CLASSIFICADORES DE REGRESSÃO LOGÍSTICA, *NAIVE* BAYES E *RANDOM FOREST* NA ANÁLISE DO TROPISMO DO HIV-1 DE SUBTIPO B

Cesar Borges Barros

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

Orientador: Flávio Fonseca Nobre

Rio de Janeiro
Março de 2019

CLASSIFICADORES DE REGRESSÃO LOGÍSTICA, *NAIVE* BAYES E *RANDOM FOREST* NA ANÁLISE DO TROPISMO DO HIV-1 DE SUBTIPO B

Cesar Borges Barros

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:

_____

Prof. Flávio Fonseca Nobre, Ph.D.

_____

Prof.ª Rosimary Terezinha de Almeida, Ph.D.

_____

Prof.ª Monica Barcellos Arruda, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

MARÇO DE 2019

Barros, Cesar Borges

Classificadores de Regressão Logística, *Naive* Bayes e *Random Forest* na Análise do Tropismo do HIV-1 de Subtipo B/ Cesar Borges Barros. – Rio de Janeiro: UFRJ/COPPE, 2019.

XII, 76 p.: il.; 29,7 cm.

Orientador: Flávio Fonseca Nobre

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia Biomédica, 2019.

Referências Bibliográficas: p. 65-76.

1. HIV-1 subtipo B. 2. Tropismo viral. 3. Classificadores baseados em aprendizado de máquina. I. Nobre, Flávio Fonseca. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Biomédica. III. Título.

Dedico este esforço à minha mãe e irmãos e, em especial, ao meu saudoso pai, José Maria de Barros, e ao meu querido cunhado, Marcelo "Mirim" Loureiro, recentemente ido.

"E aqueles que por obras valerosas
Se vão da lei da morte libertando",
(Luís de Camões em "Os lusíadas")

Agradecimentos

Agradeço ao Prof. Flávio Fonseca Nobre pela oportunidade para desenvolver o projeto de mestrado, pelos ensinamentos, notadamente na área de aprendizado de máquina, e pela orientação acadêmica.

Agradeço a todos os integrantes do Laboratório de Engenharia de Sistemas da Saúde (LESS) que contribuíram de alguma forma e/ou conviveram comigo neste período de realização do projeto. Em especial, à Prof.ª Rosimary e à Letícia.

Agradeço aos demais professores e colegas do Programa de Engenharia Biomédica (PEB) da COPPE.

Agradeço à minha família, especialmente à minha mãe e irmãos.

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

CLASSIFICADORES DE REGRESSÃO LOGÍSTICA, *NAIVE* BAYES E *RANDOM FOREST* NA ANÁLISE DO TROPISMO EM HIV-1 DE SUBTIPO B

Cesar Borges Barros

Março/2019

Orientador: Flávio Fonseca Nobre

Programa: Engenharia Biomédica

O desenvolvimento de antagonistas de correceptores – como o maraviroque – para o tratamento anti-HIV tornou mandatória a determinação clínica do tropismo viral previamente às terapias de resgate. Aspectos técnicos do Trofile™, o ensaio fenotípico referencial, dificultaram o seu uso como ferramenta de rotina para este diagnóstico. Isto levou ao desenvolvimento de algoritmos genotípicos, cujas avaliações são baseadas em sequências genéticas da região V3 da gp120 do HIV-1. Tais algoritmos se mostraram opções menos dispensiosas de custo e tempo, além de serem mais práticos para o uso na rotina clínica do que o ensaio fenotípico. Dentre eles, o geno2pheno começou a ser amplamente utilizado após apresentar uma concordância preditiva de 86,5% com o Trofile™. O presente projeto visou desenvolver modelos classificadores acurados, baseados em informações de sequências V3. Para isto, foram utilizadas 2.109 sequências de DNA da região V3 do HIV-1 de subtipo B. As sequências com os resultados do geno2pheno foram então modeladas pelos métodos de regressão logística, *naive* Bayes e *random forest*. Todos os classificadores apresentaram bons resultados preditivos, porém os modelos de *random forest* obtiveram o melhor desempenho discriminativo, sob a forma de resultados significativos de AUC. Tais resultados são encorajadores para a continuação do desenvolvimento de um algoritmo acurado e de uso prático para a predição clínica do tropismo viral, capaz de orientar a tomada de decisão em relação à utilização de antagonistas de correceptores no tratamento do HIV-1.

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

LOGISTIC REGRESSION, NAIVE BAYES AND RANDOM FOREST CLASSIFIERS IN THE ANALYSIS OF HIV-1 SUBTYPE B

Cesar Borges Barros

March/2019

Orientador: Flávio Fonseca Nobre

Programa: Engenharia Biomédica

The development of coreceptor antagonists – such as maraviroc – for HIV treatment has made mandatory the clinical determination of viral coreceptor usage prior to rescue therapy. Technical issues presented by Trofile™, the gold standard phenotypic assay, hindered its use as a routine diagnostic tool. This fact has lead to the development of genotypic algorithms, whose evaluations are based on DNA sequences of the V3 region of HIV-1 gp120. These algorithms proved to be cheaper, easier to use, and less time consuming than the phenotypic method. One of them, geno2pheno has also gained widespread use since it showed 86.5% predictive concordance with Trofile™. The present project aimed to develop accurate classification models based on V3 sequence information. For this, 2,109 DNA sequences of V3 region from HIV-1 subtype B were used. Data labeled with geno2pheno's results were then modeled by methods such as logistic regression, naive Bayes and random forest. All classifiers presented good predictive outputs, however random forest models showed the best discriminative performance, in the form of significant AUC results. These outcomes encourage us to continue the development of an easy to use and accurate algorithm for HIV-1 tropism diagnosis, capable of guiding clinical decision making regarding the use of coreceptor antagonists in HIV-1 treatment.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 01

Introdução

Causada pelo vírus da imunodeficiência humana (*Human Immunodeficiency Virus*, HIV), a síndrome da imunodeficiência adquirida (*Acquired Immunodeficiency Syndrome*, aids) se transformou em uma pandemia associada a altos índices de morbidade e mortalidade, com repercussões sociais e econômicas devastadoras [1].

Durante a progressão da pandemia, um marco importante foi o desenvolvimento de terapias anti-HIV, que reduziram o número de mortes e proporcionaram um controle significativo da aids, principalmente em países industrialmente avançados [2]. O surgimento e a transmissão de variantes virais resistentes às primeiras drogas, além dos efeitos adversos associados ao tratamento medicamentoso, comprometeram os esforços de controle da doença e compeliram o estudo por fármacos com novos alvos e/ou mecanismos moleculares anti-HIV [3–4].

Em 1996, foi demonstrado que o HIV requer a presença de um correceptor – CCR5 ou CXCR4 – para entrar nas células-alvo, cujas populações se diferenciam por expressarem níveis diferentes de um e/ou outro na superfície celular [5–7]. Desde então, uma nova classe de moléculas, com potencial de antagonizar CCR5 ou CXCR4 e bloquear a entrada do vírus nas células, vem sendo estudada [4, 8]. A primeira a receber aprovação regulatória foi o maraviroque, um antagonista específico do CCR5 [9]. Trata-se de uma droga eficaz contra variantes do HIV com tropismo exclusivo por este correceptor. A aprovação do maraviroque tornou mandatório o diagnóstico prévio do tropismo pelos correceptores das variantes virais em indivíduos infectados, de forma a assessorar a tomada de decisão clínica na administração da droga [10].

Considerado o procedimento fenotípico referencial para o diagnóstico do tropismo do HIV, o ensaio Trofile™ apresenta aspectos técnicos que dificultam a sua utilização na rotina clínica [11–13]. Alternativamente, algoritmos genotípicos validados, cujas predições se baseiam na análise de sequências genômicas da região variável V3 da gp120 do envelope viral, têm se apresentado como opções menos dispendiosas de custo e tempo, além de serem mais fáceis de operar que o Trofile™ [11–14]. Dentre os algoritmos, o geno2pheno (g2p) é o mais utilizado no diagnóstico genotípico do tropismo (genotropismo) [11–12, 14]. Baseia-se no pareamento de dados fenotípicos e

genotípicos do vírus sobre o uso dos correceptores, além da utilização de máquinas de vetores de suporte (*Support Vector Machines*, SVM) para a obtenção das predições [11–12, 15].

Diante do impacto que a aids ainda exercerá nos próximos anos [16], mantêm-se estratégicos o estudo e desenvolvimento de modelos que aliem desempenho acurado e usabilidade no genotropismo do HIV-1. O intuito é que estes modelos possam assessorar, de forma prática, segura e eficaz, a tomada de decisão na rotina clínica de tratamento anti-HIV com antagonistas dos correceptores CCR5 e CXCR4. Dentro deste escopo, a presente dissertação apresenta modelos classificadores baseados nos métodos de regressão logística (RL), *naive* Bayes (NB) e *random forest* (RF). O treinamento e a avaliação do desempenho destes modelos foram realizados a partir de sequências V3 de HIV subespécie 1 (HIV-1) previamente classificadas quanto ao tropismo pelo g2p.

1.1. Objetivos

1.1.1. Geral

Com base nos métodos de RL, NB e RF, e no uso de escalas de hidrofobicidade como preditores numéricos para os aminoácidos das sequências peptídicas utilizadas nas modelagens, propor modelos classificadores para o tropismo do HIV-1, usando como referência sequências genômicas da região V3 do envelope viral de isolados de HIV-1 subtipo B.

1.1.2. Específicos

- Seleção de variáveis explanatórias para a obtenção de modelos classificadores parcimoniosos;
- Determinação das variáveis explanatórias com associações – positivas ou negativas – significativas em relação à variável resposta;
- Comparação dos métodos e preditores numéricos usados a partir de resultados do desempenho preditivo dos modelos de classificação desenvolvidos, conforme a

análise das medidas de acurácia, sensibilidade, especificidade e área sob a curva ROC (*Area Under ROC - Receiver Operator Characteristic Curve - Curve*, AUC).

Capítulo 02

Fundamentos Teóricos

2.1. Aids

A aids foi reconhecida como uma nova doença em 1981 [17]. A possibilidade de ser transmitida por via sexual, percutânea ou perinatal, e de infligir um grave quadro clínico são características que contribuíram para que a aids se tornasse um dos maiores flagelos da história recente da humanidade [18].

O progresso da doença em adultos é caracterizado por três estágios clínicos. No primeiro, a infecção pode ser assintomática ou causar linfodenopatia (linfonodos aumentados) persistente. No segundo, há um declínio da resposta imune que resulta no aparecimento de infecções incomuns e contínuas por microrganismos como o fungo *Candida albicans*, em locais como boca, garganta e/ou vagina. Outras condições podem incluir herpes zoster, diarreia persistente, febre, leucoplasia pilosa oral (placas esbranquiçadas na mucosa oral) e certas condições cancerosas ou pré-cancerosas do colo do útero [19].

No terceiro estágio clínico, há o estabelecimento da aids propriamente dita. Ocorre um declínio muito acentuado das defesas imunológicas do indivíduo infectado, que passa a ser acometido por infecções oportunistas e cânceres de rara ocorrência. Importantes condições indicadoras da aids são as infecções por *C. albicans* do esôfago, traqueia, brônquios e pulmões, infecções oculares por citomegalovírus, tuberculose, pneumonia por *Pneumocystis*, toxoplasmose cerebral e o sarcoma de Kaposi. Em adultos soropositivos sem tratamento, a progressão da infecção inicial até a aids leva em média dez anos, culminando em óbito [19].

O Centro de Controle e Prevenção de Doenças (*Center for Disease Control and Prevention*, CDC) classifica o progresso da infecção por HIV pela população de linfócitos T $CD4^+$. O intuito desta classificação é o fornecimento de diretrizes para a administração de drogas anti-HIV, entre outras orientações para o tratamento. Em um adulto saudável, a contagem plasmática é de 800 a 1.000 linfócitos T $CD4^+/mm^3$. Uma contagem abaixo de $200/mm^3$ é considerada diagnóstica de aids, qualquer que seja o estágio clínico observado [19].

2.2. HIV-1

O HIV foi identificado como o agente etiológico da aids em 1984 [20]. Pertence ao gênero *Lentivirus* da família *Retroviridae*, e sua subespécie HIV-1 é a responsável pela quase totalidade dos casos da doença [18].

O HIV-1 compreende quatro grupos distintos, denominados M, N, O e P, cada um sendo o resultado evolutivo de eventos independentes de transmissão de variantes de vírus da imunodeficiência símia (*Simian Immunodeficiency Virus*, SIV) de primatas africanos para o homem [18]. O grupo M é o responsável pelo caráter pandêmico da doença, tendo infectado milhões de indivíduos e sendo encontrado em todos os países do mundo [1, 18].

2.2.1. Subtipos Virais

O grupo M compreende uma grande diversidade, estruturada filogeneticamente em nove subtipos (A-D, F-H, J e K) que têm funcionado como importantes marcadores moleculares ao longo da pandemia do HIV-1. Após o estabelecimento dos subtipos, verificou-se ainda a existência de isolados que apresentavam mosaicos genômicos, decorrentes de processos de recombinação genética entre variantes de subtipos distintos. Tais variantes exercem ação importante na pandemia, notadamente aqueles relacionados ao subtipo A, e são denominadas como formas recombinantes circulantes (*Circulating Recombinant Form*, CRF) [21].

Predominam no mundo os casos associados aos subtipos A e suas CRF (25% dos casos), o C (48%), seguidos do B (11%) [21]. A África, especialmente a Central, concentra a maior diversidade genética associada ao HIV-1: os subtipos A e C são os mais prevalentes, mas todos os grupos e subtipos coexistem, o que é coerente com o fato de ser este o epicentro da pandemia [22]. Na América do Norte, Europa e Austrália, o subtipo B é o que predomina. Na América do Sul, o subtipo B é o mais prevalente, seguido dos subtipos F e C [23]. O Brasil segue o mesmo padrão de ocorrência de subtipos da América do Sul, apresentando ainda uma menor proporção do subtipo D dentre os subtipos não-B mais representativos no país [24].

2.2.2. Estrutura Viral

Como outros lentivírus, o HIV possui um genoma constituído por duas cópias de RNA fita simples de senso positivo que, no citoplasma da célula hospedeira, geram moléculas intermediárias de DNA dupla fita. Este DNA viral é então integrado ao DNA cromossomal da célula hospedeira [19].

O genoma do HIV contém nove genes, que codificam dezenove proteínas. Destes genes, seis codificam proteínas que controlam a capacidade do HIV em infectar células, replicar o seu material genético e provocar a doença. Os outros três genes, *gag*, *pol* e *env*, fornecem a informação necessária para a produção de proteínas estruturais que irão compor as novas partículas virais [19].

Este genoma é envolto por um capsídeo cônico, imerso em uma matriz proteica composta pela proteína p17 [25]. O conjunto capsídeo e matriz proteica é delimitado pelo envelope viral, uma proteína complexa composta pelas glicoproteínas gp120 e gp41. Ambas são oriundas da clivagem proteolítica da gp160, produto do gene *env*. A gp120 e a gp41 permanecem ligadas não-covalentemente para formar uma estrutura trimérica na surpefície viral. Esta estrutura glicoproteica, notadamente a gp120, é a responsável pela especificidade de interação do vírus com proteínas da superfície da célula a ser infectada: primeiramente com o receptor CD4 e, em seguida, com os correceptores CCR5 ou CXCR4 [26]. A Figura 2.1 ilustra a estrutura geral do HIV.

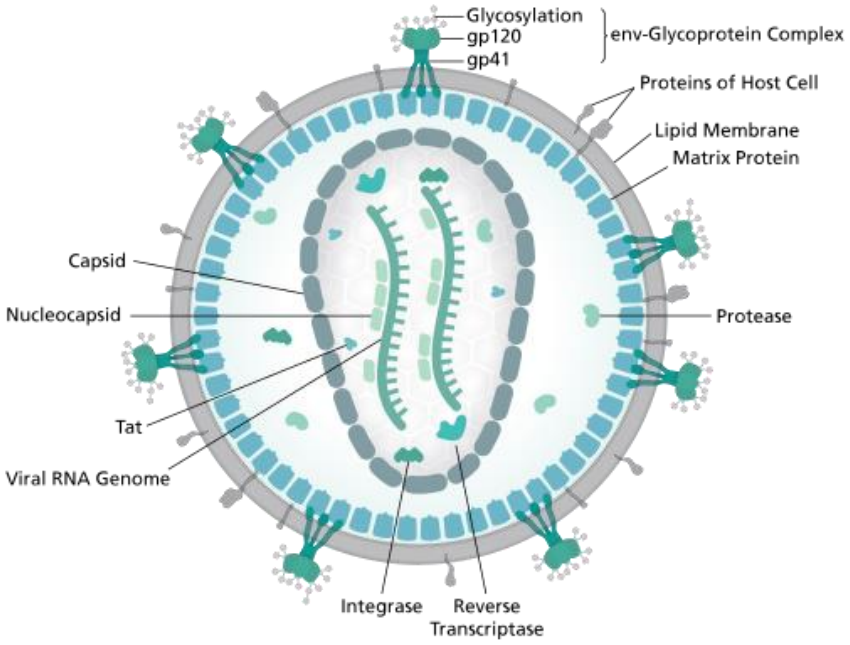

Figura 2.1. Estrutura geral do HIV (imagem permitida para uso sob a licença internacional emitida pela *Creative Commons Attribution-Share Alike* 4.0).

2.2.3. gp120

Contida no envelope viral, a gp120 se localiza externamente à membrana viral, conforme ilustrado na Figura 2.1. Com uma estrutura conformacional complexa, a glicoproteina pode ser organizada linearmente, em relação à composição de aminoácidos, em cinco regiões conservadas (C1–C5) intercaladas com cinco regiões variáveis (V1–V5). Assim, a hipervariabilidade da sequência de aminoácidos da gp120 se concentra nas cinco regiões variáveis, que se localizam na superfície mais externa da gp120. A informação sugeriu que esta variabilidade desempenharia um papel importante nas interações extracelulares do HIV [26].

A região V3 possui função crucial na interação com os correceptores CCR5 e CXCR4. Consiste aproximadamente de 34 a 36 resíduos de aminoácidos, relativos às posições 296 a 331 da gp120 da HXB2, uma cepa referencial do HIV-1 subtipo B [27]. A análise por ressonância nuclear magnética demonstrou que os peptídeos derivados da V3 exibem duas conformações distintas, que são semelhantes às apresentadas por dois grupos de moléculas do sistema imune que se ligam especificamente ao CCR5 ou CXCR4 [28]. Este conhecimento foi um dos pilares que demonstraram que a V3 seria a principal responsável pela interação seletiva com os dois correceptores. Embora outras regiões da gp120, como V1–V2 e C4, também estejam envolvidas na interação com CCR5 e CXCR4, a V3 funciona como o determinante principal para o uso dos correceptores pelo HIV [10–11, 29].

2.3. Tropismo do HIV

2.3.1. Nova Classificação Viral

Previamente à descoberta do uso dos correceptores pelo HIV, o termo “tropismo do HIV” era usado para nomear certos fenótipos virais - características expressas pelo vírus em situações específicas [10–11]. No final dos anos 80, o termo se referia à capacidade de variantes virais causarem ou não uma dada anomalia, denominada “indução de sincícios”, em células mononucleares de plasma periférico (*Peripheral Blood Mononuclear Cells*, PBMC) em cultura. Na ocasião, foi identificada uma relação deste fenótipo com dois níveis de virulência, maior ou menor, das variantes. A maior / menor

virulência se relaciona ao potencial da variante em causar no organismo um declínio mais / menos contundente de linfócitos T $CD4^+$, levando a uma progressão mais / menos rápida da infecção para o estágio clínico da aids [10–11, 30].

Posteriormente, verificou-se que variantes do HIV eram divididas em dois grupos por apresentarem uma cinética de replicação mais lenta ou mais rápida em culturas de PBMC [31]. Em seguida, foi relatada a habilidade de variantes virais em infectar culturas de macrófagos derivados de monócitos ou culturas de linfócitos T $CD4^+$ [32–33]. Houve concordância significativa entre os resultados das três características fenotípicas na discriminação dos dois níveis de virulência das variantes virais [10].

A demonstração da afinidade do HIV pelo CCR5 e/ou CXCR4 foi logo seguida da verificação de que havia também uma relação significativa entre o uso do correceptor e os dois níveis de virulência das variantes virais [34]. Como consequência, foi constatada a concordância da afinidade pelo correceptor com os resultados das três características fenotípicas: as variantes com afinidade pelo CCR5 geralmente são não-indutoras de sincício (*Non-Syncytium-Inducing*, NSI), apresentam uma cinética de replicação mais lenta em culturas de PBMC e infectam culturas de macrófagos derivados de monócitos; as variantes com afinidade pelo CXCR4 geralmente são indutoras de sincício (*Syncytium-Inducing*, SI), apresentam uma cinética de replicação mais rápida em culturas de PMBC e infectam culturas de linfócitos T [35].

Em 1998, com a manutenção do termo “tropismo do HIV”, a ampla convergência destes resultados levou a uma nova classificação do HIV, baseada no uso dos correceptores: as variantes que apresentam afinidade (tropismo) pelo CCR5, geralmente menos virulentas, são denominadas R5; as variantes com tropismo pelo CXCR4, geralmente mais virulentas, são denominadas X4; há ainda variantes intermediárias que apresentam tropismo pelos dois correceptores, e que são denominadas R5X4 [36].

2.3.2. CXCR4 e CCR5

CXCR4 e CCR5 pertencem à família de receptores celulares acoplados à proteina G e com sete domínios transmembranares em uma estrutura de alfa-hélice. Caracterizam-se ainda por apresentarem três alças extracelulares (*Extracellular Loop*, ECL) e um domínio amino-terminal, além de serem compostos por 352 aminoácidos [10–11, 37].

No sistema imune, a função primária do CXCR4 e CCR5 é atuarem como dois importantes receptores de quimiocinas – mediadoras potentes da inflamação – em populações celulares que se diferenciam por expressarem diferentes níveis de um e/ou outro receptor na superfície celular [5–7].

2.3.3. Bases Moleculares

Para se fusionar à membrana plasmática e entrar eficientemente na célula-alvo, o HIV requer CXCR4 ou CCR5 [5–7]. Porém, para que a interação do vírus com os correceptores possa ocorrer, é necessária a ligação prévia da gp120 ao receptor CD4, também presente na membrana de alguns tipos celulares. Sob a perspectiva da infecção viral, CXCR4 e CCR5 atuam como correceptores para a gp120 [10–11].

A formação do complexo CD4-gp120 leva a mudanças conformacionais no envelope viral que aumentam a exposição da região V3, inicialmente encoberta por V1 e V2, permitindo a sua interação específica com CXCR4 ou CCR5. Após esta interação, inicia-se o processo de fusão das membranas viral e celular, pela inserção do peptídeo de fusão amino terminal da gp41 à membrana da célula-alvo (Figura 2.2) [10–11].

Para as variantes X4, os domínios amino-terminal, segunda ECL (ECL-2) e ECL-3 do CXCR4 são críticos para que o vírus possa entrar na célula [38–39], enquanto que, para as R5, os domínios amino-terminal e ECL-2 do CCR5 são críticos, embora haja um contato muito íntimo da região V3 com ECL-1 [40–41]. O tropismo pelos correceptores é determinado em parte importante pela carga líquida expressa na V3, com as variantes X4 apresentando uma carga líquida positiva maior que as R5 [11]. A Figura 2.3 ilustra mais detalhes das bases moleculares do tropismo do HIV.

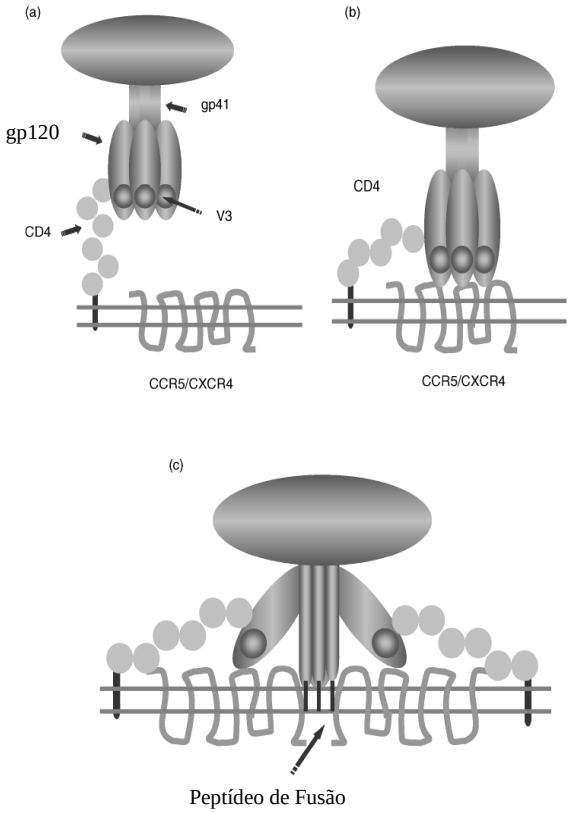

Figura 2.2. O processo de entrada do HIV na célula-alvo: (a) ligação CD4-gp120; (b) interação do complexo CD4-gp120 com CCR5 ou CXCR4; (c) fusão das membranas viral e celular, pela interação da gp41 com a membrana da célula-alvo. Adaptado de Poveda *et al.* (2006) [10]. Iimagem permitida para uso sob autorização do autor responsável, Vincent Soriano.

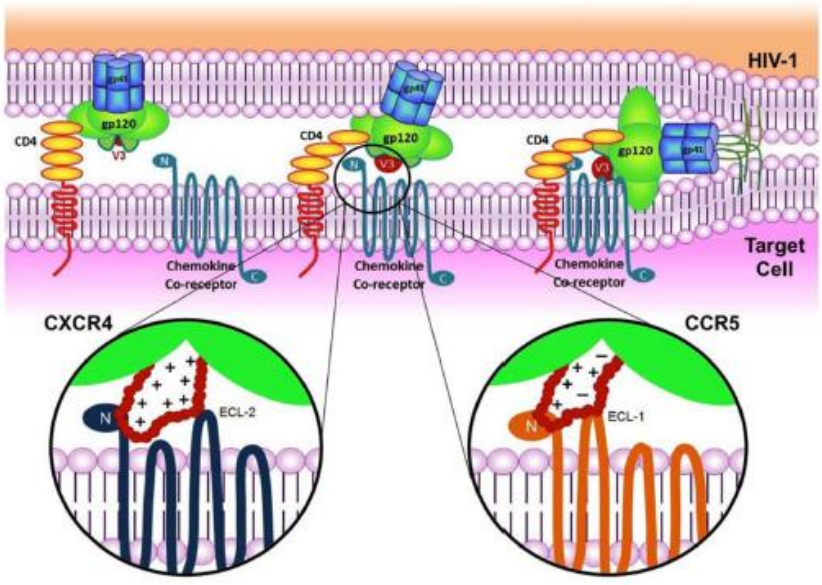

Figura 2.3. Detalhes da interação molecular entre o HIV e a célula-alvo, além de detalhes dos domínios amino-terminal (N), transmembranares e ECL dos correceptores: interação mais íntima da V3 com CXCR4 e CCR5; e ilustração das cargas líquidas de V3 durante a interação com os correceptores. Adaptado de Aiamkitsumrit *et al.* (2014) [11].

2.3.4. Progressão da Doença

O padrão de expressão do CCR5 e CXCR4 na superfície de diferentes células do sistema imune influenciam a dinâmica do tropismo viral e o decurso da doença. Uma vez que parecem ser mais eficientemente transmitidas que as variantes X4, as R5 são geralmente as responsáveis pelo início da infecção viral [42]. Por outro lado, as variantes X4 tendem a predominar em estágios mais tardios da infecção, levando à grande redução de linfócitos T $CD4^+$ e à progressão mais rápida da doença [33].

No gene que codifica CCR5, há uma mutação recessiva, relacionada à deleção – excisão – da parte que codifica ECL-2, que leva à perda de função do CCR5 como correceptor para o HIV. Os raros indivíduos homozigotos para esta mutação são imunes à infecção, mesmo que tenham sido expostos inúmeras vezes ao vírus [43–44]. Este fato sugeriu a existência de uma forte seleção a favor das variantes R5 em episódios de contágio e estabelecimento inicial da doença. Durante o contato sexual, incluindo a via retal, as variantes R5 possuem aparentemente uma maior vantagem de transmissão e infecção devido aos altos níveis de expressão de CCR5 nos tecidos da mucosa genital e também nas células do epitélio intestinal [45–46]. Por sua vez, após exposição parenteral (hemofílicos e usuários de drogas), as variantes X4 podem ser eficientemente transmitidas, tendo a princípio a possibilidade de estabelecerem uma infecção. Entretanto, na maioria destes casos, as R5 também predominam nos estágios iniciais da doença [10, 47].

A progressão mais rápida da infecção inicial para a aids é explicada em metade dos casos a uma mudança no perfil do tropismo predominante das variantes de R5 para X4 [34]. A maior virulência das variantes X4 parece ser explicada em parte pela sua habilidade em infectar timócitos, células geradas no timo e que são precursoras dos linfócitos T $CD4^+$. O CXCR4 é altamente expresso em timócitos imaturos, o que os torna muito suscetíveis às X4. Apesar de também poderem chegar ao timo, as R5 não afetam o processo de formação e desenvolvimento dos timócitos [48–49].

Para os pacientes que desevolvem a aids na ausência de variantes com tropismo pelo CXCR4, foi verificada a existência de características que distinguem as R5 "tardias" encontradas nestes indivíduos das R5 geralmente isoladas de pacientes assintomáticos. Entre outras vantagens adaptativas, tais diferenças propiciariam um aumento da eficiência de interação das R5 tardias pelo CCR5. Isto as tornaria mais virulentas para populações de células com menores níveis de expressão deste

correceptor. Tais incrementos na capacidade infectiva alavancariam a progressão da doença [50].

2.3.5. Maraviroque

O maraviroque foi a primeira droga com ação antagônica aos correceptores a receber aprovação regulatória [9]. Administrado por via oral, é um antagonista específico do CCR5, com potente atividade anti-HIV e propriedades farmacológicas bem favoráveis para o uso clínico [9, 51].

Sendo eficaz contra variantes com tropismo exclusivo pelo CCR5, o maraviroque atua de forma não competitiva com outras moléculas, bloqueando seletiva e reversivelmente a ligação da gp120 ao CCR5. Trata-se de uma inibição alostérica, onde a droga se liga a partes específicas da alfa-hélice do CCR5, induzindo mudanças conformacionais em outras partes do correceptor, especialmente na ECL-2. Assim, há uma inibição da capacidade da V3 de interagir com o CCR5, levando à interrupção do processo de entrada do vírus na célula [12, 51].

O maraviroque não é recomendado como um medicamento anti-HIV inicial, mas como alternativa de terapia de resgate em determinados casos após a ocorrência de múltiplas falhas virológicas. A falha virológica após tratamento antirretroviral específico é caracterizada pela: a) não obtenção de carga viral abaixo de 500 cópias de RNA de HIV-1 / mm^3 de plasma após 6 meses de tratamento, ou b) não manutenção de carga viral indetectável – abaixo de 50 cópias / mm^3 – após 12 meses de tratamento, pela ocorrência de rebote confirmado de carga viral acima de 500 cópias / mm^3 [52].

Desta forma, o maraviroque deve ser considerado para compor terapias anti-HIV após múltiplas falhas virológicas, quando drogas de resgate como darunavir (DRV), um inibidor de protease, dolutegravir (DTG), um inibidor da integrase, e etravirina (ETR), um inibidor de transcriptase reversa, sejam consideradas insuficientes para garantir a supressão viral [52]. Nestes casos, a utilização do maraviroque deve ser forçosamente precedida do genotropismo, além do diagnóstico genotípico de resistência a antirretrovirais (genotipagem) das variantes virais circulantes, entre outras informações clínicas, com o objetivo final de otimizar o tratamento de resgate [12, 52].

2.3.6. Métodos de Predição Referenciais

2.3.6.1. Predição Fenotípica - Trofile™

Validado para uso na rotina clínica, o Trofile™ é o ensaio referencial para a determinação fenotípica do tropismo do HIV-1. O teste foi utilizado como ferramenta diagnóstica em estudos clínicos de antagonistas de correceptores, incluindo os estudos de fase III que embasaram a aprovação regulatória do maraviroque, MOTIVATE 1 e 2 (*Maraviroc plus Optimized Therapy in Viremic Antiretroviral Treatment Experienced Patients*), além de outros estudos como o MERIT (*Maraviroc versus Efavirenz in Treatment-Naive Patients*) [12, 53]. Por outro lado, o Trofile™ apresenta limitações técnicas e logísticas que o tornam pouco prático para uso na rotina clínica [11–12].

O modo operacional do ensaio é baseado em tecnologia de recombinação viral. O RNA viral contido em uma amostra plasmática do indivíduo soropositivo é convertido em DNA complementar (cDNA), via reação com a enzima transcriptase reversa. Em seguida, os genes virais *env* contidos neste cDNA são amplificados pelo método da Reação em Cadeia da Polimerase (*Polymerase Chain Reaction*, PCR) [53].

Considerando que o produto da PCR contempla a diversidade genética das variantes virais circulantes, as sequências amplificadas de cDNA, contendo a gp120, são clonadas em cópias de um vetor de expressão (plasmídeo pCXAS-PXMX). Os plasmídeos resultantes são transfectados (introdução intencional de DNA em células eucarióticas), em conjunto com vírus HIV-1 construídos sem o gene *env*, em células humanas HEK293. Como resultado, há a produção *in vitro* de uma população de pseudovírus de HIV-1 que expressam os alelos da gp120 relativos às variantes circulantes da amostra original, sob a forma de envelopes virais [53].

As partículas pseudovirais são utilizadas para infectar linhagens de células humanas geneticamente modificadas (células U87) de modo que, além de expressarem o receptor CD4 na sua superfície, possam expressar ainda o correceptor CCR5 ou CXCR4. A depender da capacidade dos pseudovírus de entrarem em uma e/ou outra linhagem celular – verificada após 72h do inóculo, pela emissão de luz quantificável produzida pela expressão de um gene repórter específico (luciferase) presente nos pseudovírus – o diagnóstico fenotípico do tropismo é então realizado. No caso de infecção das duas culturas de células pela mesma população de pseudovírus, há uma maior chance deste resultado representar a ocorrência de uma mistura de variantes X4 e

R5 no plasma do indivíduo testado, do que ser devido à presença única de variantes R5X4 [12].

O Trofile™ é capaz de detectar variantes X4 circulantes com uma frequência de ocorrência tão baixa quanto 10% [53]. O ensaio foi aperfeiçoado em 2008 pelo desenvolvimento do *Enhanced Sensibility*-Trofile™ (ES-Trofile™), que é capaz de detectar populações minoritárias de variantes X4 com uma frequência de ocorrência tão baixa quanto 0,3% [54].

Apesar da maior sensibilidade do ES-Trofile™, não há um aumento significativo na sua habilidade em discriminar indivíduos que irão responder ou não ao tratamento clínico baseado no maraviroque, quando comparado ao Trofile™: a maioria dos pacientes com variantes X4 minoritárias, que não tenham sido detectadas pelo Trofile™, respondem favoravelmente ao tratamento de resgate com o maraviroque, atingindo a carga viral de < 50 cópias / mm^3 de plasma após 12 meses de tratamento [12, 55]. Nestes casos, a atividade de outros antirretrovirais administrados de forma personalizada – conforme a análise individual de perfis de resistência das variantes virais às drogas anti-HIV, além de informações como o tropismo viral, o uso prévio dos antirretrovirais, carga viral, entre outros dados clínicos – em conjunto ao maraviroque aparece como o principal preditor de sucesso terapêutico em pacientes com uma baixa prevalência de variantes X4 circulantes [56].

2.3.6.2. Genotropismo - Geno2pheno

Apresentando-se como uma alternativa menos dispensiosa em termos de tempo e custo em relação à predição fenotípica, o genotropismo do HIV é determinado principalmente pela sequência de aminoácidos da região V3 da gp120. Deste modo, algoritmos classificadores foram desenvolvidos para analisar diretamente sequências de DNA da V3 e/ou sua conversão em aminoácidos [10–12].

O g2p se tornou uma referência validada no genotropismo do HIV, após a concordância significativa dos seus resultados com aqueles obtidos pelo Trofile™ nos estudos MOTIVATE [11–12, 14]. Baseia-se no pareamento de dados fenotípicos e genotípicos do vírus sobre o uso dos correceptores, além da utilização do método de aprendizado estatístico SVM para a obtenção das predições. Uma vez que o SVM é um método de classificação binária, a modelagem e as predições levam em consideração o

tropismo pelo CXCR4, com as variantes X4 e R5X4 compondo a classe positiva [11–12, 15].

O algoritmo é o resultado de um procedimento de validação cruzada – conjunto de técnicas para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados [57] – utilizando um banco de 1.100 sequências V3 relativas a cepas de HIV-1, sendo a maioria do subtipo B. O tropismo fenotípico associado às sequências foi determinado pela capacidade do vírus de infectar culturas de células indicadoras. A maioria das sequências foi obtida do banco de dados de Los Alamos – um sítio referencial sobre o HIV, que contém um amplo acervo de dados de sequências genéticas e de imunologia relacionados ao vírus – e complementada por sequências da literatura científica. Estas V3 são oriundas de 332 pacientes, sendo que 769 sequências são de variantes R5, 210 de X4, e 131 de R5X4 [11–12, 15].

O g2p está disponível na internet e as predições podem ser realizadas a partir de arquivos no formato FASTA, com as sequências V3 a serem testadas representadas por nucleotídeos (DNA) ou aminoácidos. O servidor permite a predição simultânea de até 50 sequências V3, que podem ser copiadas e coladas em um campo textual, ou disponibilizadas no servidor por intermédio de um arquivo digital no formato de valores separados por vírgula (*Comma-Separated Values*, csv). Para a análise, as sequências não precisam se ater ao peptídeo usual de 34 a 36 aminoácidos da V3, visto que o algoritmo faz o alinhamento automático da sequência teste contra uma V3 referencial [11–12, 15].

Uma vez que depende de tecnologias de sequenciamento de DNA para a obtenção das V3, o algoritmo validado do g2p inclui a análise de sequências de amostras clínicas obtidas por sequenciamento de Sanger, que é capaz de detectar populações minoritárias de vírus circulantes com prevalência mínima de 10% [14, 58]. No servidor do g2p, a análise pelo algoritmo validado está indicada na opção "*original g2p coreceptor*" [15]. O aparecimento de novas tecnologias de sequenciamento, denominadas de sequenciamento de nova geração (*Next Generation Sequencing*, NGS), incrementaram a sensibilidade de detecção de variantes X4 circulantes com prevalência abaixo de 1% [59]. Neste sentido, o NGS foi também incorporado ao g2p sob a forma de uma versão alternativa ainda não validada do algoritmo, disponibilizada no servidor na opção "geno2pheno-C_NGS-Sanger" [60].

2.4. Escalas de Hidrofobicidade

Considerando que a hidrofobicidade é uma importante força de estabilização no enovelamento de proteinas, as escalas de hidrofobicidade apresentam valores desta propriedade físico-química para os vinte resíduos de aminoácidos. Estes valores são comumente usados para a predição, em proteínas transmembranares, de segmentos peptídicos que estejam imersos nas membranas biológicas. Porém, a diversidade das propriedades biofísicas destas membranas nos diferentes compartimentos celulares, entre outros aspectos, é um impedimento à formulação de uma escala de hidrofobicidade que contemple de maneira ótima tal complexidade [61].

Esta constatação levou ao desenvolvimento de várias escalas de hidrofobicidade. Escalas consensuais, com o intuito de combinar em uma única escala as vantagens de outras, também foram elaboradas [61]. Dentre estas, destacam-se a escala de Kyte e Doolittle, baseada em uma variedade de observações experimentais encontradas na literatura [62], a escala de Eisenberg e colaboradores, derivada de cinco escalas distintas [63], e a escala de Guy, desenvolvida a partir de resultados experimentais e estatísticos de inúmeros estudos [64].

As escalas de hidrofobicidade vêm sendo também muito utilizadas em projetos de aprendizado estatístico, com os valores de hidrofobicidade atuando como preditores numéricos no desenvolvimento de classificadores baseados em sequências de aminoácidos. As aplicações destes classificadores são variadas, incluindo análises do tropismo do HIV-1 [65–68].

2.5. Modelos de Classificação

Em aprendizado de máquina, os modelos de classificação visam associar uma categoria ou classe a uma dada observação, de forma a responder a pergunta: isto é “A” ou “B”? Tratam-se de modelos desenvolvidos a partir de aprendizagem supervisionada, quando há a construção de um modelo para a predição, ou estimativa de uma “saída” baseada em uma ou mais “entradas”. Estes classificadores são muito usados em bioinformática, notadamente na área de testes diagnósticos, quando se quer verificar por exemplo a presença (“A”) ou ausência (“B”) de uma dada doença em um grupo de indivíduos [69].

Nos modelos de classificação, as variáveis independentes – "entradas" – podem ser contínuas e/ou categóricas. A variável dependente – "saída" – é categórica e geralmente de natureza dicotômica ("A" ou "B"). A depender do problema de classificação, uma saída que não seja dicotômica pode ser convertida para sê-la [69]. Esta formatação ocorre na predição do tropismo do HIV-1, onde as três classificações possíveis, X4, R5X4 e R5 são readequadas conforme uma classificação binária que distingue as variantes com tropismo pelo CXCR4 daquelas com tropismo exclusivo pelo CCR5. Neste caso, as variantes com tropismo pelo CXCR4 - X4 e R5X4 - compõem a classe positiva [15].

O desenvolvimento do modelo inclui uma etapa de treinamento e outra de avaliação, ou teste de desempenho. Durante o treinamento, o objetivo é a obtenção de dados que contemplem de maneira satisfatória a complexidade do problema a ser estudado, de maneira que um modelo preditivo acurado possa ser desenvolvido a partir de métodos como regressão logística, *naive* Bayes, entre outros classificadores baseados em aprendizado de máquina. Durante a etapa de avaliação, a capacidade preditiva do modelo construído é finalmente testada a partir de novos dados, não associados à etapa prévia de treinamento [69].

2.5.1. Regressão Logística

A regressão logística (RL) consiste em um modelo linear generalizado, cuja classificação resultante é a estimativa da probabilidade de um evento ocorrer em função de um conjunto de variáveis preditoras (independentes), que podem ser qualitativas e/ou quantitativas. Este modo operacional redunda em uma classificação dicotômica de ocorrência ou não de um evento, onde os resultados de probabilidade ficam contidos no intervalo de zero a um [70].

Como exemplo, pode-se estabelecer um ponto de corte em 0,5 de forma que uma observação, associada a uma probabilidade resultante que esteja acima deste ponto, receba uma classificação "A", em função do evento "A" ter ocorrido. Por outro lado, no caso de uma probabilidade resultante que esteja abaixo de 0,5, a observação associada receberia uma classificação "B" em função de "A" não ter ocorrido. A depender do problema de classificação, outros pontos de corte podem ser estabelecidos [70].

Por se tratar de um modelo linear generalizado, a RL apresenta três componentes: um componente aleatório, relacionado à distribuição de probabilidades da variável dependente (resposta); um componente sistemático, que relaciona linearmente as variáveis independentes com os respectivos parâmetros; e uma função de ligação, *logit*, que relaciona os preditores lineares do modelo (componente sistemático) com os valores esperados da variável resposta (componente aleatório). A determinação dos parâmetros – coeficientes da regressão – é realizada pelo método da máxima verossimilhança, que gera valores que maximizam a função de verossimilhança e que geralmente apresentam propriedades matemáticas consistentes [70].

Conforme mencionado, na RL a variável resposta (y) é dicotômica, ou seja, dois valores lhe são atribuídos: 1, para o acontecimento de interesse, denominado sucesso, e 0 para o acontecimento complementar, o fracasso. A probabilidade de sucesso é dada por π_i, e a de fracasso por 1 - π_i. Considerando-se uma série de variáveis aleatórias independentes x_1, x_2, x_3, ..., x_n, e um vetor $\beta = \beta_0, \beta_1, \beta_2, ..., \beta_p$, formado por parâmetros desconhecidos do modelo, a probabilidade de sucesso é dada por:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})}{1+\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})} \quad (2.1)$$

A probabilidade de fracasso é dada por:

$$1 - \pi_i = \frac{1}{1+\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})} \quad (2.2)$$

O *logit* para o modelo de RL é dado por:

$$g(x_1) = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = x_i^T\beta = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij} \quad (2.3)$$

E o logaritmo da função de verossimilhança pode ser escrito como:

$$l(\beta) = \sum_{i=1}^{n}[y_i x_i^T\beta - \ln(1 + \exp\{x_i^T\beta\})] \quad (2.4)$$

Para se avaliar o ajuste dos modelos obtidos pela RL, um dos métodos recomendados é o uso da função desvio (*deviance*, D). A *deviance* de um modelo qualquer é definida como sendo o desvio deste modelo em relação ao modelo saturado, no qual todos os parâmetros se ajustam perfeitamente a todas as observações [70], conforme a definição:

$$\mathrm{D} = -2\ln\frac{(verossimilhança\ do\ modelo\ ajustado)}{(verossimilhança\ do\ modelo\ saturado)} \tag{2.5}$$

O numerador é a função de verossimilhança do modelo ajustado, em questão, e o denominador é a função de verossimilhança do modelo saturado. Considerando que o modelo mais simples é denominado modelo nulo, formado apenas pelo parâmetro β_0, a *deviance* é, portanto, utilizada para medir a discrepância de um modelo intermediário de *p* parâmetros em relação ao modelo saturado. Quanto menor for a *deviance*, melhor o ajuste do modelo. Entre modelos aninhados, a significância de uma diferença de ajuste via *deviance* pode ser avaliada por testes como o da razão de verossimilhança. Este teste é também utilizado para verificar a significância da contribuição de cada variável preditora para o ajuste em um dado modelo [70].

Em modelos de regressão, há a possibilidade de se selecionar variáveis preditivas a partir de métodos iterativos como a seleção *stepwise*. O intuito é a obtenção de modelos parcimoniosos que combinem um ótimo ajuste com o menor número possível de variáveis [69]. Durante as iterações na seleção *stepwise*, um dos métodos utilizados para avaliação do modelo estatístico é o critério de informação de Akaike (*Akaike Information Criterion*, AIC). Este critério considera tanto o ajuste do modelo quanto a sua simplicidade, penalizando os modelos que contenham um número maior de variáveis. Uma vez que AIC é uma medida relacionada à perda de ajuste de um determinado modelo, quanto menor for este valor, melhor o ajuste do modelo [71]. O AIC é dado pela fórmula a seguir (2.6), onde L_p é a função de máxima verossimilhança do modelo e *p* o número de parâmetros a serem estimados na modelagem.

$$\mathrm{AIC} = -2\log\left(L_p\right) + 2(p) \tag{2.6}$$

2.5.2. *Naive* Bayes

Naive Bayes (NB) é um método de classificação probabilística. Baseia-se no teorema de Thomas Bayes, que trata de problemas em que se deseja determinar a probabilidade de um evento B ocorrer na condição de que A já tenha ocorrido, conforme a definição:

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)} \tag{2.7}$$

em que P(B | A) é a probabilidade condicional de ocorrer B dado que A ocorreu; P(B) e P(A) são as probabilidades de ocorrência de B e A, respectivamente; e P(A | B) é a probabilidade condicional de ocorrer A dado que B ocorreu [72].

O algoritmo NB é denominado "ingênuo" por assumir que as variáveis preditoras (qualitativas e/ou quantitativas) são condicionalmente independentes, ou seja, a informação de um evento não é informativa sobre nenhum outro. Trata-se de um dos métodos de aprendizado estatístico mais utilizados em problemas de classificação, em função de sua simplicidade e capacidade preditiva [72].

As probabilidades *a priori* são aquelas prévias utilizadas no algoritmo NB, e que estão associadas às frequências de ocorrência das classes no conjunto de dados de treinamento. A lógica subjacente a estas probabilidades *a priori* é que as frequências de ocorrência das classes nos dados de treinamento sejam similares às frequências encontradas no conjunto de dados teste. Assim, a previsão de um desfecho não é influenciada somente pelas variáveis preditoras, mas também pela prevalência do desfecho. Probabilidades condicionais são calculadas para cada variável. Desta forma, considerando-se que NB usa o Teorema de Bayes como princípio, tem-se [72]:

$$P(y_i \mid x) = \frac{P(x \mid y_i)\, P(y_i)}{P(x)} \tag{2.8}$$

em que $P(y_i \mid x)$ é a probabilidade a posteriori, ou seja, a probabilidade de uma dada observação, com suas respectivas variáveis preditoras (x), pertencer à classe y_i; $P(x \mid y_i)$ é a probabilidade condicional (verossimilhança), ou seja, a probabilidade de verificar observações que pertencem a uma dada classe. É decomposta em probabilidades relativas a cada uma das variáveis preditoras do modelo e que são multiplicadas entre si,

no caso, $P(x^1 \mid y_i)$ x...x $P(x^n \mid y_i)$; $P(y_i)$ é a probabilidade da referida classe (prevalência); e P(x) é a probabilidade de ocorrência das variáveis preditoras em questão. Ao final, o denominador pode ser ignorado, visto que é o mesmo para todas as classes. Diferentemente da RL, em NB não há uma etapa pós-treinamento de seleção de variáveis [69, 72].

2.5.3. *Random Forest*

Em aprendizado de máquina, a partir de um mesmo conjunto de dados de treinamento, é esperado que a combinação dos resultados de vários classificadores melhore o desempenho preditivo e a confiança na tomada de decisão, se comparada à análise de um único classificador. Desta forma, há um interesse na pesquisa e desenvolvimento de métodos de modelos preditivos múltiplos (métodos *ensemble*), que se caracterizam pela geração de muitos classificadores e a combinação de seus resultados. O algoritmo *random forest* (RF) é um exemplo de método *ensemble* que utiliza classificadores do tipo árvore de decisão [73].

Na etapa de treinamento, a RF produz uma grande quantidade de árvores de decisão conforme o método *bagging*, que consiste na criação e aprendizado paralelo de preditores (ou seja, cada modelo é construído independentemente) a partir da geração repetida de amostras com reposição e de mesmo tamanho do conjunto original de dados (amostragem *bootstrap*). A utilização do método *bagging* no treinamento tem como objetivo reduzir a complexidade dos modelos, de forma a evitar a ocorrência de super ajuste dos dados por modelos muito complexos. Ademais, o *bagging* reduz a variância que interfere no desempenho de preditores gerados que sejam instáveis [74].

O algoritmo RF adiciona aleatoriedade ao modelo quando da criação das árvores, na medida que busca as melhores características para fazer a partição dos nodos, a partir de subconjuntos aleatórios das variáveis. Este procedimento gera diversidade, o que normalmente leva à formação de melhores preditores *ensemble* [73]

Ao final, cada árvore classificadora é apontada como um componente preditor. Neste sentido, a RF constroi sua decisão por meio da contagem dos votos dos componentes preditores em cada classe e, em seguida, seleciona a classe vencedora em termos de número de votos acumulados dentre todas as “árvores da floresta” [73].

2.6. Desempenho dos Modelos de Classificação

O desempenho de um modelo de classificação pode ser avaliado por meio de medidas calculadas a partir de uma matriz de confusão para duas classes. Trata-se de uma tabela de contingência 2x2, onde são representados quatro tipos de classificação, conforme os resultados de predição do modelo (Tabela 2.2). Na matriz, há duas classes: a positiva (presença da condição-alvo) e a negativa (ausência da condição-alvo), associadas a uma predição gerada por um método referencial; e há duas predições: positiva e negativa, associadas ao método / modelo preditivo a ser avaliado [75].

Para determinar o número de acertos do modelo final, é necessário estabelecer uma probabilidade denominada ponto de corte. Probabilidades estimadas pelo modelo que sejam maiores ou iguais a este ponto recebem a classificação da classe estabelecida como positiva. Probabilidades que estejam abaixo do ponto de corte, recebem a outra classificação possível pela não ocorrência da classe positiva [75].

Quando há um resultado positivo tanto para a classe quanto para a predição, tem-se um verdadeiro positivo (VP); um resultado positivo para a classe, mas negativo para a predição, tem-se um falso negativo (FN); um resultado negativo para a classe, mas positivo para a predição, tem-se um falso positivo (FP); por fim, quando há um resultado negativo tanto para a classe quanto para a predição, tem-se um verdadeiro negativo (VN) [75]. A partir destas classificações, podem ser definidas três medidas mais comuns de desempenho preditivo: acurácia, sensibilidade e especificidade.

Tabela 2.2. Matriz de confusão para duas classes.

	Classe Positiva	**Classe Negativa**
Predição Positiva	VP	FP
Predição Negativa	FN	VN

2.6.1. Acurácia

A acurácia é definida como a proporção de acertos do modelo a ser testado. É dada pela fórmula:

$$A = \frac{(VP+VN)}{(VP+VN+FP+FN)} \quad 2.9$$

Na prática, a acurácia é pouco útil, pois agrega valores que são obtidos separadamente, ou seja, mistura a sensibilidade e a especificidade [75].

2.6.2. Sensibilidade

A sensibilidade é definida como a proporção de VP dentro da classe positiva. É dada pela fórmula:

$$S = \frac{VP}{(VP+FN)} \quad 2.10$$

Um modelo com alta sensibilidade de predição raramente deixará de diagnosticar a presença da condição-alvo quando estiver realmente presente. Ademais, quanto maior a sensibilidade, menor a chance de um modelo classificar como pertencente à classe negativa um valor da classe positiva, fornecendo assim uma menor taxa de FN [75].

Os modelos e testes com alta sensibilidade são especialmente úteis quando há a necessidade de diagnosticar uma condição-alvo potencialmente grave, reduzindo a chance de outro diagnóstico possível, além de serem muito utilizados para realizar o rastreamento da condição-alvo em grupos populacionais [75].

2.6.3. Especificidade

A especificidade é definida como a proporção de VN dentro da classe negativa. É dada pela fórmula:

$$\mathrm{E} = \frac{\mathrm{VN}}{(\mathrm{VN}+\mathrm{FP})} \qquad 2.11$$

Um modelo com alta especificidade de predição raramente classificará a presença da condição-alvo quando estiver realmente ausente. Ademais, quanto maior a especificidade, menor a chance de um modelo classificar como pertencente à classe positiva um valor da classe negativa, fornecendo assim uma menor taxa de FP [75].

Os modelos e testes com alta especificidade são utilizados para excluir uma dada condição-alvo. São particularmente necessários quando o resultado FP pode lesionar o paciente de forma física, emocional e/ou financeira como, por exemplo, um teste de diagnóstico do HIV [75] ou, no caso do tropismo, a classificação errônea de variantes R5 como X4.

No g2p, a análise de predição pode ser ajustada quanto à taxa de FP (*FP Rate*, FPR), um valor que funciona como um ponto de corte: uma sequência V3 teste assinalada após a predição com uma FPR acima do ponto de corte determinado, por exemplo FPR de 10%, é classificada como R5 [15]. Quanto mais alto o valor do ponto de corte estipulado, maior a possibilidade de ocorrer a classificação errônea de variantes R5 como X4, com a consequente tendência de se excluir pacientes que poderiam se beneficiar do tratamento com base no maraviroque. Por outro lado, pontos de corte iguais ou mais baixos tendem a permitir que um maior número de pacientes com variantes X4 sejam selecionados para o tratamento com o maraviroque. Evidências suportam que o uso de FPR na faixa de 5,75 a 10% evita a exclusão de pacientes que possam se beneficiar do tratamento resgate com o maraviroque, ao mesmo tempo que garante a administração segura e eficaz da droga [76].

2.6.4. Curva ROC

Uma forma de expressar graficamente a relação entre a sensibilidade e a especificidade é através da construção da curva ROC. Desenvolvida na década de 50, o seu uso se tornou bastante comum na área médica. Sua construção é feita colocando-se os valores da sensibilidade (proporção de VP) no eixo Y, e o complemento da especificidade (1 – especificidade), ou seja, a proporção de FP, no eixo X para diferentes pontos de corte. A curva ROC pode servir como orientação para a escolha do melhor ponto de corte de um teste diagnóstico que, em geral, localiza-se no extremo da curva próximo ao canto superior esquerdo do gráfico [75]. A Figura 2.4 ilustra as informações associadas à Curva ROC.

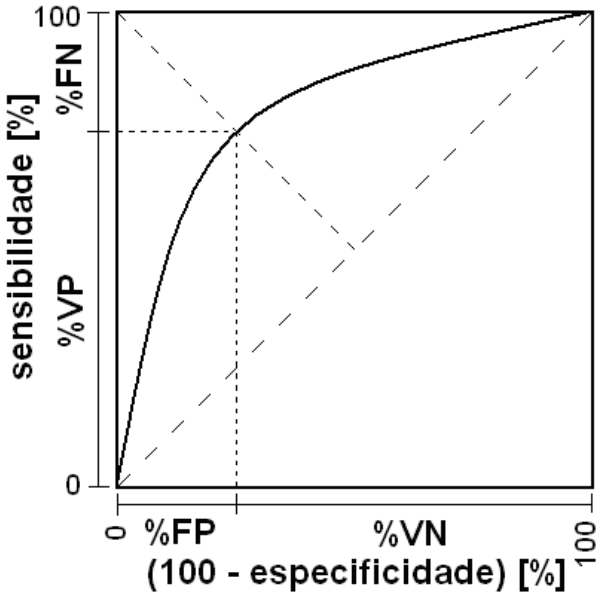

Figura 2.4. Curva ROC, e a representação da relação de reciprocidade entre a sensibilidade e a especificidade de um classificador binário. Adaptado de Sovierzoski *et al.* (2009) [77]. Imagem permitida para uso sob autorização do primeiro autor, Miguel Sovierzoski.

Além de auxiliarem na identificação do melhor ponto de corte, as curvas ROC são muito utilizadas para comparar dois ou mais testes diagnósticos para a mesma condição-alvo. Neste caso, o poder discriminatório do teste pode ser mensurado por meio da AUC. Quanto mais próxima a curva estiver do canto superior esquerdo do gráfico, maior a AUC e melhor será o poder discriminatório do teste ou modelo preditivo em questão; e quanto mais distante, até o limite da diagonal do gráfico, pior será o seu poder de predizer uma dada condição-alvo [75].

Capítulo 03

Revisão Bibliográfica

Dentre as ferramentas de genotropismo, a "regra 11/25" foi a primeira a ser desenvolvida. Fundamenta-se em informações acumuladas de que variantes do HIV, apresentando os aminoácidos básicos arginina (R) ou lisina (K) nas posições 11 e/ou 25 da sequência peptídica da V3, tendem a ser X4 / R5X4. Por outro lado, as ausências de R e K nestas posições estariam mais associadas às variantes R5 [78–81]. A revisão da regra 11/25 para "11/24/25", que classifica as variantes como X4 / R5X4 quando as posições 11, 24 e/ou 25 apresentam R ou K, objetivou melhorar o desempenho preditivo apresentado pela regra 11/25 na identificação de variantes com tropismo pelo correceptor CXCR4 [82].

Cardozo *et al.* (2007) compararam as regras 11/25 e 11/24/25 entre si, a partir da análise de um conjunto de 217 sequências V3 de isolados clínicos de HIV-1 oriundos de nove subtipos. Dentre as 217, 154 (71%) sequências eram oriundas de variantes de subtipo B. Os autores verificaram que, das 54 sequências V3 de variantes fenotipicamente classificadas como X4, 39 (72%) foram identificadas como X4 pela regra 11/25, enquanto que 48 (89%) o foram pela regra 11/24/25. Ambas as regras apresentaram a mesma especificidade de 96%, com 157 sequências V3 classificadas como R5, dentre 163 sequências [82].

A "regra da carga líquida" é também outro algoritmo de simples interpretação, que estima a carga líquida global da sequência peptídica da V3 a partir de valores de carga atribuídos a cada aminoácido. Em conformidade com o fato do tropismo ser determinado em parte importante pela carga líquida expressa na V3, com as variantes X4 apresentando uma carga líquida positiva maior que as R5 (Figura 2.3), na regra da carga líquida uma variante é classificada como X4 se a carga líquida global da V3 for igual ou maior a 5. Caso seja menor que 5, a variante é classificada como R5. A regra da carga líquida apresenta um desempenho preditivo global semelhante à regra 11/25, com a particularidade de geralmente apresentar uma sensibilidade maior e especificidade menor em relação à regra 11/25 [83].

A combinação das regras 11/25 e carga líquida originou outros dois algoritmos: a regra de Garrido, que classifica uma V3 como X4 / R5X4 se as regras 11/25 ou carga

líquida a classificarem como X4 / R5X4 [83]; e a regra de Delobel, que classifica uma V3 como X4 / R5X4 se ambas as regras a classificarem como X4 / R5X4 [59].

Seclén *et al.* (2010) compararam os desempenhos preditivos das cinco regras mencionadas, além de outros algoritmos genotípicos. A partir de amostras clínicas de 150 pacientes, 75 infectados por HIV-1 do subtipo B e 75 por subtipos não-B, os resultados de genotropismo foram comparados aos resultados do ensaio fenotípico referencial (Tabela 3.1). Os autores concluíram que as regras poderiam atuar como ferramentas para o genotropismo de variantes do subtipo B, não sendo válidas para outros subtipos devido ao seu fraco desempenho quanto à sensibilidade na análise destas variantes [83].

Tabela 3.1. Desempenho preditivo das regras no genotropismo em relação ao ensaio fenotípico referencial. Seclén *et al.* (2010).

	Total (n = 150)			**Subtipo B (n = 75)**		**Subtipo não-B (n = 75)**	
(%) Regra	**Concordância**	**Sensib.**	**Especif.**	**Sensib.**	**Especif.**	**Sensib.**	**Especif.**
11/25	83	57	89	72	86	33	92
11/24/25	83	60	89	78	86	33	92
Carga Líquida	79	70	81	78	75	58	86
Delobel	82	47	91	56	88	33	94
Garrido	79	80	79	94	74	58	84

O desenvolvimento das cinco regras foi um esforço empreendido na identificação de resíduos de aminoácidos e suas interações na V3 que estivessem envolvidos no tropismo viral. Entretanto, a análise continuada de sequências V3 de variantes X4 e R5 revelou novos padrões associados ao fenômeno. Esta complexidade genética norteou a utilização de métodos mais sofisticados, como SVM e matrizes de pontuação de posição específica (*Position-Specific Scoring Matrices*, PSSM), no desenvolvimento de algoritmos para o genotropismo [11–12].

Baseado no método PSSM, o Web-PSSM possui similaridades com o g2p, como o pareamento de dados fenotípicos e genotípicos do vírus sobre o uso dos correceptores, o alinhamento automático das sequências a serem testadas contra uma V3 referencial, além das análises serem realizadas a partir de arquivos no formato FASTA. Neste caso, o Web-PSSM possui uma processividade maior, uma vez que permite a predição simultânea de até 1.000 sequências V3 [84–85].

No Web-PSSM, cada sequência V3 teste é analisada por um sistema de pontuação definido na etapa de treinamento do algoritmo, durante procedimento de validação cruzada. No sistema, valores são atribuídos a cada um dos aminoácidos, a depender da posição da sequência V3 onde se localizam. Ao final, há a soma dos valores atribuídos pelo algoritmo a cada posição da V3 (em média, 35 posições) a ser testada, e uma pontuação final é obtida. Quanto maior a pontuação, mais a V3 se assemelha às sequências de variantes X4. Pontos de corte específicos nesta pontuação definem a classificação de uma variante como X4 / R5X4 ou R5 [84–85].

Conforme o subtipo viral B ou C da V3 teste, é possível selecionar algoritmos no Web-PSSM com matrizes de pontuação específicas. Há dois algoritmos usados para o diagnóstico de sequências V3 de HIV-1 do subtipo B. Um algoritmo, cuja matriz de pontuação resultante é denominada R5X4, é o resultado de um treinamento com um conjunto de 213 sequências V3 de 177 pacientes. As V3 são oriundas de vírus do subtipo B, classificados fenotipicamente quanto ao tropismo viral pela capacidade ou não de infectarem culturas de células indicadoras. Destas V3, 168 sequências são oriundas de variantes R5, 17 de X4 e 28 de R5X4. O outro algoritmo, cuja matriz é denominada SINSI (SI / NSI), baseia as suas predições em um treinamento com 257 sequências V3 de 107 pacientes. As V3 pertencem a vírus do subtipo B classificados quanto à capacidade ou não da variante de induzir a formação de sincício em células PBMC. Destas V3, 70 sequências são oriundas de variantes SI e 187 de NSI [84].

Para o diagnóstico de vírus do subtipo C, há um segundo algoritmo SINSI, treinado com um conjunto de 279 sequências V3 de 220 indivíduos. As V3 pertencem a vírus do subtipo C, sendo 51 sequências oriundas de variantes SI e 228 de NSI [85]. Para ambos os subtipos B e C, os sistemas de pontuação se revelaram acurados para avaliar a similaridade de uma sequência teste com sequências referenciais oriundas de vírus X4 e R5 [84–85].

A validação clínica do g2p e do Web-PSSM ocorreu após a análise retrospectiva de amostras plasmáticas armazenadas de sujeitos de pesquisa dos estudos clínicos MOTIVATE. A análise demonstrou que os dois métodos genotípicos concordaram significativamente com o Trofile^{TM} na identificação de pacientes que atingiram uma carga viral de < 50 cópias / mm^3 de sangue, após 12 meses do início da terapia de resgate anti-HIV baseada no maraviroque. Tal concordância entre os métodos de genotropismo e o Trofile^{TM} ocorreu, mesmo que as sensibilidades para detectar as variantes X4 pelo g2p, com uma FPR de 5%, e pelo Web-PSSM tenham sido

respectivamente de 63% e 59% quando comparadas à sensibilidade apresentada pelo ensaio fenotípico [14].

Na análise retrospectiva do estudo clínico MERIT, foi demonstrada a habilidade do g2p, sob uma FPR de 5,75%, em distinguir de forma similar ao ES-Trofile^{TM} os pacientes que responderam adequadamente ao tratamento baseado no maraviroc. A referida concordância ocorreu, mesmo que a sensibilidade para a detecção de variantes X4 pelo g2p tenha sido de 55%, quando comparada à sensibilidade do ensaio fenotípico [58]. A validação clínica de ferramentas como o g2p e o Web-PSSM levou a que diferentes guias clínicos nacionais e internacionais de gerenciamento da infecção por HIV incluíssem em suas diretrizes o uso de algoritmos para o genotropismo viral [52, 86–89].

Sierra *et al.* (2015) se dedicaram a uma análise mais ampla sobre a relação do ajuste da FPR do g2p com a resposta virológica ao tratamento de resgate baseado no maraviroque. A partir de um estudo prospectivo de amostras clínicas, os autores demonstraram que havia a manutenção de resposta virológica adequada em indivíduos submetidos ao tratamento de resgate com o maraviroque, independentemente se selecionados com o g2p sob o ajuste de FPR tão distintas quanto no intervalo de 1 a 20%. Porém, foi ressaltado que as FPR no intervalo de 5 a 7,5% asseguraram tanto uma resposta virológica adequada, quanto a minimização da exclusão de indivíduos que poderiam se beneficiar do tratamento de resgate [76].

Em paralelo ao uso crescente do g2p e Web-PSSM, outras regiões do envelope viral foram analisadas quanto à influência no genotropismo do HIV-1. Cashin *et al.* (2014) demonstraram uma associação significativa entre aminoácidos carregados positiva ou negativamente, em pH fisiológico, localizados nas posições 322 da região V3 e 440 da C4 na gp120. A análise foi realizada em todas as sequências do envelope viral de subtipo B disponíveis no Los Alamos e caracterizadas fenotipicamente quanto ao tropismo viral: 43 com tropismo pelo CXCR4 (23 R5X4 e 20 X4) e 223 R5. Variações significativas de tamanho entre as sequências impediram o alinhamento adequado e a análise subsequente das regiões V1, V2, V4 e V5. Porém, verificou-se que os aminoácidos K e R carregados positivamente, na posição 322, e asparagina (N) e glutamina (Q), carregados negativamente, na posição 440, ocorreram mais frequentemente em variantes X4, enquanto que N e Q, na posição 322, e R, na posição 440, ocorreram mais frequentemente em variantes R5 [90].

Os autores verificaram que a inclusão desta informação, denominada "regra 440", melhorou a sensibilidade de várias ferramentas genotípicas, como regra 11/25, g2p e Web-PSSM, sem comprometer a especificidade no genotropismo do HIV-1 subtipo B. Dentre os ajustes de FPR do g2p no intervalo de 1 a 20%, o uso da regra 440 associada a uma FPR de 5,75% promoveu ainda uma melhoria significativa na medida de desempenho AUC. Embora tenha funcionado com variantes de HIV-1 do subtipo B, a regra 440 não desempenhou da mesma forma com variantes de outros subtipos [90].

Além da análise das cinco regras, Seclén *et al.* (2010) analisaram também outras ferramentas de genotropismo, incluindo o g2p, ajustado sob diferentes FPR, e os dois algoritmos do Web-PSSM com as matrizes de pontuação associadas ao genotropismo de variantes do subtipo B: R5X4 e SINSI-B. De forma semelhante àquela verificada nas cinco regras, os resultados mostraram o bom desempenho dos algoritmos principalmente para o genotropismo de variantes do subtipo B. Já os resultados de sensibilidade para variantes não-B foram bem inferiores àqueles relacionados a variantes do subtipo B (Tabela 3.2) [83].

Tabela 3.2. Desempenho preditivo dos algoritmos g2p e Web-PSSM no genotropismo em relação ao ensaio fenotípico referencial. Seclén *et al.* (2010).

	Total (n = 150)			**Subtipo B (n = 75)**		**Subtipo não-B (n = 75)**	
(%) Algoritmo	**Concordância**	**Sensib.**	**Especif.**	**Sensib.**	**Especif.**	**Sensib.**	**Especif.**
Web-PSSM							
R5X4	85	77	87	89	86	58	87
SINSI-B	83	73	86	89	81	50	90
G2P							
FPR 1%	79	23	93	28	91	17	94
FPR 2,5%	83	70	86	83	79	50	92
FPR 5%	77	80	77	94	68	58	84
FPR 10%	69	80	66	94	51	58	79
FPR 15%	68	83	64	94	49	67	78
FPR 20%	63	83	58	94	46	67	70

Frente ao desempenho insatisfatório do g2p no genotropismo de variantes do HIV-1 de subtipos não-B, pesquisadores buscaram avaliar mais detalhadamente a capacidade do algoritmo, entre outras ferramentas de genotropismo, em analisar sequências V3 de subtipos não-B. Riemenschneider *et al.* (2016) demonstraram uma

alta acurácia do g2p no genotropismo de sequências de vírus do subtipo C. Em relação à referência fenotípica, constituída de um conjunto de 56 sequências V3 de variantes X4 e 359 de R5 oriundas de HIV-1 do subtipo C, o g2p sob uma FPR de 5% apresentou uma sensibilidade de 87,5% e especificidade de 97,8%. Dentre os algoritmos do Web-PSSM, a matriz SINSI-B apresentou uma acurácia (95,2%) ainda melhor que as matrizes R5X4 (91,9%) e SINSI-C (91,2%), com sensibilidade de 71,4%, contra respectivamente 75% e 89,3%. A especificidade da SINSI-B foi de 98,9%, contra respectivamente 94,5% e 91,5% [91].

Por outro lado, os métodos genotípicos não apresentaram um bom resultado na predição do tropismo em sequências oriundas de HIV-1 do subtipo A e suas CRF, notadamente a CRF02_AG. Neste caso, a referência fenotípica foi um conjunto de 209 sequências V3 de variantes X4 e 190 de R5. A despeito do bom desempenho obtido quanto à especificidade pelo g2p, sob o ajuste de FPR de 5% (97,9%), e pelas matrizes R5X4 (93,9%) e matriz SINSI-B (98%), suas sensibilidades foram respectivamente de 15,8%, 15,3 e 11,5% [91]. A Tabela 3.3 sumariza os resultados da análise dos subtipos C e A mencionados acima.

Tabela 3.3. Desempenho preditivo dos algoritmos g2p e Web-PSSM no genotropismo em relação ao referencial fenotípico. Riemenschneider *et al.* (2016).

Subtipo	Algoritmo	Acurácia (%)	Sensibilidade (%)	Especificidade (%)
C	**Web-PSSM**			
	R5X4	91,9	75,0	94,5
	SINSI-B	95,2	71,4	98,9
	SINSI-C	91,2	89,3	91,5
	G2P			
	FPR 5%	96,4	87,5	97,8
A	**Web-PSSM**			
	R5X4	53,6	15,3	93,9
	SINSI-B	53,7	11,5	98,0
	SINSI-C	47,9	37,8	58,6
	G2P			
	FPR 5%	54,4	15,8	97,9

Jeanne *et al.* (2015) estudaram certas limitações técnicas associadas ao NGS e suas repercussões no genotropismo: a geração de artefatos durante o NGS e a capacidade de diferenciá-los de sequências com informação biológica relevante [92]. Procedimentos padrão do NGS tratam a questão pelo estabelecimento de corte arbitrários de sensibilidade de 1 a 2%, abaixo dos quais variantes sequenciais minoritárias são descartadas. Porém, os erros gerados durante o NGS estão mais associados a características das próprias sequências a serem sequenciadas do que a aspectos técnicos e aleatórios do procedimento. As falhas são mais comuns em regiões homopoliméricas – partes das sequências com repetição de um mesmo nucleotídeo – também presentes no envelope viral [93].

Com o intuito de minimizar este problema, os autores desenvolveram um programa que processa automaticamente as sequências do envelope viral de HIV-1 oriundas do NGS. O programa tem dois módulos que atuam como filtros de sequências: um biológico e outro estatístico. O biológico descarta sequências sem função biológica, principalmente aquelas que apresentam artefatos na forma de erros de leitura na sequência, imputadas pela inserção ou deleção de nucleotídeos, principalmente em regiões homopoliméricas da sequência. O segundo filtro, estatístico, utiliza-se de propriedades associadas à distribuição de Poison para descartar artefatos nas sequências na forma de mutações pontuais. Com resultados significativamente concordantes com um ensaio fenotípico de alta sensibilidade, o programa de processamento de sequências conseguiu replicar a detecção de variantes X4 minoritárias em três misturas artificiais de vírus com proporções conhecidas de variantes X4 e R5 [92].

Métodos como RL, NB e RF também foram utilizados em estudos sobre o tropismo do HIV-1. Schapiro *et al.* (2011), utilizando o método de RL, propuseram um critério mais robusto para a indicação de indivíduos que se beneficiariam do tratamento baseado no maraviroque. Para isto, os autores desenvolveram dois sistemas de pontuação para medir a susceptibilidade das variantes de HIV-1 de cada indivíduo aos antirretrovirais. As duas pontuações se basearam em teste fenotípico ou genotípico de resistência medicamentosa, em associação aos dados de uso prévio das drogas anti-HIV. Nestas pontuações, houve a consideração da contribuição específica de cada droga na capacidade de reduzir a carga viral do HIV-1, ao contrário da metodologia tradicional de avaliação de susceptibilidade também analisada no estudo [56].

Os autores verificaram a resposta virológica associada aos três métodos, considerando como covariáveis preditoras as medições de suscetibilidade balanceadas

(genotípica ou fenotípica) ou sem balanço (tradicional), uso do maraviroque, contagem de células T $CD4^+$, carga viral e tropismo do HIV-1, em modelagens utilizando o método de RL. Os autores verificaram que os métodos balanceados são similares entre si, e superiores à metodologia tradicional na predição de susceptibilidade virológica aos antirretrovirais. As variáveis preditoras que apresentaram uma associação mais significativa com a resposta virológica ao tratamento de resgate com o maraviroque foram ambas as novas pontuações balanceadas, além da contagem de células T $CD4^+$ [56].

Arif *et al.* (2017) utilizaram o método de RL para analisar possíveis preditores para a mudança do tropismo de R5 para X4, a partir de amostras de DNA proviral de 66 pacientes avaliados longitudinalmente. Para isto, as sequências V3 obtidas foram analisadas pelo g2p, e os resultados da predição, na forma de FPR, foram utilizados como uma das variáveis preditoras analisadas em modelagens com RL. Foi demonstrado que os pacientes com sequências V3 com resultado igual ou superior à FPR de 40,6% tenderam a manter estas FPR estáveis ao longo do tempo. Em contrapartida, indivíduos com vírus relacionados a FPR menores que 40,6% tenderam a apresentar um decaimento progressivo destas FPR, dando origem a variantes com tropismo pelo CXCR4, em um tempo de evolução médio de 27,3 meses (8,9 a 64,6 meses). Os autores concluíram que uma FPR igual ou acima a 40,6% pode ser um indicador para que estes pacientes sejam dispensados da realização de genotropismo adicional com o intuito de verificar a progressão da doença e/ou embasar a tomada de decisão clínica para o uso de tratamento baseado no maraviroc [94].

Díez-Fuertes *et al.* (2013) desenvolveram uma ferramenta de genotropismo baseada em um classificador Bayesiano, que permite relações de dependência entre as variáveis preditoras, no caso, 26 posições nucleotídicas ao longo de todo o gene *env*. O classificador foi gerado a partir de um procedimento de validação cruzada, utilizando sequências oriundas de variantes de diversos subtipos do HIV-1, classificados fenotipicamente quanto ao tropismo viral. Na validação, o algoritmo apresentou uma acurácia de 95,6%, uma especificidade de 99,4% e uma sensibilidade de 80,5%. Os autores relataram ainda um desempenho preditivo do algoritmo significativamente melhor que aqueles apresentados pelo g2p e Web-PSSM [95].

Xu *et al.* (2007) colaboradores usaram o método de RF para predizer o tropismo do HIV-1, além da predição da característica SI / NSI, usando 37 variáveis, incluindo as 35 posições de aminoácidos da V3, a carga líquida total e informações sobre a

polaridade da sequência. A partir de treinamento e teste com sequências V3 oriundas de subtipos B, C, entre outros (não-B e não-C), o algoritmo de RF desenvolvido apresentou melhor desempenho preditivo em relação ao Web-PSSM tanto na predição do tropismo viral, quanto da característica SI / NSI [96]. A Tabela 3.4 sumariza os resultados do estudo.

Tabela 3.4. Desempenho dos algoritmos de RF e Web-PSSM no genotropismo e predição da característica SI / NSI em relação ao referencial fenotípico. Xu *et al.* (2007).

Subtipo	Algoritmo	Acurácia (%)	Sensibilidade (%)	Especificidade (%)
		Tropismo Viral		
B	**RF**	94	80	98,3
	Web-PSSM	89,4	60	98,3
C	**RF**	96,7	76,5	100
	Web-PSSM	92,5	76,5	95,1
Não-B e não-C	**RF**	96,6	96,4	96,7
	Web-PSSM	86,3	85,5	86,8
		SI / NSI		
B	**RF**	90,6	90,9	90,4
	Web-PSSM	89,6	87	91,1
C	**RF**	92	80	98,9
	Web-PSSM	90,3	60	92,5
Não B e C	**RF**	95,3	92,3	96,3
	Web-PSSM	90,7	88,5	91,4

Heider *et al.* (2014) apresentaram o T-CUP 2.0, um modelo que é o resultado de classificadores de RF gerados em dois níveis de aprendizado, a partir de procedimentos de validação cruzada em 1351 sequências V3, classificadas quanto ao tropismo fenotípico. Destas, 200 são oriundas de variantes com tropismo pelo CXCR4 (34 R5X4, 166 X4) e 1151 de variantes R5. A maioria das sequências é de HIV-1 do subtipo B (52%), além do subtipo C (17%) e subtipo D (9%), com 22% das sequências pertencendo a outros subtipos. No primeiro nível, classificadores foram gerados a partir de treinamento utilizando preditores numéricos relacionados ao potencial eletrostático

da V3 – informação estrututal da sequência baseada nas suas alças –, e à hidrofobicidade, a partir da escala de KyteDoolittle. O segundo nível de aprendizado combinou a informação estrutural e sequencial contida nos resultados fornecidos por ambos os classificadores, para o treinamento e teste do novo algoritmo RF. Com esta abordagem, os autores conseguiram resultados bem superiores a algoritmos como o g2p, principalmente no que concerne à sensibilidade, ressaltando assim o potencial preditivo da combinação das informações estruturais e sequenciais da V3 em modelagens com o método de RF [68].

Em abordagem semelhante àquela utilizada por Heider *et al.* (2014) [68], Lochel *et al.* (2018) desenvolveram um algoritmo baseado no método de RF, com dois níveis de aprendizado, combinando informações estruturais e sequenciais da V3, obtendo boa acurácia no genotropismo do subtipo A e suas CRF. Para isto, os autores utilizaram um conjunto de 182 sequências V3 do subtipo A e CRF, sendo 147 de variantes R5 e 35 de variantes com tropismo pelo CXCR4. O bom desempenho obtido quanto à especificidade foi similar aos outros algoritmos como o g2p e o Web-PSSM. No que tange à sensibilidade, enquanto o g2p apresentou um resultado de 15,8%, e os algoritmos do Web-PSSM – R5X4, SINSI-B e SINSI-C – obtiveram respectivamente os resultados de 15,3%, 11,5% e 37,8%, o algoritmo RF desempenhou de forma significativamente superior, apresentando uma sensibilidade de 47,7% [97].

Capítulo 04

Materiais e Métodos

4.1. Conjunto de Dados

O conjunto de dados foi obtido em janeiro de 2018 junto ao banco de Los Alamos, cujo endereço é https://www.hiv.lanl.gov/content/index. Trata-se de uma amostra composta por 2.333 sequências de DNA de HIV-1 subtipo B, relativas à região V3 da gp120 viral e com tamanho de até 105 nucleotídeos.

Dentro do Los Alamos, o caminho para se chegar ao conjunto de dados foi "*Sequence Database* / *Programs and Tools* / *Search Interface"*, onde foram selecionadas as informações relacionadas ao vírus (HIV-1), subtipo (B), tamanho da sequência quanto ao número de nucleotídios (105) e região genômica de interesse (V3). A seguir, há o exemplo de duas sequências V3 obtidas no formato FASTA fornecido pelo sítio (Tabela 4.1).

Tabela 4.1. Exemplos de duas V3 obtidas no sítio Los Alamos, no formato FASTA fornecido, com suas identificações e sequências de DNA (105 nucleotídeos).

>Identificação da V3 **Sequência Nucleotídica**
>B.JP.-.SUBJECT_4.AB001137 TGTACAAGACCCAACAACAATACA---AGAAAAGGTATAAATATA------GGACCAGGGAGAG---CA---TTATTTTA TGCAACA---GACATAATAGGAGATATAAGACAAGCACATTGT
>B.JP.-.SUBJECT_4.AB001142 TGTACAAGACCCAACAACAATACA---AGAAAAGGTATACATATA------GGACCAGGGAGAG---CAGTATT---TTA TGCAACA---GACATAATAGGAGATATAAGACAAGCACATTGT

4.1.1. Predição do Tropismo Viral pelo Geno2pheno

As 2.333 sequências foram submetidas ao genotropismo pelo g2p, cujo endereço é https://coreceptor.geno2pheno.org/. As análises no algoritmo foram realizadas sob uma probabilidade de FPR de 10%, que está dentro da faixa de FPR recomendada por Sierra *et al.* (2015) e por guias europeus de boas práticas clínicas [76, 89]. A predição

associada a uma FPR igual ou abaixo de 10% foi indicativa de que a sequência era X4 ou R5X4, denominadas a partir daqui como não-R5 (NR5). Acima desta FPR de 10%, a sequência foi classificada como R5.

Para a predição das sequências contidas nos arquivos, foram selecionadas as informações relacionadas à escolha do método validado de predição ("*original g2p coreceptor*"), nível de significância ("*false positive rate* = 10%"), escolha do arquivo FASTA em questão ("escolher arquivo"), e a solicitação da predição propriamente dita ("*action: align and predict"*). A seguir, há o exemplo do resultado da análise pelo g2p das duas V3 apresentadas na Tabela 4.1. Além de fazer a predição, o algoritmo também forneceu as sequências, convertidas em aminoácidos (peptídeos), em arquivos digitais no formato csv (Tabela 4.2).

Tabela 4.2. Resultados da análise pelo g2p das V3 apresentadas na Tabela 4.1, com as identificações, sequências de aminoácidos e FPR. As classificações não integraram o arquivo digital csv com os resultados citados.

Identificação	Sequência V3 (Peptídica)	Subtipo	FPR	Classif.
B.JP.-.SUBJECT_4.AB001137	CTRPNNNTRKGINIGPGRALFYATDIIGDIRQAHC	B	29,8	R5
B.JP.-.SUBJECT_4.AB001142	CTRPNNNTRKGIHIGPGRAVFYATDIIGDIRQAHC	B	8,1	NR5

O g2p não conseguiu classificar três sequências de DNA, não reconhecidas como V3, que foram retiradas da análise. Assim, foram obtidas 2.330 sequências peptídicas relativas à região V3, classificadas cada uma quanto ao tropismo viral pelo algoritmo. Todas as sequências apresentavam até 35 posições de aminoácidos, codificados por letras conforme a simbologia internacional (Tabela 4.3). As etapas seguintes foram realizadas com o uso do programa R, versão 3.4.3.

4.1.2. Conjunto de Dados para a Modelagem

As sequências peptídicas de V3 que possuíam mais de um aminoácido em uma dada posição e/ou tamanho menor do que 35 aminoácidos foram removidas da análise. Dentre as 2.109 sequências peptídicas que permaneceram, com 1.110 sequências distintas entre si, 568 eram NR5 (classe positiva) e 1541 eram R5.

As 35 posições de aminoácidos em cada uma das 2.109 sequências V3, nomeadas aqui de "Posição 1" (P1) a P35, compuseram as 35 variáveis preditoras

utilizadas na modelagem. Os aminoácidos foram então convertidos em dados numéricos, conforme as escalas de hidrofobicidade de Eisenberg, Guy ou KyteDoolittle (Tabela 4.3).

Tabela 4.3. Valores de hidrofobicidade atribuídos aos 20 aminoácidos pelas escalas de Eisenberg, Guy e KyteDoolittle (KD).

Aminoácidos	Símbolo	Eisenberg	Guy	KD
Alanina	A	0,62	0,10	1,80
Cisteína	C	0,29	-1,42	2,50
Ácido Aspártico	D	-0,90	0,78	-3,50
Ácido Glutâmico	E	-0,74	0,83	-3,50
Fenilalanina	F	1,19	-2,12	2,80
Glicina	G	0,48	0,33	-0,40
Histidina	H	-0,40	-0,50	-3,20
Isoleucina	I	1,38	-1,13	4,50
Lisina	K	-1,50	1,40	-3,90
Leucina	L	1,06	-1,18	3,80
Metionina	M	0,64	-1,59	1,90
Asparagina	N	-0,78	0,48	-3,50
Prolina	P	0,12	0,73	-1,60
Glutamina	Q	-0,85	0.83	-3,50
Arginina	R	-2,53	1,91	-4,50
Serina	S	-0,18	0,52	-0,80
Treonina	T	-0,05	0,07	-0,70
Valina	V	1,08	-1,27	4,20
Triptofano	W	0,81	-0,51	-0,90
Tirosina	Y	0,26	-0,21	-1,30

4.1.2.1. Conjuntos de Treinamento e Teste

Com base na validação pelo método *Holdout* [57], as sequências foram aleatoriamente divididas em dois conjuntos mutuamente exclusivos, sob a proporção usual de 70% dos dados para o conjunto de treinamento e 30% para o conjunto teste. Seguindo esta proporção de 0,7:0,3, as 2.109 sequências foram então divididas em um conjunto de treinamento composto por uma amostra de 1.477 sequências, com 398 NR5 e 1.079 R5, enquanto que o conjunto de teste resultou em uma amostra de 632 sequências, 170 NR5 e 462 R5.

O balanceamento dos dados de treinamento foi obtido pela divisão aleatória das 1.079 sequências R5 em três subconjuntos, contendo respectivamente 360, 360 e 359 sequências. Cada um destes subconjuntos foi combinado ao mesmo subconjunto de 398 sequências NR5, apresentando uma proporção final de 0,47:0,53. Ao término deste procedimento, para cada uma das escalas de hidrofobicidade utilizadas nas modelagens foram obtidos três conjuntos de treinamento A, B e C.

Durante o método *Holdout* e no preparo dos subcontuntos de treinamentos, a aleatoriedade inicialmente utilizada para a divisão das sequências foi replicada nos demais experimentos pelo uso da função "set.seed()" do pacote simEdv1.0.3 do R, resultando sempre nos mesmos conjuntos e subconjuntos de treino e teste de V3.

4.1.2.2. Tratamento das Variáveis Explicativas

Os conjuntos de treinamento foram analisados separadamente acerca das variáveis preditoras (P1 a P35) que apresentaram um mesmo aminoácido em mais de 95% ou de 98% das sequências. As Figuras 4.1 a 4.3 ilustram as frequências dos aminoácidos que mais se repetiram em cada uma das 35 variáveis dos subconjuntos de treinamento A, B e C, "Eisenberg", conforme os dois escrutínios de 95% e 98%. Deve ser ressaltado que a escala de Eisenberg possui um valor de hidrofobicidade específico para cada aminoácido, conforme mostra a Tabela 4.3, permitindo assim a análise da frequência de aminoácidos mencionada acima.

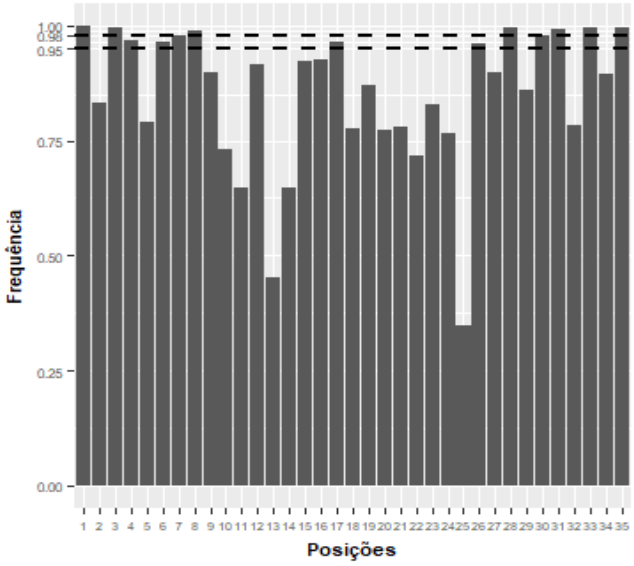

Figura 4.1. Frequências dos aminoácidos que mais se repetiram em cada uma das 35 variáveis do subconjunto de treinamento A "Eisenberg". Os escrutínios de 95% e 98% estão destacados.

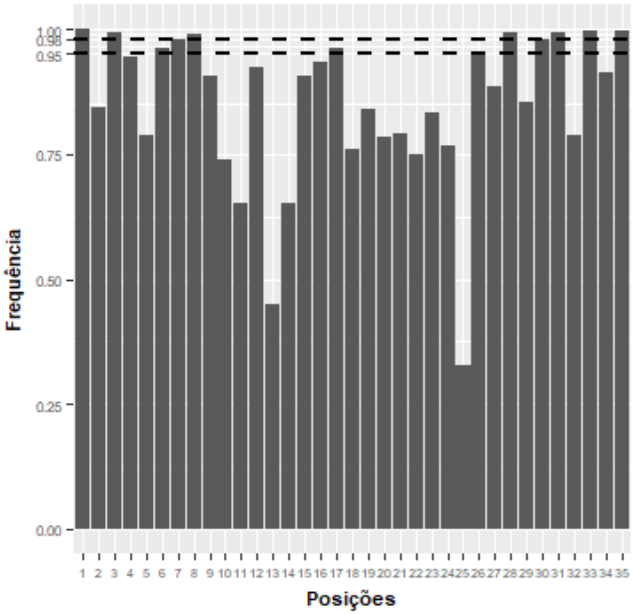

Figura 4.2. Frequências dos aminoácidos que mais se repetiram em cada uma das 35 variáveis do subconjunto de treinamento B “Eisenberg”. Os escrutínios de 95% e 98% estão destacados.

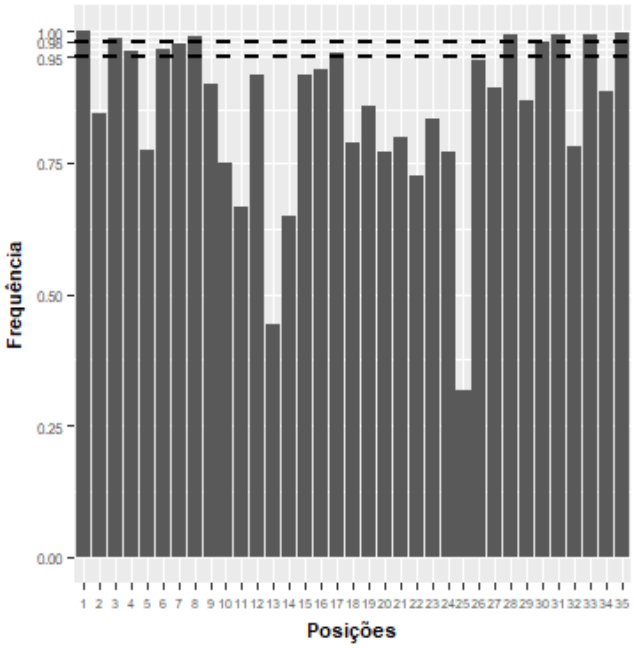

Figura 4.3. Frequências dos aminoácidos que mais se repetiram em cada uma das 35 variáveis do subconjunto de treinamento C “Eisenberg”. Os escrutínios de 95% e 98% estão destacados.

A seguir, a Tabela 4.4 sumariza as variáveis que apresentaram um mesmo aminoácido que se repetiu em mais de 95% ou 98% das sequências V3, em cada subconjunto "Eisenberg" A, B e C.

Tabela 4.4. Variáveis P1 a P35 que apresentaram um mesmo aminoácido repetido em mais de 95% ("95") e 98% ("98") das sequências V3, em cada subconjunto de treinamento "Eisenberg" A, B e C.

Subconjunto	Escrutínio	
	95	98
A	P1, P3, P4, P6, P7, P8, P17, P26, P28, P30, P31, P33 e P35	P1, P3, P8, P28, P31, P33 e P35
B	P1, P3, P6, P7, P8, P17, P26, P28, P30, P31, P33 e P35	P1, P3, P8, P28, P30, P31, P33 e P35
C	P1, P3, P4, P6, P7, P8, P17, P28, P30, P31, P33 e P35	P1, P3, P8, P28, P31, P33 e P35

Com o intuito de selecionar os modelos mais parcimoniosos possíveis, os modelos com as variáveis P1 a P35, ditos "completos", foram comparados com os modelos sem as respectivas variáveis assinaladas na Tabela 4.4, ou seja, "95" e "98". Utilizando o método de RL, e considerando o modelo "completo" sempre como referência, a comparação foi realizada a partir do teste da razão de verossimilhança, com o uso das *deviances* residuais relativas aos ajustes dos modelos a cada um dos subconjuntos de treinamento. Neste teste de hipótese, a H_0 se referiu a não existência de diferença de ajuste entre os modelos "95" ou "98" em relação ao "completo". Para valores $p < 0,05$, a hipótese nula de igualdade de ajuste foi rejeitada em favor de um melhor ajuste do modelo "completo".

Seis dentre nove modelos "95" apresentaram um ajuste inferior aos modelos "completos" correspondentes ($p < 0,05$). Com exceção de um único modelo "98" ("Guy" e treinado com o subconjunto C), onde a H_0 foi rejeitada ($p = 0,026$), os outros oito modelos "98" apresentaram um ajuste similar aos modelos "completos" correspondentes ($p > 0,05$). Assim, os conjuntos de variáveis "98" foram aqueles selecionados para a modelagem subsequente do tropismo do HIV-1 pelos métodos de RL, NB e RF. As Tabelas 4.5 a 4.7 apresentam os desfechos dos testes de hipótese que embasaram esta escolha.

Tabela 4.5. Comparação do ajuste dos modelos de RL "completos", "Eisenberg", frente aos ajustes dos modelos de RL "95" e "98", "Eisenberg". H_0 refere-se à não existência de diferença de ajuste entre os modelos, sendo rejeitada quando $p < 0,05$ em prol do melhor ajuste do modelo completo.

Sub-conjunto de Treino	**A**			**B**			**C**		
Medida \ Modelo	**Completo**	**95**	**98**	**Completo**	**95**	**98**	**Completo**	**95**	**98**
Valor P	**Referência**	**p = 0,003009**	**p = 0,2016**	**Referência**	**p = 0,05106**	**p = 0,3116**	**Referência**	**p = 0,01019**	**p = 0,07323**
***Deviance* Residual**	562,8	594,2	572,6	547,7	567,3	556,0	546,2	570,8	557,7
AIC	634,8	640,2	630,6	617,7	615,3	612,0	616,2	618,8	615,7

Tabela 4.6. Comparação do ajuste dos modelos de RL "completos", "Guy", frente aos ajustes dos modelos de RL "95" e "98", "Guy". H_0 refere-se à não existência de diferença de ajuste entre os modelos, sendo rejeitada quando $p < 0,05$ em prol do melhor ajuste do modelo completo.

Sub-conjunto de Treino	**A**			**B**			**C**		
Medida \ Modelo	**Completo**	**95**	**98**	**Completo**	**95**	**98**	**Completo**	**95**	**98**
Valor P	**Referência**	**p = 0,00059**	**p = 0,1689**	**Referência**	**p = 0,3857**	**p = 0,5682**	**Referência**	**p = 0,0185**	**p = 0,02615**
***Deviance* Residual**	543,1	579,1	553,5	512,4	524,1	518,1	537,8	560,7	552,2
AIC	615,1	625,1	611,5	582,4	572,1	574,1	607,8	608,7	610,2

Tabela 4.7. Comparação do ajuste dos modelos de RL "completos", "KyteDoolittle", frente aos ajustes dos modelos RL "95" e "98", "KyteDoolittle". H_0 refere-se à não existência de diferença de ajuste entre os modelos, sendo rejeitada quando $p < 0,05$ em prol do melhor ajuste do modelo completo.

Sub-conjunto de Treino	**A**			**B**			**C**		
Medida \ Modelo	**Completo**	**95**	**98**	**Completo**	**95**	**98**	**Completo**	**95**	**98**
Valor P	**Referência**	**p = 4,333.e^{-07}**	**p = 0,1154**	**Referência**	**p = 7,472.e^{-05}**	**p = 0,1521**	**Referência**	**p = 0,06593**	**p = 0,06831**
***Deviance* Residual**	486,7	541,5	498,2	497,4	535,5	508,1	514,9	533,6	526,6
AIC	558,7	587,5	556,2	567,4	583,5	564,1	584,9	581,6	584,6

4.2. Modelagem

As modelagens foram realizadas no programa R, versão 3.4.3. Para cada método de modelagem, foram gerados três grupos, baseados nas escalas de hidrofobicidade – Eisenberg, Guy e KyteDoolittle –, de classificadores. Dentro de cada grupo, três classificadores – A, B e C – foram construídos e assim nomeados pelo treinamento com os subconjuntos balanceados de sequências A, B e C, contendo a informação sequencial da V3 na forma de preditores numéricos da escala de hidrofobicidade correspondente ao grupo em questão. Cada conjunto de três classificadores – A, B e C – foi testado com um mesmo conjunto teste, contendo a informação sequencial da V3 da escala correspondente ao grupo. Para as modelagens, as variantes NR5 foram selecionadas como a classe positiva. A Figura 4.4 apresenta os 27 modelos "98" utilizados na presente dissertação.

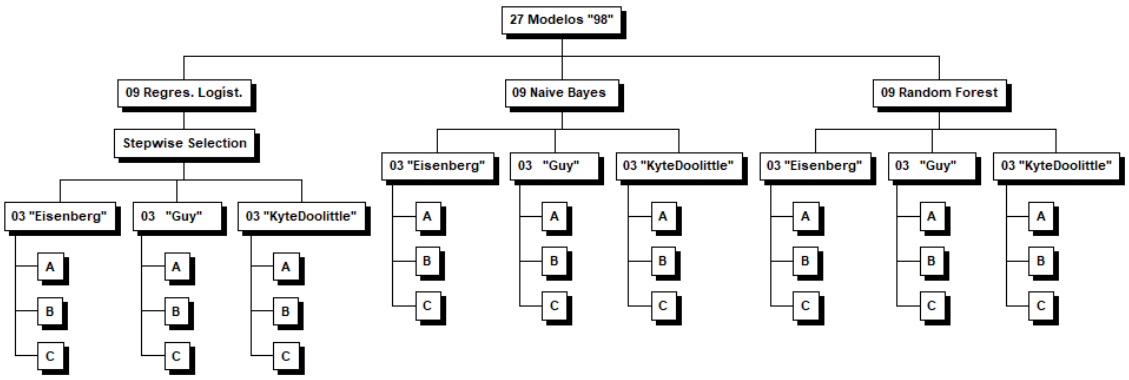

Figura 4.4. Sumário dos modelos, com as variáveis "98", utilizados na análise do tropismo do HIV-1 de subtipo B.

4.2.1. Regressão Logística

Os treinamentos pelo método de RL foram realizados com o uso da função "glm()" do pacote *stats* do R, utilizada para o ajuste de modelos lineares generalizados. Cada modelo foi então submetido à seleção de variáveis pelo método iterativo *stepwise*, com abordagem *backward*, via função "step()" do pacote *stats*.

Conforme a abordagem *backward*, a seleção foi sempre iniciada com todas as variáveis independentes do modelo em questão, e testadas uma a uma para serem excluídas sempre que não houvesse uma perda do ajuste [69]. A função "step()" compara os modelos, com e sem uma dada variável a ser testada, utilizando o AIC como critério de avaliação do ajuste de cada modelo.

Na etapa de análise de desempenho dos modelos de RL, as predições foram realizadas pelo uso da função "predict()" do pacote *stats*, que utiliza um dado modelo ajustado para a análise de um conjunto de teste específico. Os objetos criados nesta função foram utilizados para a obtenção das respectivas curvas ROC e matrizes de confusão. Neste sentido, cada curva ROC foi obtida pelo uso da função "roc()" do pacote *pROC*, enquanto que as matrizes de confusão são o resultado do uso da função "confusionMatrix()" do pacote *caret*.

Assim, os modelos de RL parcimoniosos, oriundos da seleção de variáveis, foram analisados quanto ao seu poder preditivo com base nos resultados das medidas de acurácia, sensibilidade, especificidade e AUC. Intervalos de confiança de 95% (IC95%) para cada uma das medidas foram obtidos a partir de 2.000 réplicas de *bootstrap* estratificado, que se tratam de estimativas pontuais obtidas a partir de reamostragens com reposição para a construção de uma estimativa intervalar de valores. ICs95% para acurácia foram um dos resultados previamente obtidos com a função "confusionMatrix()". ICs95% para sensibilidade e especificidade foram gerados com o uso da função "ci.thresholds()", enquanto que ICs95% de AUC foram gerados pela função "ci.auc()". Ambas as funções são oriundas do pacote *pROC*. A aleatoriedade contida na geração das réplicas de *bootstrap* foi replicada no método de RL e demais classificadores pelo uso da função "set.seed()".

No R, a análise de cada sequência V3 pelo método de RL resulta em uma estimativa da probabilidade do evento "NR5" ocorrer, em função do conjunto de 35 variáveis preditoras P1 a P35.

4.3.2. *Naive* Bayes

Os treinamentos pelo método NB foram realizados com o uso de função "naiveBayes()" do pacote *e1071* do R. Usando o teorema de Bayes, a função computou as probabilidades a posteriori da classe NR5, dadas as variáveis independentes de treinamento.

Com o uso das mesmas funções utilizadas para o método de RL na parte de análise de desempenho, os modelos de NB foram então analisados quanto ao seu poder preditivo com base nos resultados das medidas de acurácia, sensibilidade, especificidade e AUC. Para cada uma das medidas, ICs95% foram obtidos a partir de 2.000 réplicas de *bootstrap* estratificado.

No R, a análise de cada V3 pelo método NB resulta em estimativas das probabilidades complementares dos eventos "NR5" e "R5" ocorrerem, em função do conjunto de 35 variáveis preditoras P1 a P35.

4.3.3. *Random Forest*

Os treinamentos pelo método de RF foram realizados com o uso da função "randomForest()", do pacote "*randomForest*" do R, que implementa o algoritmo original de RF desenvolvido por Breiman (73). Durante cada treinamento, 500 árvores de classificação foram geradas, cada uma composta por 05 variáveis explicativas selecionadas de forma aleatória. A aleatoriedade contida na geração das primeiras 500 árvores de classificação foi replicada nas demais modelagens com o método de RF pelo uso da função "set.seed()".

Com o uso das mesmas funções utilizadas para o método de RL na parte de análise de desempenho, os modelos de RF foram então analisados quanto ao seu poder preditivo com base nos resultados das medidas de acurácia, sensibilidade, especificidade e AUC. Para cada uma das medidas, ICs95% foram obtidos a partir de 2.000 réplicas de *bootstrap* estratificado.

No R, a análise de cada V3 pelo método de RF resulta em estimativas das probabilidades complementares dos eventos "NR5" e "R5" ocorrerem, em função do conjunto de 35 variáveis preditoras P1 a P35.

Capítulo 05

Resultados

5.1. Regressão Logística

5.1.1. Treinamento e Seleção de Variáveis

Após o ajuste com cada subconjunto de treinamento, os modelos de RL foram submetidos à etapa de seleção de variáveis. Para cada escala de hidrofobicidade utilizada, os modelos parcimoniosos resultantes foram designados como A, B e C, a depender da origem do subconjunto de treinamento, conforme previamente ilustrado (Figura 4.4). As Tabelas 5.1 a 5.3 apresentam as variáveis selecionadas dos modelos de RL, com os seus respectivos coeficientes, bem como as suas significâncias quanto à associação positiva ou negativa com a classe positiva NR5.

5.1.2. Análise do Desempenho Preditivo

As matrizes de confusão dos modelos de RL apresentam resultados referentes ao uso de dois pontos de corte: 0,5 e ótimo, que fornece a maior soma dos valores pontuais de sensibilidade e especificidade. As matrizes estão ilustradas nas Tabelas 5.4 a 5.6.

As curvas ROC dos modelos estão demonstradas na Figura 5.1. As medidas de desempenho de acurácia, sensibilidade, especificidade e AUC, além dos seus respectivos IC95%, estão apresentadas na Tabela 5.7. Para acurácia, sensibilidade e especificidade, as estimativas pontuais e intervalares foram realizadas usando o ponto de corte ótimo.

Tabela 5.1. Modelos de RL, "Eisenberg", após seleção s*tepwise,* com as variáveis selecionadas com os respectivos coeficientes e significâncias quanto à existência de associação com a classe positiva NR5, indicada por um valor p < 0,05. As variáveis com p < 0,001 estão identificadas como "***"; 0,001 < p < 0,01, "**"; 0,01 < p < 0,05, "*"; 0,05 < p < 0,1, "•"; e 0,1 < p, "‒". Para os 03 modelos, as variáveis P2, P5 e P26 não foram selecionadas pela *stepwise*. NA significa que a referida variável não foi selecionada para o modelo parcimonioso. No consenso, NC refere-se à inexistência de uma mesma variável nos três modelos com um valor p de, ao menos, 0,01 < p < 0,05, "*". Nas demais variáveis, a menor significância dentre os três modelos é destacada. Os valores de *deviance* residual e AIC de cada modelo estão respectivamente apresentados entre parênteses.

Modelo \ Preditor	**Intercept**	**P4**	**P6**	**P7**	**P9**	**P10**	**P11**	**P12**	**P13**	**P14**	**P15**	**P16**	**P17**
A (577,6; 623,6)	7,13	-2,79	1,96	7,54	1,23	0,46	-0,85	NA	1,29	-0,84	-3,84	1,11	-1,28
Significância	*	*	**	*	***	*	***		***	*	•	*	‒
B (559,5; 601,5)	5,97	-2,35	1,65	7,18	1,46	0,45	-0,71	-2,33	1,28	-0,77	NA	1,57	NA
Significância	*	*	*	*	***	*	**	*	***	*		**	
C (567,8; 603,8)	4,20	-3,70	NA	NA	0,70	0,77	-0,68	-3,95	1,11	-1,06	NA	1,21	NA
Significância	*	**			***	**	***	***	***	**		*	
Consenso A, B e C	*****	*****	**NC**	**NC**	*******	*****	******	**NC**	*******	*****	**NC**	*****	**NC**

Modelo \ Preditor	**P18**	**P19**	**P20**	**P21**	**P22**	**P23**	**P24**	**P25**	**P27**	**P29**	**P32**	**P34**
A	0,47	0,75	3,17	NA	0,71	-1,29	-2,69	-1,02	-0,50	-3,58	-1,52	-0,62
Significância	***	‒	***		•	***	***	***	‒	**	***	‒
B	0,49	1,24	3,63	0,46	NA	-1,41	-3,33	-1,21	-0,57	-5,12	-1,66	NA
Significância	***	**	***	‒		***	***	***	‒	***	***	
C	0,73	1,46	3,22	0,78	NA	-1,35	-3,65	-1,04	NA	NA	-2,02	-1,88
Significância	***	***	***	*		***	***	***			***	***
Consenso A, B e C	*******	**NC**	*******	**NC**	**NC**	*******	*******	*******	**NC**	**NC**	*******	**NC**

Tabela 5.2. Modelos de RL, "Guy", após seleção *stepwise*, com as variáveis selecionadas com os respectivos coeficientes e significâncias quanto à existência de associação com a classe positiva NR5, indicada por um valor p < 0,05. As variáveis com p < 0,001 estão identificadas como "***"; 0,001 < p < 0,01, "**"; 0,01 < p < 0,05, "*"; 0,05 < p < 0,1, "•"; e 0,1 < p, "‒". Para os 03 modelos, as variáveis P7, P13, P22 e P27 não foram selecionadas pela *stepwise*. NA significa que a referida variável não foi selecionada para o modelo parcimonioso. No consenso, NC refere-se à inexistência de uma mesma variável nos três modelos com um valor p de, ao menos, 0,01 < p < 0,05, "*". Nas demais variáveis, a menor significância dentre os três modelos é destacada. Os valores de *deviance* residual e AIC de cada modelo estão respectivamente apresentados entre parênteses.

Modelo \ Preditor	**Intercept**	**P2**	**P4**	**P5**	**P6**	**P9**	**P10**	**P11**	**P12**	**P14**	**P15**	**P16**
A (561,1; 603,1)	1,25	-0,57	NA	-2,42	-2,57	-0,61	-1,31	1,98	5,14	1,92	10,06	-0,54
Significância	‒	•		*	*	*	***	***	‒	***	***	*
B (524,7; 562,7)	3,66	NA	0,76	-3,98	-2,65	-1,01	-1,43	1,73	NA	2,90	1,94	-0,65
Significância	•		•	**	*	**	***	***		***	*	*
C (559,3; 599,3)	2,76	-0,51	NA	-4,07	-1,71	-0,51	-1,64	1,74	NA	1,21	3,13	-0,59
Significância	‒	•		***	•	*	***	***		*	*	*
Consenso A, B e C	**NC**	**NC**	**NC**	*****	**NC**	*****	*******	*******	**NC**	*****	*****	*****

Modelo \ Preditor	**P17**	**P18**	**P19**	**P20**	**P21**	**P23**	**P24**	**P25**	**P26**	**P29**	**P32**	**P34**
A	20,74	-1,02	-0,86	1,19	NA	0,78	3,85	2,33	2,61	2,30	2,82	NA
Significância	*	***	*	***		*	***	***	•	*	***	
B	NA	-1,03	-1,24	1,85	NA	1,04	4,39	2,49	NA	4,00	3,98	1,50
Significância		***	**	***		**	***	***		**	***	**
C	NA	-1,48	-0,84	1,02	-0,31	0,77	4,99	2,30	2,01	2,79	3,34	NA
Significância		***	•	***	•	*	***	***	•	*	***	
Consenso A, B e C	**NC**	*******	**NC**	*******	**NC**	*****	*******	*******	**NC**	*****	*******	**NC**

Tabela 5.3. Modelos de RL, "KyteDoolittle", após seleção s*tepwise*, com as variáveis selecionadas com os respectivos coeficientes e significâncias quanto à existência de associação com a classe positiva NR5, indicada por um valor $p < 0,05$. As variáveis com $p < 0,001$ estão identificadas como "***"; $0,001 < p < 0,01$, "**"; $0,01 < p < 0,05$, "*"; $0,05 < p < 0,1$, "•"; e $0,1 < p$, "—". Para os 03 modelos, as variáveis P2, P5, P14, P21, P22, P26, P27 e P33 não foram selecionadas pela s*tepwise*. NA significa que a referida variável não foi selecionada para o modelo parcimonioso. No consenso, NC refere-se à inexistência de uma mesma variável nos três modelos com um valor p de ao menos $0,01 < p < 0,05$, "*". Nas demais variáveis, a menor significância dentre os modelos é destacada. Os valores de *deviance* residual e AIC de cada modelo estão respectivamente apresentados entre parênteses.

Modelo \ Preditor	**Intercept**	**P4**	**P6**	**P7**	**P9**	**P10**	**P11**	**P12**	**P13**	**P15**	**P16**
A (506,0; 542,0)	168,1	-0,71	4,25	46,4	0,79	0,30	-0,93	-0,90	0,84	-1,37	NA
Significância	—	•	***	—	***	•	***	—	***	*	
B (512,9; 584,9)	165,8	-0,47	0,86	49,5	0,83	NA	-0,93	NA	0,83	-0,79	0,28
Significância	—	**	*	—	***		***		***	**	*
C (534,5; 570,5)	5,49	-0,25	NA	NA	0,49	0,58	-0,59	-2,77	0,69	-0,57	0,50
Significância	—	•			***	***	***	***	***	**	***
Consenso A, B e C	**NC**	**NC**	**NC**	**NC**	*******	**NC**	*******	**NC**	*******	*****	**NC**

Modelo \ Preditor	**P17**	**P18**	**P19**	**P20**	**P23**	**P24**	**P25**	**P29**	**P32**	**P34**
A	-2,12	0,30	NA	1,12	-0,93	-1,66	-0,38	-1,18	-2,02	NA
Significância	—	**		***	***	***	***	*	***	
B	0,37	0,44	0,27	1,14	-0,99	-2,19	-0,46	-1,53	-1,09	NA
Significância	—	***	*	***	***	***	***	**	*	
C	NA	0,56	0,41	0,92	-0,60	-2,03	-0,34	-0,16	-2,14	-0,50
Significância		***	**	***	***	***	***	•	***	***
Consenso A, B e C	**NC**	******	**NC**	*******	*******	*******	*******	**NC**	*****	**NC**

Tabela 5.4. Matrizes de confusão dos modelos de RL, "Eisenberg", com os pontos de corte (*cut-off*) de 0,5 e ótimo.

Modelo	**A**						**B**						**C**					
Cut-off	**0,500**			**0,452**			**0,500**			**0,404**			**0,500**			**0,396**		
	Referência			**Referência**			**Referência**			**Referência**			**Referência**			**Referência**		
Predição	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**
NR5	134	48	**182**	146	57	**203**	137	68	**205**	151	96	**247**	140	64	**204**	149	83	**232**
R5	36	414	**450**	24	405	**429**	33	394	**427**	19	366	**385**	30	398	**428**	21	379	**400**
Total	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**

Tabela 5.5. Matrizes de confusão dos modelos de RL, "Guy", com os pontos de corte (*cut-off*) de 0,5 e ótimo.

Modelo	**A**						**B**						**C**					
Cut-off	**0,500**			**0,446**			**0,500**			**0,538**			**0,500**			**0,433**		
	Referência			**Referência**			**Referência**			**Referência**			**Referência**			**Referência**		
Predição	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**
NR5	145	65	**210**	150	66	**216**	144	62	**206**	143	45	**188**	140	74	**214**	156	87	**243**
R5	25	397	**422**	20	396	**416**	26	400	**426**	27	417	**444**	30	388	**418**	14	375	**389**
Total	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**

Tabela 5.6. Matrizes de confusão dos modelos de RL, "KyteDoolittle", com os pontos de corte (*cut-off*) de 0,5 e ótimo.

Modelo	**A**						**B**						**C**					
Cut-off	**0,500**			**0,629**			**0,500**			**0,592**			**0,500**			**0,724**		
	Referência			**Referência**			**Referência**			**Referência**			**Referência**			**Referência**		
Predição	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**
NR5	145	80	**225**	134	30	**164**	143	82	**229**	140	47	**187**	131	65	**196**	124	21	**145**
R5	25	382	**407**	36	432	**468**	27	380	**403**	30	415	**445**	39	397	**436**	46	441	**487**
Total	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**

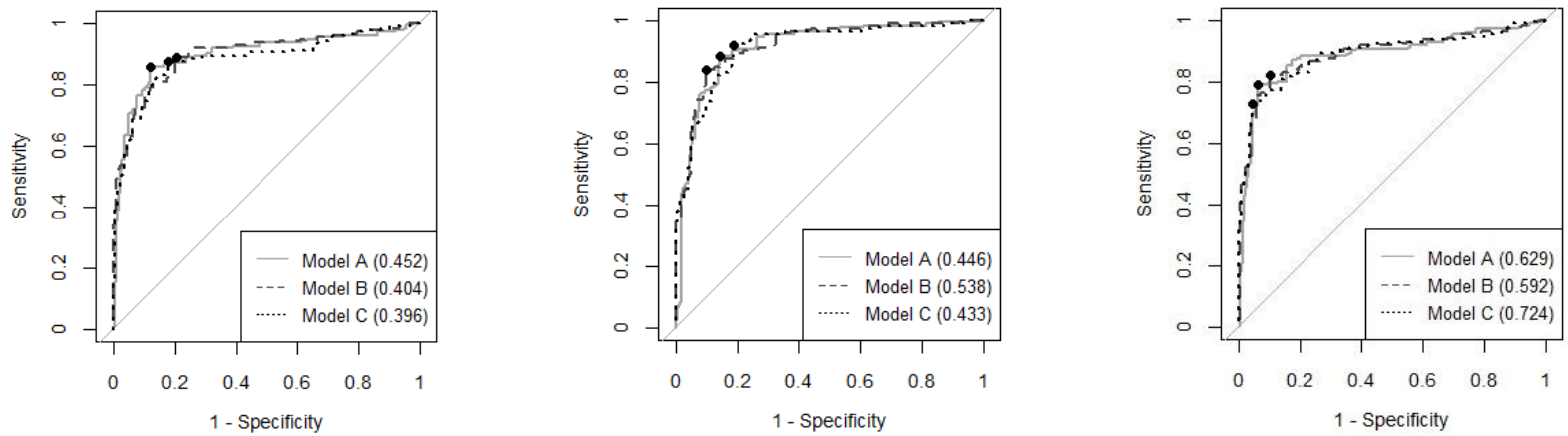

Figura 5.1. Curvas ROC dos modelos de RL "Eisenberg", "Guy" e "KyteDoolittle", respectivamente. Os pontos de corte ótimos estão entre parênteses e indicados nas curvas por círculos.

Tabela 5.7. Desempenhos preditivos dos modelos de RL, parcimoniosos, sob os pontos de corte (*cut-off*) ótimos.

Modelo	**A**			**B**			**C**		
Escala	**Eisenberg**	**Guy**	**KyteDoolittle**	**Eisenberg**	**Guy**	**KyteDoolittle**	**Eisenberg**	**Guy**	**KyteDoolittle**
Medida \ *Cut-off*	**(0,452)**	**(0,446)**	**(0,629)**	**(0,404)**	**(0,538)**	**(0,592)**	**(0,396)**	**(0,433)**	**(0,724)**
Acurácia (%)	87,2	86,4	89,6	81,8	88,6	87,8	83,5	84,0	89,4
[IC95%]	[84,3-89,7]	[83,5-89,0]	[86,9-91,8]	[78,6-84,7]	[85,9-91,0]	[85,0-90,3]	[80,4-86,4]	[80,9-86,8]	[86,7-91,7]
Sensibilidade (%)	85,9	88,2	78,8	88,8	84,1	82,4	87,7	91,8	72,9
[IC95%]	[80,6-91,2]	[83,5-92,9]	[72,9-84,7]	[84,1-93,5]	[78,2-89,4]	[77,1-87,7]	[82,9-92,4]	[87,7-95,3]	[66,5-79,4]
Especificidad (%)	87,7	85,7	93,5	79,2	90,3	89,8	82,0	81,2	95,5
[IC95%]	[84,4-90,7]	[82,5-89,0]	[91,1-95,7]	[75,1-82,9]	[87,7-92,9]	[87,0-92,6]	[78,4-85,5]	[77,7-84,6]	[93,3-97,2]
AUC	0,894	0,913	0,885	0,896	0,918	0,890	0,882	0,912	0,888
[IC95%]	[0,860-0,928]	[0,887-0,939]	[0,850-0,920]	[0,864-0,928]	[0,893-0,943]	[0,855-0,925]	[0,847-0,917]	[0,886-0,939]	[0,853-0,923]

5.2. *Naive* Bayes

As covariáveis usadas no treinamento dos modelos no método de NB foram as mesmas utilizadas para o treinamento dos modelos "98" correspondentes no método de RL, conforme a Tabela 4.4 e a configuração ilustrada na Figura 4.4.

As probabilidades *a posteriori* obtidas após os ajustes dos modelos foram geralmente bem próximas de um, quando a sequência V3 teste foi classificada como NR5, ou geralmente bem próximas de zero, quando a V3 foi classificada como R5. Em função destes resultados e o maior número de sequências R5 no conjunto teste, os pontos de corte ótimos gerados foram bem próximos de zero – com exceção de um modelo "KyteDoolittle", subconjunto de treinamento C. Um exemplo de ponto ótimo de corte obtido foi o do modelo "Eisenberg", treinado com o subconjunto de treinamento A, de valor $3{,}25197e^{-05}$.

De forma a evitar estes valores de uso pouco prático, as matrizes de confusão foram construídas considerando apenas o ponto de corte de 0,5. As matrizes de confusão dos modelos de NB estão apresentadas nas Tabelas 5.8 a 5.10. As curvas ROC dos modelos estão ilustradas na Figura 5.2. As medidas de desempenho de acurácia, sensibilidade, especificidade e AUC, além dos seus respectivos IC95%, estão apresentadas na Tabela 5.11.

Tabela 5.8. Matrizes de confusão dos modelos de NB, "Eisenberg", com os pontos de corte de 0,5.

Modelo	A			B			C		
	Referência			Referência			Referência		
Predição	NR5	R5	Total	NR5	R5	Total	NR5	R5	Total
NR5	101	17	**118**	111	23	**134**	113	27	**140**
R5	69	445	**514**	59	439	**498**	57	435	**492**
Total	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**

Tabela 5.9. Matrizes de confusão dos modelos de NB, "Guy", com os pontos de corte de 0,5.

Modelo	A			B			C		
	Referência			Referência			Referência		
Predição	NR5	R5	Total	NR5	R5	Total	NR5	R5	Total
NR5	104	22	**126**	105	21	**126**	106	19	**125**
R5	66	440	**506**	65	441	**506**	64	443	**507**
Total	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**

Tabela 5.10. Matrizes de confusão dos modelos de NB, "KyteDoolittle", com os pontos de corte de 0,5.

Modelo	A			B			C		
	Referência			Referência			Referência		
Predição	NR5	R5	Total	NR5	R5	Total	NR5	R5	Total
NR5	90	7	**97**	99	25	**124**	116	26	**142**
R5	80	455	**435**	71	437	**508**	54	436	**490**
Total	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**

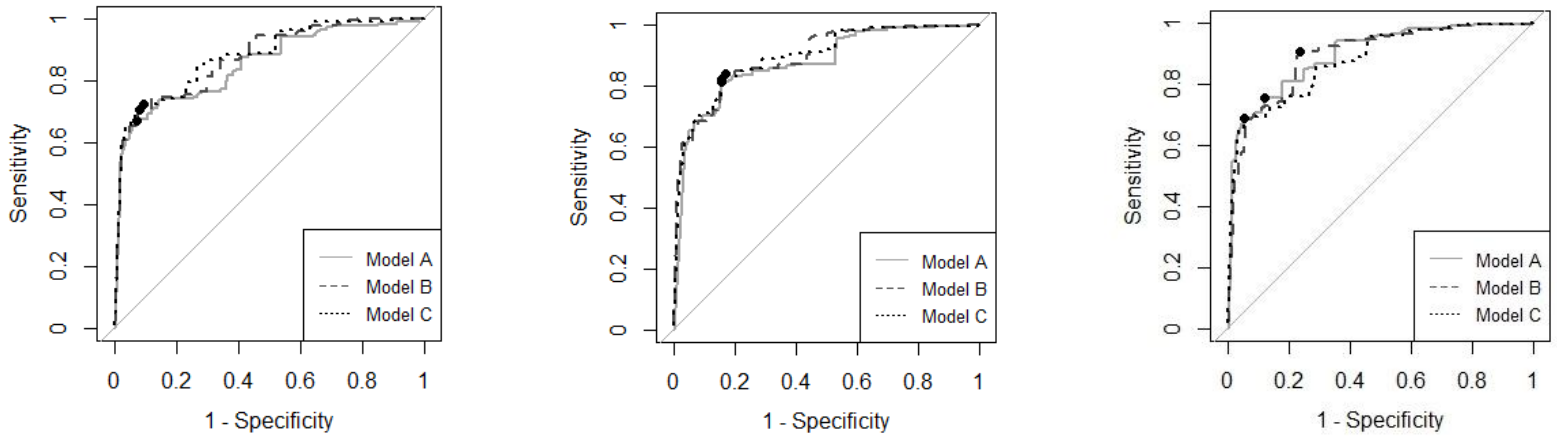

Figura 5.2. Curvas ROC dos modelos de NB "Eisenberg", "Guy" e "KyteDoolittle", respectivamente. Os pontos de corte de 0,5 estão indicados nas curvas por círculos.

Tabela 5.11. Desempenho preditivo dos modelos de NB, sob os pontos de corte (*cut-off*) de 0,5.

Modelo	**A**			**B**			**C**		
Escala	**Eisenberg**	**Guy**	**KyteDoolittle**	**Eisenberg**	**Guy**	**KyteDoolittle**	**Eisenberg**	**Guy**	**KyteDoolittle**
Medida									
Acurácia (%)	86,4	86,1	86,2	87,0	86,4	84,8	86,7	86,9	87,3
[IC95%]	[83,5-89,0]	[83,1-88,7]	[83,3-88,8]	[84,2-89,6]	[83,5-89,0]	[81,8-87,5]	[83,8-89,3]	[84,0-89,4]	[84,5-89,8]
Sensibilidade (%)	59,4	61,2	52,9	65,3	61,8	58,2	66,5	62,4	68,2
[IC95%]	[51,8-66,5]	[53,5-68,2]	[45,3-60,6]	[57,7-71,8]	[54,1-68,8]	[50,6-65,3]	[59,4-73,0]	[54,7-69,4]	[61,2-75,3]
Especificidad (%)	96,3	95,2	98,5	95,0	95,5	94,6	94,2	95,9	94,4
[IC95%]	[94,4-98,1]	[93,3-97,2]	[97,2-99,6]	[93,1-97,0]	[93,5-97,2]	[92,4-96,5]	[92,0-96,1]	[94,2-97,6]	[92,2-96,3]
AUC	0,851	0,874	0,895	0,874	0,894	0,893	0,878	0,892	0,878
[IC95%]	[0,814-0,888]	[0,842-0,907]	[0,866-0,923]	[0,843-0,906]	[0,865-0,922]	[0,864-0,921]	[0,846-0,910]	[0,863-0,922]	[0,846-0,909]

5.3. *Random Forest*

As covariáveis usadas no treinamento dos modelos no método de RF foram as mesmas utilizadas para o treinamento dos modelos "98" correspondentes no método de RL, conforme a Tabela 4.4 e a configuração ilustrada na Figura 4.4.

As matrizes de confusão foram construídas considerando os pontos de corte de 0,5 e ótimos (Tabelas 5.12 a 5.14). As curvas ROC dos modelos de RF estão ilustradas na Figura 5.3. As medidas de desempenho de acurácia, sensibilidade, especificidade e AUC, além dos seus respectivos IC95%, estão apresentadas nas Tabelas 5.14. Para acurácia, sensibilidade e especificidade, as estimativas pontuais e intervalares foram realizadas usando o ponto de corte ótimo.

5.4. Análise dos IC95% de AUC

A relação dos subconjuntos de treinamento, das escalas de hidrofobicidade e dos métodos de classificação no desempenho dos modelos foi verificada a partir da análise dos IC95% das medidas de AUC. Neste sentido, não houve diferença estatisticamente significativa, dentro de cada método, entre os subconjuntos de treinamento e entre as escalas de hidrofobicidade (Figuras 5.4 a 5.6).

Em relação aos métodos, enquanto RL e NB não se diferenciaram quanto ao desempenho preditivo, RF foi o método que gerou os classificadores com a melhor capacidade discriminativa, de maneira significativa em todos os casos, além de estimativas de AUC mais precisas. Para ilustrar este resultado, a Figura 5.7 apresenta os IC95% da medida AUC dos modelos "Guy", treinados com o conjunto A, em cada método.

Tabela 5.12. Matrizes de confusão dos modelos de RF, "Eisenberg", com os pontos de corte (*cut-off*) de 0,5 e ótimo.

Modelo	**A**						**B**						**C**					
Cut-off	**0,500**			**0,448**			**0,500**			**0,633**			**0,500**			**0,399**		
	Referência			**Referência**			**Referência**			**Referência**			**Referência**			**Referência**		
Predição	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**
NR5	159	04	**163**	161	06	**167**	160	07	**167**	159	02	**161**	159	03	**162**	163	09	**172**
R5	11	458	**469**	09	456	**465**	10	455	**465**	11	460	**471**	11	459	**470**	07	453	**460**
Total	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**

Tabela 5.13. Matrizes de confusão dos modelos de RF, "Guy", com os pontos de corte (*cut-off*) de 0,5 e ótimo.

Modelo	**A**						**B**						**C**					
Cut-off	**0,500**			**0,345**			**0,500**			**0,502**			**0,500**			**0,452**		
	Referência			**Referência**			**Referência**			**Referência**			**Referência**			**Referência**		
Predição	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**
NR5	159	04	**163**	163	11	**174**	162	10	**172**	162	10	**172**	159	06	**165**	163	10	**173**
R5	11	458	**469**	07	451	**458**	08	452	**460**	08	452	**460**	11	456	**467**	07	452	**459**
Total	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**

Tabela 5.14. Matrizes de confusão dos modelos de RF, "KyteDoolittle", com os pontos de corte (*cut-off*) de 0,5 e ótimo.

Modelo	**A**						**B**						**C**					
Cut-off	**0,500**			**0,437**			**0,500**			**0,543**			**0,500**			**0,491**		
	Referência			**Referência**			**Referência**			**Referência**			**Referência**			**Referência**		
Predição	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**	**NR5**	**R5**	**Total**
NR5	160	07	**167**	161	07	**168**	161	09	**170**	161	06	**167**	161	5	**166**	162	5	**167**
R5	10	455	**465**	09	455	**464**	09	453	**462**	09	456	**465**	09	457	**466**	08	457	**465**
Total	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**	**170**	**462**	**632**

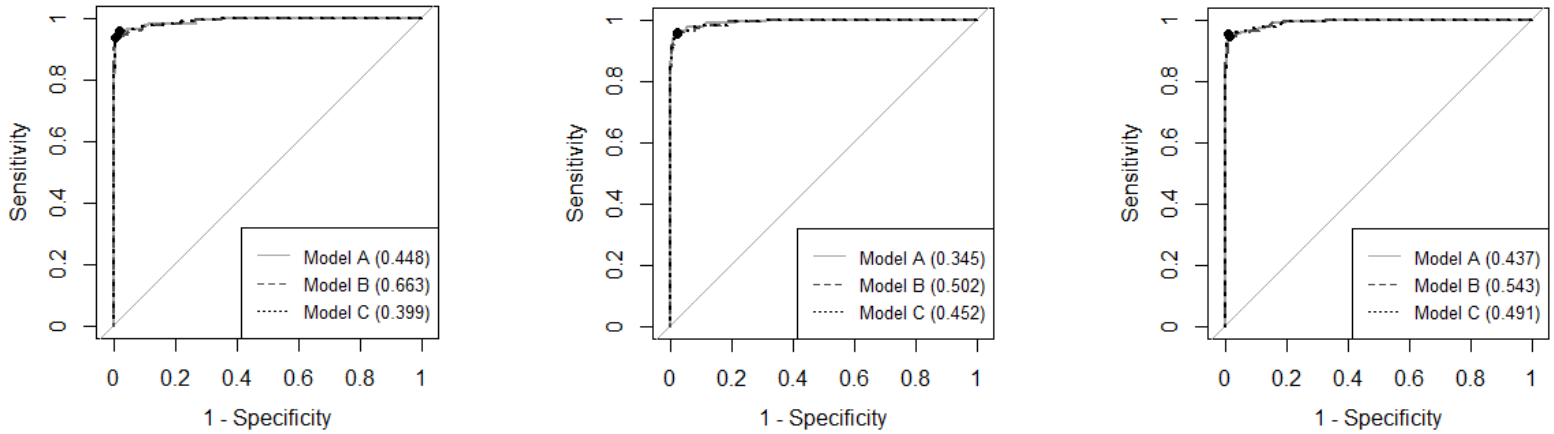

Figura 5.3. Curvas ROC dos modelos de RF "Eisenberg", "Guy" e "KyteDoolittle", respectivamente. Os pontos de corte ótimos estão entre parênteses e indicados nas curvas por círculos.

Tabela 5.15. Desempenho preditivo dos modelos de RF, sob os pontos de corte (*cut-off*) ótimos.

Modelo	**A**			**B**			**C**		
Escala	**Eisenberg**	**Guy**	**KyteDoolittle**	**Eisenberg**	**Guy**	**KyteDoolittle**	**Eisenberg**	**Guy**	**KyteDoolittle**
Medida \ *Cut-off*	**(0,448)**	**(0,345)**	**(0,437)**	**(0,633)**	**(0,502)**	**(0,543)**	**(0,399)**	**(0,452)**	**(0,491)**
Acurácia (%)	97,6	97,2	97,5	98,0	97,2	97,6	97,5	97,3	97,9
[IC95%]	[96,1-98,7]	[95,5-98,3]	[95,9-98,6]	[96,5-98,9]	[95,5-98,3]	[96,1-98-7]	[95,9-98,6]	[95,7-98,4]	[96,5-98,9]
Sensibilidade (%)	94,7	95,9	94,7	93,5	95,3	94,7	95,9	95,9	95,3
[IC95%]	[91,2-97,7]	[92,9-98,8]	[91,2-97,7]	[90,0-97,1]	[92,4-98,2]	[91,2-97,7]	[92,9-98,2]	[92,9-98,8]	[91,8-98,2]
Especificid. (%)	98,7	97,6	98,5	99,6	97,8	98,7	98,1	97,8	98,9
[IC95%]	[97,6-99,6]	[96,3-98,9]	[97,2-99,6]	[98,9-100,0]	[96,3-99,1]	[97,6-99,6]	[96,5-99,1]	[96,3-99,1]	[97,8-99,8]
AUC	0,992	0,994	0,992	0,991	0,993	0,992	0,992	0,993	0,992
[IC95%]	[0,985-0,998]	[0,989-0,999]	[0,987-0,998]	[0,985-0,998]	[0,988-0,998]	[0,987-0,998]	[0,986-0,998]	[0,988-0,998]	[0,987-0,998]

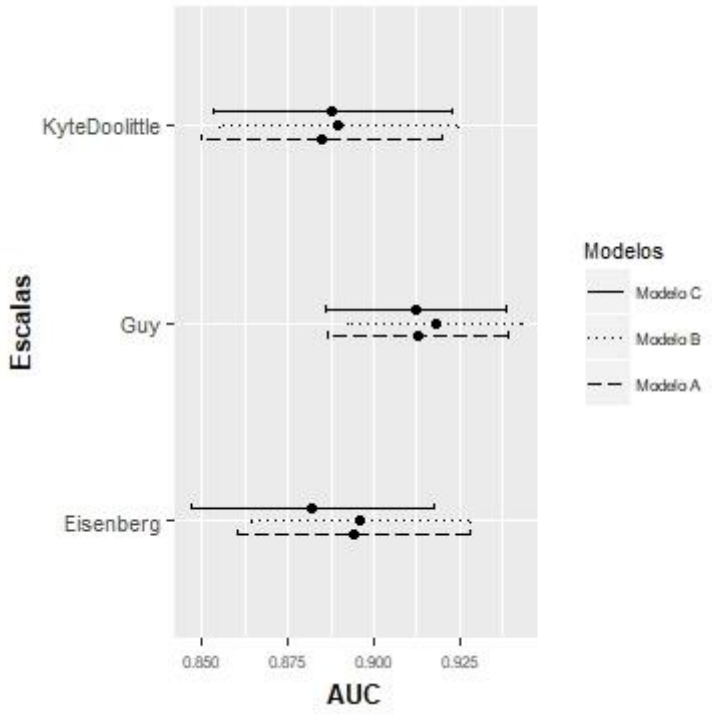

Figura 5.4. Comparação entre os ICs95% da medida AUC dos modelos de RL.

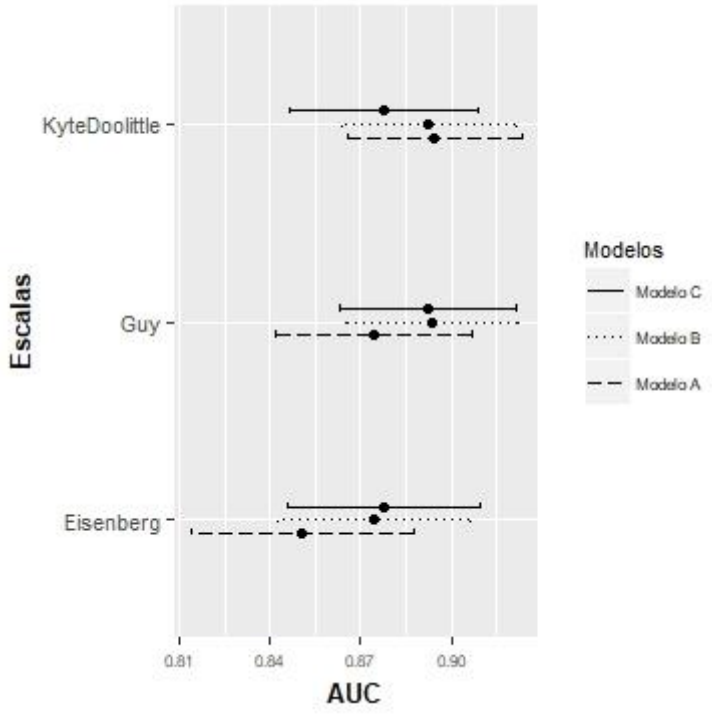

Figura 5.5. Comparação entre os ICs95% da medida AUC dos modelos de NB.

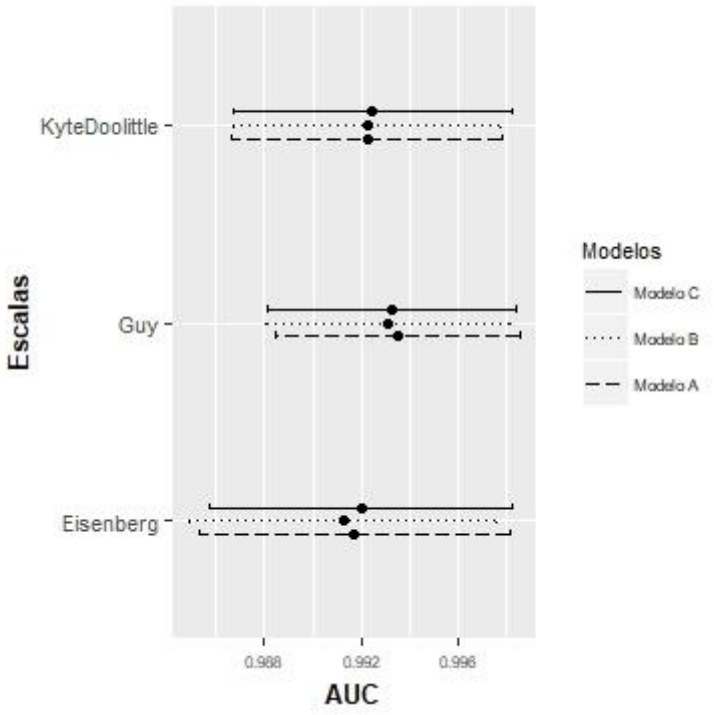

Figura 5.6. Comparação entre os ICs95% da medida AUC dos modelos de RF.

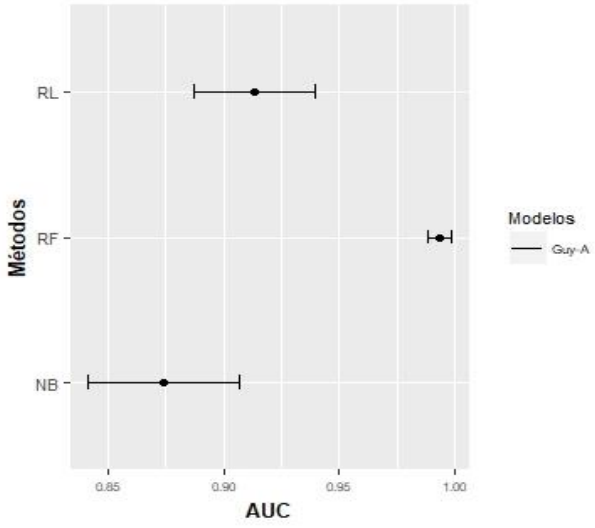

Figura 5.7. Comparação entre os ICs95% da medida AUC dos modelos treinados com o subconjunto A "Guy" em cada método (RL, RF e NB).

Capítulo 06

Discussão

Os ensaios Trofile™ e ES-Trofile™ se estabeleceram como referências fenotípicas no diagnóstico do tropismo do HIV [10–12]. Porém, as suas limitações técnicas e logísticas levaram ao desenvolvimento de algoritmos de genotropismo, que são ferramentas mais práticas e com uma relação de custo-benefício mais favorável para o uso na rotina clínica [11–12].

O desempenho do g2p repercutiu inicial e positivamente após a verificação de uma concordância preditiva de 86,5% com o Trofile™, na análise de amostras clínicas de HIV-1 [98]. Em coerência com o fato do algoritmo ser treinado sobretudo por sequências V3 de HIV-1 do subtipo B, análises subsequentes demonstraram que o bom desempenho preditivo do g2p em vírus de subtipo B não era reproduzido em vírus de subtipos não-B analisados em conjunto [83]. Contudo, análise recente mais detalhada demonstrou uma boa acurácia do g2p na predição específica de sequências V3 de HIV-1 do subtipo C [91]. Em virtude dos subtipos B e C representarem cerca de 60% das infecções por HIV-1 ao redor do mundo [21], o estudo corroborou a importância do g2p como ferramenta de genotropismo. Nesta perspectiva, o subtipo B é o mais prevalente no Brasil, distribuindo-se em todas as suas regiões. Dentre os dois subtipos não-B que prevalecem no país, embora bem menos frequentes que o B, além do subtipo F, inclui-se o subtipo C [24].

Na dissertação, foram incluídas 2.109 sequências V3 de HIV-1 do subtipo B. Dentre as 2.109 sequências V3, 1.110 são distintas entre si. Trata-se de um número praticamente idêntico ao de 1.100 sequências V3 de HIV-1 utilizadas no procedimento de validação cruzada do g2p [11–12, 15]. Para a modelagem, os subconjuntos de treinamento foram balanceados quanto ao número de sequências NR5 e R5. O intuito foi evitar a geração de modelos com melhor desempenho a favor de uma classe mais numerosa, o que ocorreria com dados desbalanceados. Neste sentido, dados balanceados minimizam a perda de robustez e a capacidade de generalização dos modelos, características consideradas fundamentais em análises preditivas efetivas [99].

O presente estudo promove uma comparação, aparentemente não disponível na literatura científica indexada, dos métodos de RL, NB e RF, além do uso de três escalas

consensuais de hidrofobicidade como preditores numéricos das sequências peptídicas de V3, na modelagem e análise do tropismo do HIV-1. Não houve diferenças significativas entre os resultados de AUC dos modelos quanto aos preditores numéricos das três escalas utilizadas, em cada um dos métodos e respectivos subconjuntos de treinamento (Figuras 5.4 a 5.6).

Os modelos de RL, NB e, notadamente, RF apresentaram uma boa concordância com os resultados do g2p, sendo que que RL e NB não apresentaram diferença significativa entre os seus desempenhos preditivos (Figuras 5.4, 5.5 e 5.7). Enquanto que estudos na literatura indexada sobre o uso do método de RL na predição do tropismo do HIV-1 não foram encontrados, Díez-Fuertes *et al.* (2013) desenvolveram uma ferramenta de genotropismo baseada em um classificador Bayesiano, utilizando para isto 26 posições nucleotídicas ao longo de todo o gene *env*. Apesar do bom desempenho preditivo do classificador [95], deve-se ressaltar que o uso, como variáveis preditoras, de 26 posições nucleotídicas ao longo de todo o gene da gp120 traz os problemas técnicos e de custos associados ao sequenciamento de sequências de cDNA de maior extensão, em comparação com a análise baseada na região V3.

Apesar dos métodos de RL e NB não terem apresentado diferenças significativas entre as suas predições, o método de RL permite a análise inferencial, onde relações específicas entre as variáveis explanatórias e a variável resposta podem ser estabelecidas [69]. A partir desta análise, de forma consensual entre os modelos de RL "Eisenberg", verificou-se que treze variáveis explanatórias demonstraram uma associação significativa ($p < 0,05$) com o desfecho NR5, sendo que oito variáveis apresentaram um valor $p < 0,001$. Dentre estas variáveis, quatro (P9, P13, P18 e P20) apresentaram uma associação positiva, enquanto que quatro (P23, P24, P25, P32) apresentaram uma associação negativa com o desfecho NR5 (Tabela 5.1). Entre os modelos "Guy", quatorze variáveis demonstraram uma associação significativa ($p < 0,05$) com o desfecho NR5, sendo que seis variáveis apresentaram um valor $p < 0,001$. Dentre estas variáveis, cinco (P11, P18, P20, P24 e P25) apresentaram uma associação positiva, enquanto que uma (P10) apresentou uma associação negativa com o desfecho NR5 (Tabela 5.2). Entre os modelos "KyteDoolittle", dez variáveis demonstraram uma associação significativa ($p < 0,05$) com o desfecho NR5, sendo que sete variáveis apresentaram um valor $p < 0,001$. Dentre estas variáveis, duas (P9, P13 e P20) apresentaram uma associação positiva, enquanto que quatro (P11, P23, P24 e P25) apresentaram uma associação negativa com o desfecho NR5 (Tabela 5.3). De forma

consensual dentre todos os modelos de RL, oito variáveis (P9, P11, P18, P20, P23, P24, P25 e P32) apresentaram associação positiva ou negativa significativa ($p < 0,05$) com o desfecho NR5. Dentre estas oito variáveis, três (P11, P24 e 25) são a base do modo operacional de predição das regras 11/25 e 11/24/25 [78–81], consideradas as primeiras ferramentas de genotropismo do HIV [11–12].

Esta informação fornecida pelo método de RL é mais um exemplo da existência de uma variabilidade na região V3 mais complexa que aquela preconizada pelas regras pioneiras, e que se reflete na determinação do genotropismo [12, 68, 97]. Como consequência, tais regras foram sendo substituídas por algoritmos mais sofisticados como o g2p e o Web-PSSM [11–12]. A análise inferencial pelo método de RF deve confirmar a existência de associação significativa com o desfecho NR5 de, ao menos, parte das variáveis preditoras indicadas pelo método de RL.

Enquanto os métodos de RL e NB não se diferenciaram quanto à predição, o RF apresentou um desempenho preditivo significativamente superior aos dois métodos, gerando assim os melhores classificadores, além de estimativas de AUC mais precisas (Figuras 5.4 a 5.7). Tal desempenho é mais uma amostra do potencial do método de RF no genotropismo do HIV-1. Neste sentido, Xu *et al.* (2007) haviam demonstrado um melhor desempenho do método de RF no genotropismo frente ao Web-PSSM [96]. Por sua vez, Heider *et al.* (2014) atingiram novos patamares preditivos com o uso combinado de informações estruturais e sequencias da V3, a partir de classificadores de RF gerados em dois níveis de aprendizado [68]. Recentemente, Lochel *et al.* (2018), utilizando uma abordagem de modelagem semelhante àquela de Heider *et al.* (2014), desenvolveram um algoritmo de RF com boa acurácia no genotropismo de HIV-1 do subtipo A e suas CRF, superando assim as restrições do g2p e Web-PSSM na análise deste subtipo viral [97]. Desta forma, o modo operacional do método de RF, baseado na combinação de vários classificadores do tipo árvore de decisão [73], apresenta-se como uma ferramenta preditiva muito robusta, capaz de aumentar a confiança na tomada de decisão em relação ao tratamento de resgate com o maraviroque, a partir da análise de sequências V3 oriundas de HIV-1 de subtipo B e não-B [68, 97].

No presente estudo, durante cada treinamento com o método de RF, 500 árvores classificadores foram geradas, com cada árvore sendo composta por cinco variáveis explicativas selecionadas de forma aleatória. Tal aleatoriedade associada ao modo operacional do algoritmo proporciona a diversidade que leva à obtenção final dos melhores classificadores possíveis [73]. Na perspectiva do genotropismo, é possível

especular que a geração de classificadores com várias combinações de variáveis da V3 aumenta a chance do algoritmo de RF em relação a outros métodos de interpretar mais adequadamente a complexidade da variabilidade genética intrínseca à região V3, refletindo-se assim no seu desempenho na determinação do tropismo em HIV-1.

Diante dos resultados encorajadores, as próximas etapas de trabalho incluem a curto prazo o treinamento e teste do método de RF com um conjunto de sequências V3 de HIV-1 do subtipo B classificadas quanto ao tropismo por caracterizações fenotípicas referenciais, de forma que um comparativo direto de desempenho possa ser realizado com algoritmos como o g2p e o Web-PSSM. Nesta etapa, sequências de HIV-1 do subtipo F podem ser incluídas, de forma a contemplar a análise preditiva do método de RF naqueles subtipos de maior ocorrência no Brasil. A depender destes resultados, pode-se conjeturar a médio prazo o desenvolvimento de um algoritmo baseado no método de RF, capaz de classificar sequências V3 clínicas de HIV-1, oriundas de NGS, de forma prática, segura e eficaz, cujo intuito seja assessorar a tomada de decisão clínica quanto ao uso de antagonistas de correceptores em tratamentos de resgate em indivíduos infectados no Brasil pelos subtipos de HIV-1 que predominam no país.

Capítulo 07

Conclusão

Enquanto que as escalas de hidrofobicidade de Eisenberg, Guy e KyteDoolittle não apresentaram diferenças significativas entre si como preditores numéricos para as modelagens das sequências peptídicas de V3, na comparação entre os métodos de RL, NB e RF na análise do tropismo do HIV-1 de subtipo B, os modelos classificadores de RF apresentaram os melhores resultados preditivos, sob a forma de resultados significativos na principal medida de desempenho de AUC, além de apresentar IC95% sempre mais precisos.

Os resultados são encorajadores para a continuação do projeto de desenvolvimento de um classificador baseado no método de RF, e validado com informação fenotípica do tropismo, que seja prático, seguro e eficaz, com potencial para assessorar a tomada de decisão clínica do uso de antagonistas de correceptores em pacientes infectados no Brasil com os subtipos de HIV-1 mais prevalentes no país.

Referências Bibliográficas

[1] MERSON, M., O'MALLEY, J., SERWADDA, D., et al. "The History and Challenge of HIV Prevention", **Lancet** v. 372, n. 9637, pp. 475–488, Ago. 2008.

[2] FAUCI, A. "HIV and AIDS: 20 Years of Science", **Nat Med** v. 9, n. 7, pp. 839–843, Jul. 2003.

[3] CLAVEL, F., HANCE, A. "HIV Drug Resistance", **N Engl J Med** v. 350, n. 10, pp. 1023–1035, Mar. 2004.

[4] MAEDA, K., NAKATA, H., OGATA, H., et al. "The Current Status of, and Challenges in, the Development of CCR5 Inhibitors as Therapeutics for HIV-1 Infection", **Curr Opin Pharmacol** v. 4, n. 5, pp. 447–452, Out. 2004.

[5] DENG, H., LIU, R., ELLMEIER, W., et al. "Identification of a Major correceptor for Primary Isolates of HIV-1", **Nature** v. 381, n. 6584, pp. 661–666, Jun. 1996.

[6] DRAGIC, T., LITWIN, V., ALLAWAY, G., et al. "HIV-1 Entry into $CD4^+$ Cells is Mediated by the Chemokine Receptor CC-CKR-5", **Nature** v. 381, n. 6584, pp. 667–673, Jun. 1996.

[7] FENG, Y., BRODER, C., KENNEDY, P., et al. "HIV-1 Entry Cofactor: Functional cDNA Cloning of a Seven-Transmembrane, G Protein-Coupled Receptor", **Science** v. 272, n. 5263, pp. 872–877, Mai. 1996.

[8] SEIBERT, C., SAKMAR, T. "Small-Molecule Antagonists of CCR5 and CXCR4: a Promising New Class of Anti-HIV Drugs", **Curr Pharm Des** v. 10, n. 17, pp. 2041–2062, 2004.

[9] ANÔNIMO. "FDA Approves Maraviroc Tablets", **AIDS Patient Care STDS** v. 21, n. 9, pp. 702, Set. 2007.

[10] POVEDA, E., BRIZ, V., QUIÑONES-MATEU, M., et al. “HIV Tropism: Diagnostic Tools and Implications for Disease Progression and Treatment with Entry Inhibitors”, **AIDS** v. 20, n. 10, pp. 1359–1367, Jun. 2006.

[11] AIAMKITSUMRIT, B., DAMPIER, W., ANTELL, G., et al. “Bioinformatic Analysis of HIV Entry and Pathogenesis”, **Curr HIV Res** v. 12, n. 2, pp. 132–161, 2014.

[12] 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 Rev** v. 12, n. 3, pp. 135–148, Jul. 2010.

[13] VERHOFSTEDE, C., BRUDNEY, D., REYNAERTS J, et al. “Concordance between HIV-1 Genotypic Coreceptor Tropism Predictions Based on Plasma RNA and Proviral DNA”, **HIV Med** v. 12, n. 9, pp. 544–552, Out. 2011.

[14] McGOVERN, R., THIELEN, A., MO, T., et al. “Population-based V3 Genotypic Tropism Assay: a Retrospective Analysis Using Screening Samples from the A4001029 and MOTIVATE Studies”, **AIDS** v. 24, n. 16, pp. 2517–2525, Out. 2010.

[15] LENGAUER, T., SANDER, O., SIERRA, S., et al. “Bioinformatics Prediction of HIV Coreceptor Usage”, **Nat Biotechnol** v. 25, n. 12, pp. 1407–1410, Dez. 2007.

[16] STOVER, J., BOLLINGER, L., IZAZOLA, J., et al. “What is required to End the AIDS Epidemic as a Public Health Threat by 2030? The Cost and Impact of the Fast-Track Approach”, **Plos One** v. 11, n. 5, e0154893, Mai. 2016.

[17] CENTER FOR DISEASE CONTROL. “Kaposi’s Sarcoma and *Pneumocystis* Pneumonia Among Homosexual Men – New York City and California”, **MMWR Morb Mortal Wkly Rep** v. 30, n. 25, pp. 305–308, Jul. 1981.

[18] SHARP, P., HAHN, B. “Origins of HIV and the AIDS Pandemic”, **Cold Spring Harb Perspect Med** v. 1, n. 1, pp. a006841, Set. 2011.

[19] TORTORA, G., FUNKE, B., CASE, C., **Microbiologia**. 12 ed. Porto Alegre, Artmed, 2016.

[20] GALLO, R., MONTAGNIER, L. “The Discovery of HIV as the Cause of AIDS”, **N Engl J Med** v. 349, n. 24, pp. 2283–2285, Dez. 2003.

[21] PEETERS, M. “The Genetic Variability of HIV-1 and its Implications”, **Transfus Clin Biol** v. 8, n. 3, pp. 222–225, Jun. 2001.

[22] VIDAL, N, PEETERS, M., MULANGA-KABEYA, C., et al. “Unprecedented Degree of Human Immunodeficiency Virus Type 1 (HIV-1) Group M Genetic Diversity in the Democratic Republic of Congo Suggests that the HIV-1 Pandemic Originated in Central Africa”, **J Virol** v. 74, n. 22, pp. 10498–10507, Nov. 2000.

[23] GILBERT, M., RAMBAUT, A., WLASIUK, G., et al. “The Emergence of HIV/AIDS in the Americas and Beyond”, **Proc Natl Acad Sci USA** v. 104, n. 47, pp. 18566–18570, Nov. 2007.

[24] BONGERTZ, V., BOU-HABIB, D., BRÍGIDO, L., et al. “HIV-1 Diversity in Brazil: Genetic, Biologic, and Immunologic Characterization of HIV-1 Strains in Three Potential HIV Vaccine Evaluation Sites. Brazilian Network for HIV Isolation and Characterization", **J Acquir Immune Defic Syndr** v. 23, n. 2, pp. 184–193, Fev. 2000.

[25] CACCURI, F., MARSICO, S., FIORENTINI, S., et al. “HIV-1 Matrix Protein p17 and its Receptors”, **Curr Drug Targets** v. 17, n. 1, pp. 23–32, 2016.

[26] ARRILDT, K., JOSEPH, S., SWANSTROM, R. “The HIV-1 Env Protein: a Coat of Many Colors”, **Curr HIV/AIDS Rep** v. 9, n. 1, pp. 52–63, Mar. 2012.

[27] SIROIS, S., SING, T., CHOU, K. “HIV-1 gp120 V3 Loop for Structure-Based Drug Design”, **Curr Protein Pept Sci** v. 6, n. 5, pp. 413–422, Out. 2005.

[28] SHARON, M., KESSLER, N., LEVY, R., et al. Alternative Conformations of HIV-1 V3 Loops Mimic Beta Hairpins in Chemokines, Suggesting a Mechanism for Coreceptor Selectivity”, **Structure** v. 11, n. 2, pp. 225-236, Fev. 2003.

[29] CORMIER, E., DRAGIC, T. “The Crown and Stem of the V3 Loop Play Distinct Roles in HIV Type 1 Envelope Glycoprotein Interactions with the CCR5 Coreceptor”, **J Virol** v. 76, n. 17; pp. 8953–8957, Set. 2002.

[30] TERSMETTE, M., GOEDE, R., AL, B., et al. “Differential Syncytium-Inducing Capacity of HIV Isolates: Frequent Detection of Syncytium-Inducing Isolates in Patients with AIDS and AIDS-Related Complex”, **J Virol** v. 62, n. 6, pp. 2026-2032, Jun. 1988.

[31] FENYÖ, E., MORFELDT-MANSON, L., CHIODI, F., et al. “Distinct Replicative and Cytopathic Characteristics of HIV Isolates”, **J Virol** v. 62, n. 11, pp. 4414–4419, Nov. 1988.

[32] COLLMAN, R., HASSAN, N., WALKER, R., et al. “Infection of Monocyte-Derived Macrophages with HIV-1”, **J Exp Med** v. 170, n. 4, pp. 1149–1163, Out. 1989.

[33] SCHUITEMAKER, H., KOOT, M., KOOTSTRA, N., et al. “Biological Phenotype of HIV Type 1 Clones at Different Stages of Infection: Progression of Disease is Associated with a Shift from Monocytotropic to T-Cell-Tropic Virus Populations”, **J Virol** v. 66, n. 3, pp. 1354–1360, Mar. 1992.

[34] CONNOR, R., SHERIDAN, K., CERADINI, D., et al. “Change in Coreceptor Use Coreceptor Use Correlates with Disease Progression in HIV-1-Infected Individuals”, **J Exp Med** v. 185, n. 4, pp. 621–628, Fev. 1997.

[35] BJÖRNDAL, A., DENG, H., JANSSON, M., et al. “Coreceptor Usage of Primary HIV Type 1 Isolates Varies According to Biological Phenotype”, **J Virol** v. 71, n. 10, pp. 7478–7487, Out. 1997.

[36] BERGER, E., DOMS, R., FENYÖ, E., et al. “A New Classification for HIV-1”, **Nature** v. 391, n. 6664, pp. 240, Jan. 1998.

[37] MURPHY, P., BAGGIOLINI, M., CHARO, I., et al. “International Union of Pharmacology. XXII. Nomenclature for Chemokine Receptors”, **Pharmacol Rev** v. 52, n. 1, pp. 145–176, Mar. 2000.

[38] DORANZ, B., ORSINI, M., TURNER, J., et al. “Identification of CXCR4 Domains that Support Coreceptor and Chemokine Receptor Functions”, **J Virol** v. 73, n. 4, pp. 2752–2761, Abr. 1999.

[39] ZHOU, N., LUO, Z., LUO, J., et al. “Structural and Functional Characterization of Human CXCR4 as a Chemokine Receptor and HIV-1 Co-Receptor by Mutagenesis and Molecular Modeling Studies”, **J Biol Chem** v. 276, n. 46, pp. 42826–42833, Nov. 2001.

[40] WU, L., LAROSA, G., KASSAM, N., et al. “Interaction of Chemokine Receptor CCR5 with its Ligands: Multiple Domains for HIV-1 gp120 Binding and a Single Domain for Chemokine Binding”, **J Exp Med** v. 186, n. 08, pp. 1373–1381, Out. 1997.

[41] DRAGIC, T., TRKOLA, A., LIN, S., et al. “Amino-Terminal Substitutions in the CCR5 Coreceptor Impair gp120 Binding and HIV Type 1 Entry”, **J Virol** v. 72, n. 1, pp. 279–285, Jan. 1998.

[42] POPE, M., HAASE, A. “Transmission, Acute HIV-1 Infection and the Quest for Strategies to Prevent Infection”, **Nat Med** v. 9, n. 7, pp. 847–852, Jul. 2003.

[43] LIU, R., PAXTON, W., CHOE, S., et al. “Homozygous Defect in HIV-1 Coreceptor Accounts for Resistance of Some Multiply-Exposed Individuals to HIV-1 Infection”, **Cell** v. 86, n. 3, pp. 367–377, Ago. 1996.

[44] SAMSON, M., LIBERT, F., DORANZ, B., et al. “Resistance to HIV-1 Infection in Caucasian Individuals Bearing Mutant Alleles of the CCR5 Chemokine Receptor Gene”, **Nature** v. 382, n. 6593, pp. 722–725, Ago. 1996.

[45] VEAZEY, R., LACKNER, A. “The Mucosal Immune System and HIV-1 Infection”, **AIDS Rev** v. 5, n. 4, pp. 245–252, Out. 2003.

[46] BOMSEL, M., DAIVID, V. “Mucosal Gatekeepers: Selecting HIV Viruses for Early Infection”, **Nat Med** v. 8, n. 2, pp. 114–116, Fev. 2002.

[47] MENDOZA, C., RODRIGUEZ, C., GARCÍA, F., et al. “Prevalence of X4 Tropic Viruses in Patients Recently Infected with HIV-1 and Lack of Association with Transmission of Drug Resistance”, **J Antimicrob Chemother** v. 59, n. 4, pp. 698–704, Abr. 2007.

[48] CORREA, R., MUÑOZ-FERNÁNDEZ, M. “Viral Phenotype Affects the Thymic Production of New T Cells in HIV-1-Infected Children”, **AIDS** v. 15, n. 15, pp. 1959–1963, Out. 2001.

[49] HAZEMBERG, M., OTTO, S., HAMANN, D., et al. “Depletion of Naive CD4 T Cells by CXCR4-using HIV-1 Variants Occurs Mainly Through Increased T-cell Death and Activation”, **AIDS** v. 17, n. 10, pp. 1419–1424, Jul. 2003.

[50] GRAY, L., STERJOVSKI, J., CHURCHILL, M., et al. “Uncoupling Coreceptor Usage of HIV-1 from Macrophage Tropism Reveals Biological Properties of CCR5-Restricted HIV-1 Isolates from Patients with AIDS”, **Virology** v. 337, n. 2, pp. 384–398, Jul. 2005.

[51] DORR, P., WESTBY, M., DOBBS, S., et al. “Maraviroc (UK-427,857), a Potent, Orally Bioavailable, and Selective Small-Molecule Inhibitor of Chemokine

Receptor CCR5 with Broad-Spectrum Anti-Human Immunodeficiency Virus Type 1 Activity", **Antimicrob Agents Chemother** v. 49, n. 11, pp. 4721–4732, Nov. 2005.

[52] MINISTÉRIO DA SAÚDE. **Protocolo Clínico e Diretrizes Terapêuticas para o Manejo da Infecção pelo HIV em Adultos**. Brasília, 2018.

[53] WHITCOMB, J., HUANG, W., FRANSEN, S., et al. "Development and Characterization of a Novel Single-Cycle Recombinant-Virus Assay to Determin HIV Type 1 Coreceptor Tropism", **Antimicrob Agents Chemother** v. 51, n. 2, pp. 566–575, Fev. 2007.

[54] TRINH, L., HAN, D., HUANG, W., et al. "Validation of an Enhanced Sensitivity Trofile HIV-1 Co-Receptor Tropism Assay for Selecting Patients for Therapy with Entry Inhibitors Targeting CCR5", **Journal of the International AIDS Society** v. 11, Suppl I, P197, Nov. 2008.

[55] RAYMON, S., DELOBEL, P., MAVIGNER, M., et al. "Development and Performance of a New Recombinant Virus Phenotypic Entry Assay to Determine HIV-1 Coreceptor Usage", **J Clin Virol** v. 47, n. 2, pp. 126–130, Fev. 2010.

[56] SCHAPIRO, J., BOUCHER, C., KURITZKES, D., et al. "Baseline CD(+) T-Cell Counts and Weighted Background Susceptibility Scores Strongly Predicts Response to Maraviroc Regimens in Experienced Patients", **Antiv Ther** v. 16, n. 3, pp. 395–404, Jan. 2011.

[57] KOHAVI, R. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", **Proceedings of the Fourtennth International Joint Conference on Artificial Intelligence**, v. 2, n. 12, pp. 1137–1143, 1995.

[58] McGOVERN., R., THIELEN, A., PORTSMOUTH, S., et al. "Population-Based Sequencing of the V3-Loop Can Predict the Virological Response to Maraviroc

in Treatment-Naive Patients of the MERIT Trial", **J Acquir Immune Defic Syndr** v. 61, n. 3, pp. 279–286, Nov. 2012.

[59] DELOBEL, P., NUGEYRE, M., CAZABAT, M., et al. "Population-Based Sequencing of the V3 Region of *Env* for Predicting Usage of Human Immunodeficiency Virus Type 1 Quasiespecies", **J Clin Microbiol** v. 45, n. 5, pp. 1572–1580, Mai. 2007.

[60] PFEIFER, N., LENGAUER, T. "Improving HIV Coreceptor Usage Prediction in the Clinic Using Hints from Next-Generation Sequencing Data", **Bioinformatics** v. 28, n. 18, pp. i589-i595, Set. 2012.

[61] KOEHLER, J., WOETZEL, N., STARITZBICHLER, R., et al. "A Unified Hidrophobicity Scale for Multispan Membrane Proteins", **Proteins** v. 76, n. 1, pp. 13–29, Jul. 2009.

[62] KYTE, J., DOOLITTLE, R. "A Simple Method for Displaying the Hydrophatic Character of a Protein", **J Mol Biol** v. 157, n. 1, pp. 105–132, Mai. 1982.

[63] EISENBERG, D., WEISS, R. M, TERWILLIGER, T. C. "The Hydrophobic Moment Detects Periodicity in Protein", **Proc Natl Acad Sci USA** v. 81, n. 1, pp. 140–144, Jan. 1984.

[64] GUY, H. "Amino Acid Side-Chain Partition Energies and Dsitribution of Residues in Soluble Proteins", **Biophys J** v. 47, n. 1, pp. 61–70, Jan. 1985.

[65] CHOWRIAPPA, P., DUA, S., KANNO, J., et al. "Protein Structure Classification Based on Conserved Hydrophobic Residues", **IEEE/ACM Trans Comput Biol Bioinformatics** v. 6, n. 4, pp. 639–651, Out. 2009.

[66] HEIDER, D., HAUKE, S., PYKA, M., et al. "Insights into the Classification of Small GTPases", **Adv Appl Bioinformatics Chem** v. 3, pp. 15–24, 2010.

[67] HEIDER, D., VERHEYEN, J., HOFFMANN, D. “Predicting Bevirimat Resistance of HIV-1 from Genotype”, **BMC Bioinformatics** v. 11, n. 1, pp. 1–9, Jan. 2010.

[68] HEIDER, D., DYBOWSKI, J., WILMS, C., et al. “A Simple Structure-Based Model for the Prediciton of HIV-1 Co-Receptor Tropism, **BioData Min** v. 7, n. 1, pp. 1–11, Ago. 2014.

[69] JAMES, G., WITTEN, D., HASTIE, T., et al., **An Introduction to Statistical Learning**. 6 ed. Nova Iorque, Springer, 2015.

[70] MONTGOMERY, D., PECK, E., VINING, G., **Introduction to Linear Regression Analysis**. 5 ed. Nova Iorque, John Wiley & Sons, 2012.

[71] AKAIKE, H. “A New Look at the Statistical Model Identification”, **IEEE Trans. Automatic Control**, v. AC-19, pp. 716–723, 1974.

[72] ZHANG, Z. “Naive Bayes Classification in R”, **Ann Transl Med** v. 4, n. 12, pp. 241–245, Jun. 2016.

[73] BREIMAN, L. “Random Forests”, **Machine Learning** v. 45, n. 1, pp. 5–32, Out. 2001.

[74] BREIMAN, L. “Bagging Predictors”, **Machine Learning** v. 26, n. 2, pp. 123–140, 1996.

[75] HAYNES, R. B., SACKETT, D. L., GUYATT, G. H., et al., **Epidemiologia Clínica: Como Realizar Pesquisa Clínica na Prática**. 3 ed. Artmed, Porto Alegre, 2008.

[76] SIERRA, S., DYBOWSKI, J., PIRONTI, A., et al. “Parameters Influencing Baseline HIV-1 Genotypic Tropism Testing Related to Clinical Outcome in Patients on Maraviroc”, **Plos One**, v. 10, n. 5, e0125502, Mai. 2015.

[77] SOVIERZOSKI, M., SCHWARZ, L., AZEVEDO, F. “Identificação em Sinais de EEG de Eventos Epileptiformes e da Piscada Palpebral com um Classificador Neural Binário”, **XXIX Congresso da Sociedade Brasileira de Computação**, Jul. 2009.

[78] FOUCHIER, R., GROENINK, M, KOOTSTRA, N., et al. “Phenotype-Associated Sequence Variation in the Third Variable Domain of the Human Immunodeficiency Virus Type 1 gp120 Molecule”, **J Virol** v. 66, n. 5, pp. 3183–3187, Mai. 1992.

[79] JONG, J., 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”, **J Virol** v. 66, n. 11, pp. 6777–6780, Nov. 1992.

[80] MILICH, L., MARGOLIN, B., SWANSTROM, R. “V3 Loop of the Human Immunodeficiency Virus Type 1 Env Protein: Interpreting Sequence Variability”, **J Virol** v. 67, n. 9, pp. 5623–5634, Set. 1993.

[81] XIAO, L., OWEN, S., 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, pp. 83–92, Jan. 1998.

[82] CARDOZO, T., KIMURA, T., PHILPOTT, S., et al. “Structural Basis for Coreceptor Selecting by the HIV Type 1 V3 Loop”, AIDS Res Hum Retroviruses v. 23, n. 3, pp. 415–426, Mar. 2007.

[83] SECLÉN, E., GARRIDO, C., GONZÁLEZ, M., et al. “High Sensitivity of Specific Genotypic Tools for Detection of X4 Variants in Antiretroviral-Experienced Patients Suitable to be Treated with CCR5 Antagonists”, **J Antimicrob Chemother** v. 65, n. 7, pp. 1486–1492, Jul. 2010.

[84] JENSEN, M., LI, F., WOUT, A., et al. “Improved Coreceptor Usage Prediction and Genotypic Monitoring of R5-to-X4 Transition by Motif Analysis of HIV Type 1 Env V3 Loop Sequence”, **J Virol** v. 77, n. 24, pp. 13376-13388, Dez. 2003.

[85] JENSEN, M., COETZER, M., WOUT, A., et al. “A Reliable Phenotype Predictor for HIV Type 1 Subtype C Based on Envelope V3 Sequences”, **J Virol** v. 80, n. 10, pp. 4698-4704, Mai. 2006.

[86] BHIVA GUIDELINES FOR THE ROUTINE INVESTIGATION AND MONITORING OF ADULT HIV-1-INFECTED INDIVIDUAL (2016). Disponível em: https://www.bhiva.org/file/DqZbRxfzlYtLg/2016-BHIVA-Monitoring-Guidelines.pdf. Acesso em: Janeiro de 2019.

[87] DOCUMENTO DE CONSENSO DE GESIDA SOBRE CONTROL Y MONITORAZACIÓN DE LA INFECTIÓN POR EL HIV (2018). Disponível em: http://gesida-seimc.org/wp-content/uploads/2018/01/gesida_DC_Control_y_Monitorizacion_b23_01_18.pdf Acesso em: Janeiro de 2019.

[88] GUIDELINES FOR THE USE OF ANTIRETROVIRAL AGENTS IN ADULTS AND ADOLESCENTS LIVING WITH HIV (2018). Disponível em: https://aidsinfo.nih.gov/contentfiles/lvguidelines/adultandadolescentgl.pdf. Acesso em: Janeiro de 2019.

[89] VANDEKERCKHOVE, L., WENSING, A., KAISER, R., et al. “European Guidelines on the Clinical Management of HIV-1 Tropism Testing”, **Lancet Infect Dis** v. 11, n. 5, pp. 394–407, Mai. 2008.

[90] CASHIN, K., STERJOVSKI, J., HARVEY, K., et al. “Covariance of Charged Amino Acids at Positions 322 and 440 of HIV-1 Env Contributes to Coreceptor Specificity of Subtype B Viruses, and Can Be Used to Improve the Performance of V3 Sequence-Based Coreceptor Usage Prediction Algorithms”, **Plos One** v. 09, n. 10, e109771, Out. 2014.

[91] RIEMENSCHNEIDER, M., CASHIN, K., BUDEUS, B., et al. “Genotypic Prediction of Subtypes A and C”, **Sci Rep** v. 6, n. 24883, pp. 1–8, Abr. 2016.

[92] JEANNE, N., SALIOU, A., CARCENAC, R., et al. “Position-Specific Automated Processing of V3 Env Ultra Deep Pyrosequencing Data for Predicting HIV-1 Tropism”, **Sci Rep** v. 05, n. 16944, pp. 1–10, Nov. 2015.

[93] GILLES, A., MEGLÉCZ, E., PECH, N., et al. “Accuracy and Quality Assesment of 454 GS-FLX Titanium and Quality Assessment of 454 GS-FLX Titanium Pyrosequencing”, **BMC Genomics** v. 12, n. 245, pp. 1–11, Mai. 2011.

[94] ARIF, M., HUNTER, J., LÉDA, A., et al. “Pace of Tropism Switch in HIV-1-Infected after Recent Infection”, **J Virol** v. 91, n. 19, e00793-17, Set. 2017.

[95] DÍEZ-FUERTES, F., DELGADO, E., VEJA, Y., et al. “Improvement of HIV-1 Coreceptor Tropism Prediction by Employing Selected Nucleotide Positions of the Env Gene in a Bayesian Network Classifier”, **J Antimicrob Chemother** v. 68, n. 7, pp. 1471–1485, Jul. 2013.

[96] XU, S., HUANG, X., ZHANG, C. “Improved Prediction of Coreceptor Usage and Phenotype of HIV-1 Based on Combined Features of V3 Loop Sequence Using Random Forest”, **J Microbiol** v. 45, n. 5, pp. 441–446. Out. 2007.

[97] LÖCHEL, H., RIEMENSCHENEIDER, M., FRISHMAN, D., et al. “SCOTCH: Subtype A Coreceptor Tropism Classification in HIV-1”, **Bioinformatics** v. 34, n. 15, pp. 2575–2580, Ago. 2018.

[98] SKRABAL, K., LOW, A., DONG, W., et al. “Determining Human Immunodeficiency Virus Coreceptor Use in a Clinical Setting: Degree of Correlation between Two Phenotypic Assays and a Bioinformatic Model”, **J Clin Microbiol** v. 45, n. 2, pp. 279–284, Fev. 2007.

[99] HE, H., GARCIA, E. “Learning from imbalanced data”, **IEEE Transactions on Knowledge and Data Engineering** v. 21, n. 9, pp. 1263–1284, Jun. 2009.