Helices in Amino Acid Sequences of Membrane Proteins", v. 15, p. 35, 1986. .

FANALES-BELASIO, E., RAIMONDO, M., SULIGOI, B., *et al.* "HIV virology and pathogenetic mechanisms of infection: a brief overview", **Annali dell'Istituto Superiore Di Sanita**, v. 46, n. 1, p. 5–14, 2010. DOI: 10.4415/ANN_10_01_02. .

FOLEY, B. T., KORBER, B. T. M., LEITNER, T. K., *et al.* **HIV sequence compendium 2018**. . [S.l.], Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2018.

FOUCHIER, R. A. M., GROENINK, M., KOOTSTRA, N. A., *et al.* "Phenotype-Associated Sequence Variation in the Third Variable Domain of the Human Immunodeficiency Virus Type 1 gpl20 Molecule", **J. VIROL.**, v. 66, p. 5, 1992. .

FRANKEL, A. D., YOUNG, J. A. T. "HIV-1: Fifteen Proteins and an RNA", **Annual Review of Biochemistry**, v. 67, n. 1, p. 1–25, jun. 1998. DOI: 10.1146/annurev.biochem.67.1.1. .

GAC BLOOD. "Human Immunodeficiency Virus (HIV)", **Transfusion Medicine and Hemotherapy**, v. 43, n. 3, p. 203–222, 2016. DOI: 10.1159/000445852. .

GAGOLEWSKI, M. **R package stringi: Character string processing facilities**. [S.l: s.n.], 2019. Disponível em: http://www.gagolewski.com/software/stringi/.

GALLO, R., SARIN, P., GELMANN, E., *et al.* "Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS)", **Science**, v. 220, n. 4599, p. 865–867, 20 maio 1983. DOI: 10.1126/science.6601823. .

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. **The Elements of Statistical Learning**. New York, NY, Springer New York, 2009. Disponível em: http://link.springer.com/10.1007/978-0-387-84858-7. Acesso em: 4 maio 2020. (Springer Series in Statistics).

HEALTH, U. D. of, SERVICES, H., OTHERS. "Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents", **http://aidsinfo. nih. gov/OrderPublication/OrderPubsBrowseSearchResultsTable. aspx? ID= 115**, 2009. .

HEIDER, D. **Interpol: Interpolation of amino acid sequences**. [S.l: s.n.], 2012. Disponível em: https://CRAN.R-project.org/package=Interpol.

HEIDER, D., DYBOWSKI, J. N. **TCUP2: Two-level Co-receptor Usage Prediction**. [S.l: s.n.], 2013.

HEIDER, D., DYBOWSKI, J. N., WILMS, C., *et al.* "A simple structure-based model for the prediction of HIV-1 co-receptor tropism", **BioData Mining**, v. 7, n. 1, p. 14, dez. 2014. DOI: 10.1186/1756-0381-7-14. .

IVANOFF, L. A., DUBAY, J. W., MORRIS, J. F., *et al.* "V3 Loop region of the HIV-1 gpl20 envelope protein is essential for virus infectivity", **Virology**, v. 187, n. 2, p. 423–432, abr. 1992. DOI: 10.1016/0042-6822(92)90444-T. .

JAMES, G., WITTEN, D., HASTIE, T., *et al.* **An Introduction to Statistical Learning**. New York, NY, Springer New York, 2013. v. 103. Disponível em: http://link.springer.com/10.1007/978-1-4614-7138-7. Acesso em: 8 maio 2020. (Springer Texts in Statistics).

KAGAN, R. M., JOHNSON, E. P., SIAW, M. F., *et al.* "Comparison of Genotypic and Phenotypic HIV Type 1 Tropism Assay: Results from the Screening Samples of Cenicriviroc Study 202, a Randomized Phase II Trial in Treatment-Naive Subjects",

AIDS Research and Human Retroviruses, v. 30, n. 2, p. 151–159, fev. 2014. DOI: 10.1089/aid.2013.0123. .

KANMOGNE, G., WOOLLARD, S. "Maraviroc: a review of its use in HIV infection and beyond", **Drug Design, Development and Therapy**, p. 5447, out. 2015. DOI: 10.2147/DDDT.S90580. .

KASSAMBARA, A. **ggpubr: "ggplot2" Based Publication Ready Plots**. [S.l: s.n.], 2020. Disponível em: https://CRAN.R-project.org/package=ggpubr.

KUHN, M., WESTON, S., WILLIAMS, A., *et al.* **caret: Classification and Regression Training**. [S.l: s.n.], 2019. Disponível em: https://CRAN.R-project.org/package=caret.

KYTE, J., DOOLITTLE, R. F. "A simple method for displaying the hydropathic character of a protein", **Journal of Molecular Biology**, v. 157, n. 1, p. 105–132, 1982. DOI: https://doi.org/10.1016/0022-2836(82)90515-0. .

LENGAUER, T., SANDER, O., SIERRA, S., *et al.* "Bioinformatics prediction of HIV coreceptor usage", **Nature Biotechnology**, v. 25, n. 12, p. 1407–1410, dez. 2007. DOI: 10.1038/nbt1371. .

LEVY, J. A., KAMINSKY, L. S., MORROW, W. J. W., *et al.* "Infection by the Retrovirus Associated With the Acquired Immunodeficiency Syndrome: Clinical, Biological, and Molecular Features", **Journal of Urology**, v. 136, n. 2, p. 546–547, ago. 1986. DOI: 10.1016/S0022-5347(17)44958-5. .

LIAW, A., WIENER, M. "Classification and Regression by randomForest", **R News**, v. 2, n. 3, p. 18–22, 2002. .

LIEBERMAN-BLUM, S. S., FUNG, H. B., BANDRES, J. C. "Maraviroc: A CCR5-receptor antagonist for the treatment of HIV-1 infection", **Clinical Therapeutics**, v. 30, n. 7, p. 1228–1250, jul. 2008. DOI: 10.1016/S0149-2918(08)80048-3. .

LIN, N. H., KURITZKES, D. R. "Tropism testing in the clinical management of HIV-1 infection:", **Current Opinion in HIV and AIDS**, v. 4, n. 6, p. 481–487, nov. 2009. DOI: 10.1097/COH.0b013e328331b929. .

LOW, A. J., MCGOVERN, R. A., HARRIGAN, P. R. "Trofile HIV co-receptor usage assay", **Expert Opinion on Medical Diagnostics**, v. 3, n. 2, p. 181–191, mar. 2009. DOI: 10.1517/17530050802708981. .

MADDON, P. J., DALGLEISH, A. G., MCDOUGAL, J. S., *et al.* "The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain", **Cell**, v. 47, n. 3, p. 333–348, 1986. .

MCDOUGAL, J., KENNEDY, M., SLIGH, J., *et al.* "Binding of HTLV-III/LAV to T4+ T cells by a complex of the 110K viral protein and the T4 molecule", **Science**, v. 231, n. 4736, p. 382–385, 24 jan. 1986. DOI: 10.1126/science.3001934. .

MILICH, L., MARGOLIN, B., SWANSTROM, R. "V3 loop of the human immunodeficiency virus type 1 Env protein: interpreting sequence variability.", **Journal of Virology**, v. 67, n. 9, p. 5623–5634, 1993. DOI: 10.1128/JVI.67.9.5623-5634.1993. .

MONNO, L., SARACINO, A., SCUDELLER, L., *et al.* "Impact of Mutations Outside the V3 Region on Coreceptor Tropism Phenotypically Assessed in Patients Infected with HIV-1 Subtype B", **Antimicrobial Agents and Chemotherapy**, v. 55, n. 11, p. 5078–5084, nov. 2011. DOI: 10.1128/AAC.00743-11. .

MOORE, J. P., TRKOLA, A., DRAGIC, T. "Co-receptors for HIV-1 entry", **Current Opinion in Immunology**, v. 9, n. 4, p. 551–562, ago. 1997. DOI: 10.1016/s0952-7915(97)80110-0. .

MUSHAHWAR, I. K., "Human Immunodeficiency Viruses: Molecular Virology, Pathogenesis, Diagnosis and Treatment". **Perspectives in Medical Virology**, [S.l.], Elsevier, 2006. v. 13. p. 75–87. DOI: 10.1016/S0168-7069(06)13005-0. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0168706906130050. Acesso em: 15 jan. 2020.

OSORIO, D., RONDON-VILLARREAL, P., TORRES, R. "Peptides: A Package for Data Mining of Antimicrobial Peptides", **The R Journal**, v. 7, n. 1, p. 4–14, 2015. .

PAGÈS, H., ABOYOUN, P., GENTLEMAN, R., *et al.* **Biostrings: Efficient manipulation of biological strings**. [S.l: s.n.], 2018.

POVEDA, E., ALCAMÍ, J., PAREDES, R., *et al.* "Genotypic Determination of HIV Tropism - Clinical and Methodological Recommendations to Guide the Therapeutic Use of CCR5 Antagonists", **AIDS Reviews.**, p. 14, 2010. .

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, R Foundation for Statistical Computing, 2018. Disponível em: https://www.R-project.org/.

RESCH, W., HOFFMAN, N., SWANSTROM, R. "Improved Success of Phenotype Prediction of the Human Immunodeficiency Virus Type 1 from Envelope Variable Loop 3 Sequence Using Neural Networks", **Virology**, v. 288, n. 1, p. 51–62, set. 2001. DOI: 10.1006/viro.2001.1087. .

RICHMAN, D., FISCHL, M., GRIECO, M., *et al.* "The toxicity of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. A double-blind, placebo-controlled trial", 1987. DOI: doi.org/10.1056/NEJM198707233170402. Disponível em: https://www.nejm.org/doi/full/10.1056/NEJM198707233170402?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dpubmed.

RIEMENSCHNEIDER, M., CASHIN, K. Y., BUDEUS, B., *et al.* "Genotypic Prediction of Co-receptor Tropism of HIV-1 Subtypes A and C", **Scientific Reports**, v. 6, n. 1, p. 24883, abr. 2016. DOI: 10.1038/srep24883. .

ROBIN, X., TURCK, N., HAINARD, A., *et al.* "pROC: an open-source package for R and S+ to analyze and compare ROC curves", **BMC Bioinformatics**, v. 12, p. 77, 2011. .

SANDER, O., SING, T., SOMMER, I., *et al.* "Structural Descriptors of gp120 V3 Loop for the Prediction of HIV-1 Coreceptor Usage", **PLoS Computational Biology**, v. 3, n. 3, p. 10, 2007. .

SARMATI, L., PARISI, S. G., ANDREONI, C., *et al.* "Switching of Inferred Tropism Caused by HIV during Interruption of Antiretroviral Therapy", **Journal of Clinical Microbiology**, v. 48, n. 7, p. 2586–2588, 2010. DOI: 10.1128/JCM.02125-09. .

SATTENTAU, Q J, MOORE, J. P. "Conformational changes induced in the human immunodeficiency virus envelope glycoprotein by soluble CD4 binding.", **The Journal of Experimental Medicine**, v. 174, n. 2, p. 407–415, 1 ago. 1991. DOI: 10.1084/jem.174.2.407. .

SATTENTAU, Quentin J, MOORE, J. P. "The role of CD4 in HIV binding and entry", **Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences**, v. 342, n. 1299, p. 59–66, 29 out. 1993. DOI: 10.1098/rstb.1993.0136. .

SHARON, M., KESSLER, N., LEVY, R., *et al.* "Alternative Conformations of HIV-1 V3 Loops Mimic β Hairpins in Chemokines, Suggesting a Mechanism for Coreceptor Selectivity", **Structure**, v. 11, n. 2, p. 225–236, fev. 2003. DOI: 10.1016/S0969-2126(03)00011-X. .

SHEN, H.-S., YIN, J., LENG, F., *et al.* "HIV coreceptor tropism determination and mutational pattern identification", **Scientific Reports**, v. 6, n. 1, p. 21280, ago. 2016. DOI: 10.1038/srep21280. .

SHPAER, E. G., DELWART, E. L., KUIKEN, C. L., *et al.* "Conserved V3 Loop Sequences and Transmission of Human Immunodeficiency Virus Type 1", **AIDS Research and Human Retroviruses**, v. 10, n. 12, p. 1679–1684, dez. 1994. DOI: 10.1089/aid.1994.10.1679. .

STAMBUK, D., YOULE, M., HAWKINS, D., *et al.* "The Efficacy and Toxicity of Azidothymidine (AZT) in the Treatment of Patients with AIDS and AIDS-related Complex (ARC): An Open Uncontrolled Treatment Study", **QJM: An International Journal of Medicine**, fev. 1989. DOI: 10.1093/oxfordjournals.qjmed.a068309. Disponível em: https://academic.oup.com/qjmed/article/70/2/161/1507778/The-Efficacy-and-Toxicity-of-Azidothymidine-AZT-in. Acesso em: 15 jan. 2020.

STANFIELD, R. L., GORNY, M. K., ZOLLA-PAZNER, S., *et al.* "Crystal Structures of Human Immunodeficiency Virus Type 1 (HIV-1) Neutralizing Antibody 2219 in Complex with Three Different V3 Peptides Reveal a New Binding Mode for HIV-1 Cross-Reactivity", **Journal of Virology**, v. 80, n. 12, p. 6093–6105, 15 jun. 2006. DOI: 10.1128/JVI.00205-06. .

TIN KAM HO. "Random decision forests". In: **3rd International Conference on Document Analysis and Recognition**, 1, 1995. **Anais** [...] Montreal, Que., Canada, IEEE Comput. Soc. Press, 1995. p. 278–282. DOI: 10.1109/ICDAR.1995.598994. Disponível em: http://ieeexplore.ieee.org/document/598994/. Acesso em: 14 mar. 2020.

TRKOLA, A., DRAGIC, T., ARTHOS, J., *et al.* "CD4-dependent, antibody-sensitive interactions between HIV-1 and its co-receptor CCR-5", **Nature**, v. 384, n. 6605, p. 184–187, nov. 1996. DOI: 10.1038/384184a0. .

TROUPLIN, V., SALVATORI, F., CAPPELLO, F., *et al.* "Determination of Coreceptor Usage of Human Immunodeficiency Virus Type 1 from Patient Plasma Samples by Using a Recombinant Phenotypic Assay", **Journal of Virology**, v. 75, n. 1, p. 251–259, 1 jan. 2001. DOI: 10.1128/JVI.75.1.251-259.2001. .

UNAIDS. **The Joint United Nations Programme on HIV/AIDS (UNAIDS) Data 2019**. . [S.l: s.n.], 4 dez. 2019. Disponível em: http://aidsinfo.unaids.org/.

WHITCOMB, J. M., HUANG, W., FRANSEN, S., *et al.* "Development and Characterization of a Novel Single-Cycle Recombinant-Virus Assay To Determine Human Immunodeficiency Virus Type 1 Coreceptor Tropism", **Antimicrobial Agents and Chemotherapy**, v. 51, n. 2, p. 566–575, 1 fev. 2007. DOI: 10.1128/AAC.00853-06. .

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. [S.l.], Springer-Verlag New York, 2016. Disponível em: https://ggplot2.tidyverse.org.

WICKHAM, H. "Reshaping Data with the reshape Package", **Journal of Statistical Software**, v. 21, n. 12, p. 1–20, 2007. .

WICKHAM, H. **stringr: Simple, Consistent Wrappers for Common String Operations**. [S.l: s.n.], 2019. Disponível em: https://CRAN.R-project.org/package=stringr.

WILEN, C. B., TILTON, J. C., DOMS, R. W. "HIV: Cell Binding and Entry", **Cold Spring Harbor Perspectives in Medicine**, v. 2, n. 8, p. a006866–a006866, 1 ago. 2012. DOI: 10.1101/cshperspect.a006866. .

XIAO, L., OWEN, S. M., GOLDMAN, I., *et al.* "CCR5 Coreceptor Usage of Non-Syncytium-Inducing Primary HIV-1 Is Independent of Phylogenetically Distinct Global HIV-1 Isolates: Delineation of Consensus Motif in the V3 Domain That Predicts CCR-5 Usage", **Virology**, v. 240, n. 1, p. 83–92, jan. 1998. DOI: 10.1006/viro.1997.8924. .

XU, S., HUANG, X., XU, H., *et al.* "Improved prediction of coreceptor usage and phenotype of HIV-1 based on combined features of V3 loop sequence using random forest", **Journal of Microbiology (Seoul, Korea)**, v. 45, n. 5, p. 441–446, out. 2007. .

YOON, V., FRIDKIS-HARELI, M., MUNISAMY, S., *et al.* "The GP120 molecule of HIV-1 and its interaction with T cells", **Current Medicinal Chemistry**, v. 17, n. 8, p. 741–749, 2010. DOI: 10.2174/092986710790514499. .