UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS
CURSO DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

DISSERTAÇÃO

REDES NEURAIS APLICADA NO DESENVOLVIMENTO DE MODELO PARA APOIO A DECISÃO NA TERAPIA ANTIRRETROVIRAL EM PORTADORES DO HIV-1

THUANY CHRISTINE LESSA DE AZEVEDO VIEIRA

2015

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

REDES NEURAIS APLICADA NO DESENVOLVIMENTO DE MODELO PARA APOIO A DECISÃO NA TERAPIA ANTIRRETROVIRAL EM PORTADORES DO HIV-1

THUANY CHRISTINE LESSA DE AZEVEDO VIEIRA

Sob a Orientação do Professor
Dr. Robson Mariano da Silva

e Co-orientação do Professor
Dr. Angel Ramon Sanchez Delgado

Dissertação submetida como requesito parcial para obtenção do grau de **Mestre em Ciências**, no Curso de Pós-Graduação em Modelagem Matemática e Computacional, Área de Concentração em Modelagem Matemática e Computacional.

Seropédica, RJ
Abril de 2015
UFRRJ / Biblioteca Central / Divisão de Processamentos Técnicos

006.3
V658r
T

Vieira, Thuany Christine Lessa de Azevedo, 1991-

Redes neurais aplicada no desenvolvimento de modelo para apoio a decisão na terapia antirretroviral em portadores do HIV-1 / Thuany Christine Lessa de Azevedo Vieira – 2015.

61 f.: il.

Orientador: Robson Mariano da Silva.

Dissertação (mestrado) – Universidade Federal Rural do Rio de Janeiro, Curso de Pós-Graduação em Modelagem Matemática e Computacional.

Bibliografia: f. 49-52.

1. Redes neurais (Computação) – Teses. 2. Inteligência artificial – Teses. 3. Inteligência artificial – Aplicações médicas – Teses. 4. Agentes antirretrovirais – Teses. 5. HIV (Vírus) – Tratamento – Teses. I. Silva, Robson Mariano da, 1963-. II. Universidade Federal Rural do Rio de Janeiro. Curso de Pós-Graduação em Modelagem Matemática e Computacional. III. Título.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS
CURSO DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

THUANY CHRISTINE LESSA DE AZEVEDO VIEIRA

Dissertação submetida como requisito parcial para obtenção do grau de Mestre em Ciências, no Curso de Pós-graduação em Modelagem Matemática e Computacional, área de concentração em Modelagem Matemática e Computacional.

Dissertação aprovada em 15 / 04 / 15

Robson Mariano da Silva. Prof. Dr. UFRRJ
(Orientador)

Angel Ramon Sanchez Delgado. Prof. Dr. UFRRJ
(Co-Orientador)

Wagner de Souza TaSsinari. Prof. Dr. UFRRJ

Rodney Cézar de Albuquerque. Prof. Dr. IFRJ

Dedicatória

Dedico este trabalho aos meus pais, Maria Lessa e Nilton, meus avós, Sandra e Camilo, e ao meu esposo, Felipe, por todas as palavras de incentivo, carinho, amor e atitudes que fizeram com que chegasse até aqui.

Agradecimentos

A Deus, pela persistência que me foi dada no decorrer desses dois anos.

A minha família, principalmente meus pais e avós, pelo incentivo, compreensão nos momentos de ausência e estresse. E mais, por tudo que representam em minha vida.

Ao Felipe, meu companheiro em cada etapa que passei, por toda a paciência, amor e ajuda depreendida.

Ao Prof. Robson Mariano da Silva, por sua orientação paciente, mesmo em momentos de dificuldade.

Ao Programa de Pós-Graduação em Modelagem Matemática e Computacional da UFRRJ, em especial a Janaína Gama por sempre ouvir nossos desabafos e constante disposição em ajudar-me.

Aos meus amigos, por sempre estarem dispostos a ajudar no que fosse necessário e alegrar meus dias apaziguando as preocupações.

A Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro-FAPERJ, pela bolsa concedida.

A todos que fazem parte da minha vida, e, em vários momentos, contribuíram para o que sou hoje. Muito obrigada.

RESUMO

Durante a última década a terapia antirretroviral (TARV) contribuiu para a redução da taxa de mortalidade e morbidade entre as pessoas infectadas pelo HIV-1. Contudo, a falha terapêutica relacionada ao surgimento de resistências aos retrovirais em função das mutações e/ou pela não adesão à terapia antirretroviral é um problema de saúde pública. Torna-se de fundamental importância a compreensão dos padrões de resistências e dos mecanismos a eles associados, possibilitando a escolha de um tratamento terapêutico apropriado que considere a frequência de mutação, quantidade de partículas virais (CV) e células $CD4^+$ entre os subtipos B e C. Portanto, o objetivo desse trabalho é desenvolver um modelo baseado em inteligência computacional para auxiliar a tomada de decisão e proporcionar melhor suporte a prática clínica e de pesquisa daqueles que lidam diretamente com pacientes. Foram utilizadas 923 amostras para esse estudo, obtidas juntos ao Laboratório de Virologia Molecular da Universidade Federal do Rio de Janeiro pertencente à rede de genotipagem do Ministério da Saúde. Inicialmente foi realizado um estudo do perfil de mutações dos subtipos B e C. Para tal foi feito um corte com pacientes com entrada no sistema a partir de 1998, com frequência de mutações na protease maior ou igual a 5% e submetidos a uma única terapia HAART com apenas um inibidor de protease, Nelfinavir (NFV), ou sem nenhum inibidor de protease. Foram realizadas 50 simulações para cada um dos subtipos usando as posições da sequência da protease como dados entrada juntamente com as taxas de carga viral e $CD4^+$. Através dos estudos foi possível observar que o subtipo C possui caráter diferenciado do subtipo B tanto em nível de CV e $CD4^+$ quanto ao número de mutações no gene da protease, fato esse que enfatiza a necessidade de tratamentos específicos para cada subtipo pelos profissionais da saúde. Além disso, o modelo demonstrou um desempenho satisfatório, possuindo um bom índice de acertos.

Palavras-chave: *Tratamento antirretroviral, HIV, Redes Neurais Artificiais.*

ABSTRACT

During the last decade, the antiretroviral therapy (ART) contributed to the reduction of the mortality and morbidity among HIV-1 affected people. However, the therapeutic flaw related to the appearance of resistence to the retrovirals due to the mutations and/or the not adherence to the antiretroviral therapy, is a problem of public health. It becomes of extreme importance, the comprehension of the resistence patterns and the mechanism related to them, enabling the choice of a suited therapeutic treatment that considers the mutation frequency, the quantity of viral particles and CD4+ cells among subtypes B and C. Therefore, the goal of the paper is to develop a model based on computational intelligence to help make decisions and give a better support to the clinic practice and research for those who deal with the patients. 923 samples were used for thus study, obtained together with the Laboratory of Molecular Virology of the Federal University of Rio de Janeiro, that belongs to the genotyping network of the Health Ministry. Initially, it was done a study of the profile of the mutations of subtypes B and C. To do so, it was made a cut of the patients that entered from 1998 on, with mutation frequency in the protease equal or greater than 5% and submitted to only one HAART therapy with just one protease inhibitor, Nelfinavir (NFV), or without any kind of protease inhibitor. Through these studies, it was possible to observe that the subtype C has a different character from subtype B, not only when it comes to the CV and CD4+ level but also the numbers of mutations in the protease gene, this fact emphasizes the necessity of specific treatments, from health professionals, for each subtype Key-words: antiretroviral treatment, HIV, Artificial Neural Network.

Palavras-chave: *Antiretroviral therapy, HIV, Artificial Neural Network*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Desde os anos 80, quando surgiram os primeiros relatos de uma nova patologia, que seria conhecida como Síndrome da Imunudodeficiência Adquirida (*em inglês: Acquired Immunodeficiency Syndrome* – AIDS), cientistas estudam as características e possíveis soluções de tratamento contra o Vírus da Imunudodeficiência Humana (*em inglês: Human Immunodeficiency Virus* – HIV), causador da AIDS. A partir dos estudos foi possível compreender a doença, que no início parecia ocorrer apenas em homossexuais, através de relações sexuais.

Ao longo dos anos identificaram como o vírus se espalhava e que seu alvo eram as células de defesa do indivíduo, tornando-o vulnerável a doenças oportunistas. Outra descoberta importante, e preocupante, foi quanto a sua variedade genética, isto é, perceberam que a estrutura genética do vírus era mutável e que isso ocorria em uma grande frequência tornando impossível encontrar uma cura para a AIDS.

Dadas essas descobertas os estudiosos da área encontraram modos de evitar que o vírus se reproduzisse dentro do organismo humano de modo que, mesmo não havendo uma cura, a expectativa e qualidade de vida do portador do HIV aumentasse, dando início assim aos Tratamentos Antiretrovirais (TARV).

Segundo boletim da UNAIDS, mais de 35 milhões de pessoas viviam com HIV no mundo e apenas 13.6 milhões delas tiveram acesso à TARV (UNAIDS, 2014). Ainda pela UNAIDS, constata-se que houve 29% de redução em mortes relacionadas à AIDS no período de 2005 a 2012. Esses dados validam a importância dos regimes terapêuticos para os pacientes infectados pelo vírus. Contudo, mesmo com resultados relevantes, os mesmos não chegam a ser 100% eficaz, devido ao alto índice de mutação decorrente das falhas terapêuticas, sendo necessário um novo regime terapêutico para o paciente.

Sendo assim, a resistência do vírus aos antirretrovirais representa um desafio no tratamento de pacientes infectados pelo HIV-1. Essa dificuldade tem sido alvo de muitos estudos com o intuito de melhorar a qualidade das terapias. Em 2007, KALMAR desenvolveu um trabalho de avaliação da resistência do vírus às drogas em pacientes que tinham interrompido o tratamento. Mais tarde, CARDOSO (2007), realizou um mapeamento no Estado de Goiás com um estudo de genotipagem para resistências de pacientes portadores do HIV-1, com a intenção de observar as características do vírus na cidade e de melhorar o tratamento de combate ao vírus. Um trabalho similar foi desenvolvido em 2010 por MACÊDO tendo como foco os pacientes oriundos dos Estados do Pará e do Amazonas.

A partir dessas pesquisas foi observada a dificuldade na elaboração de um plano de tratamento para o paciente em terapia antirretroviral, uma vez que além das mutações do vírus há também a resistência criada em relação a algum medicamento, muitas vezes consequência das variações genéticas do HIV, dando origem a estudos *in silico* com o intuito de minimizar essas imprecisões e apoiar os profissionais da saúde nesta decisão, como por exemplo o trabalho realizado por BEERENWINKEL, SHMIDT, WALTER, KAISER, LENGAUER, HOFFMANN, KORN e SELBIG (2001), no qual foi desenvolvido um sistema inteligente, o *Geno2pheno*, que prevê resistência genotípica analisando a sequência genômica do vírus. Com base nesses dados, mostrou-se necessário desenvolver um modelo que possa servir como ferramenta auxiliar aos médicos no processo de tratamento. BALASUBRAMANIE e FLORENCE (2009) também contribuíram para a problemática com a criação de um modelo de especificação de regimes terapêuticos fazendo uso de Base Radial. Ainda relatam ter desenvolvido também um modelo MLP que obteve resultados ainda melhores do que com Base Radial.

2. JUSTIFICATIVA

O HIV-1 foi detectado há aproximadamente 30 anos, e ainda é uma das principais causas de morte no mundo. Segundo Ministério da Saúde, no Brasil estima-se que cerca de 630 mil indivíduos de 15 a 49 anos vivem com HIV, sendo que 255 mil não sabem que são portadores do vírus (MINISTÉRIO DA SAÚDE, 2012).

Mesmo depois de tanto tempo os números ainda são altos, apesar das campanhas de conscientização e dos tratamentos antirretrovirais, que hoje são fornecidos pelo ministério da saúde gratuitamente em determinadas regiões do Brasil, o vírus continua se espalhando e as pesquisas apontam um crescimento na taxa de infectados, principalmente entre os jovens do sexo masculino.

A UNAIDS e seus parceiros estabeleceram metas para 2015, dentre elas estão aumentar o acesso à terapia antirretroviral para alcançar 15 milhões de pessoas e reduzir as taxas de transmissão do vírus. Como não existe uma cura para o HIV, os tratamentos são dados com o intuito de melhorar a qualidade de vida do indivíduo infectado, contudo, um dos grandes problemas enfrentado é a falha terapêutica, que obriga uma alteração no regime antirretroviral, fazendo com que o médico ou especialista da área tenha que montar uma nova terapia.

Esse processo pode se tornar complicado dependendo da situação do paciente, como tempo de diagnóstico ou de vida. Devido a essas dificuldades, modelos matemáticos, estatísticos e computacionais vêm sendo cada vez mais utilizados como ferramenta no processo de tratamento do vírus.

Os modelos, em geral, são usados para predição de falhas terapêuticas ou de tendência a resistências em determinadas posições da sequência da protease de modo que o profissional da saúde possa usar como um auxílio no caminho a seguir com seu paciente com o intuito de fornecer uma vida com mais qualidade. Com esta ideia, o presente trabalho visa à criação de um modelo usando redes neurais artificiais de perceptron múltiplas camadas para auxiliar na escolha da terapia antirretroviral a ser escolhida.

3. OBJETIVOS

3.1. Objetivo Geral

Elaborar um modelo computacional estruturado em Redes Neurais Artificiais, que possa auxiliar na orientação dos regimes terapêuticos em pacientes portadores do HIV-1, com falha terapêutica aos inibidores antirretroviral.

3.2. Objetivos Específicos

- Observar os regimes terapêuticos dos pacientes portadores do HIV-1;
- Estudar as tendências de mutações em posições clássicas relacionadas aos regimes;
- Verificar a capacidade de uma Rede Neural Artificial de Perceptron de Múltiplas Camadas em classificar resistências ao inibidor de protease, Nelfinavir;
- Avaliar a metodologia de corte na frequência de mutações na sequência da protease;
- Examinar a contagem de taxa de carga viral e CD4+ nos subtipos B e C.

4. FUNDAMENTAÇÃO TEÓRICA

4.1. HIV

4.1.1. Conceito

O vírus da imunodeficiência humana (HIV) é um retrovírus da família *Retroviridae*, do gênero *Lentivirus* e causa a síndrome da imunodeficiência adquirida (AIDS) (GALLO 1984). Como todos os vírus da subfamília *lentiviridae*, possui período de incubação prolongado antes que os sintomas da doença apareçam, infectam células do sangue e do sistema nervoso e suprimem o sistema imunológico do indivíduo infectado. Uma de suas principais características é sua diversidade genética, chegando a ser superior a 10% em uma pessoa infectada (PINTO e STRUCHINER, 2006). Devido a essa variabilidade genética o vírus está organizado em três grupos, M, N e O, sendo do M o grupo principal e mais estudado, podendo ser subdividido filogeneticamente em vários subtipos, como A, B, C, D, F, G, H, J e K, conforme ilustra a Figura 1. Além de diversas formas recombinantes circulantes (CRF).

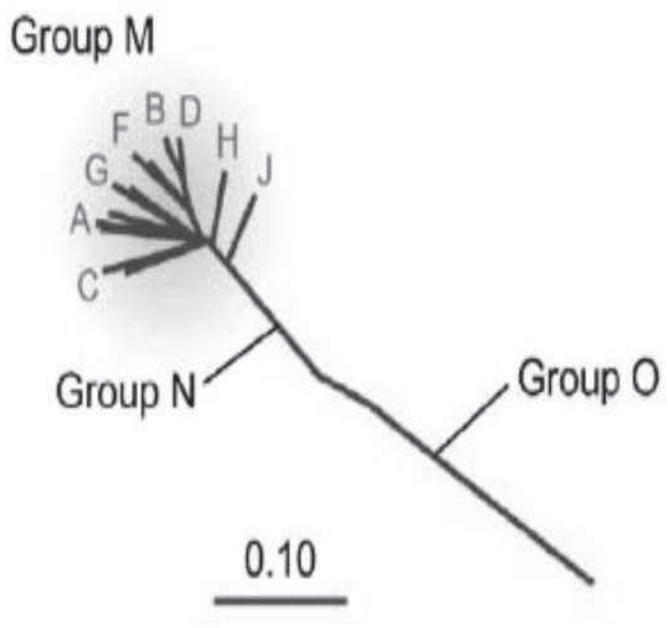

Figura 1: Classificação da diversidade do HIV-1.

No Brasil, além do subtipo B, que é predominante, estudos apontam a presença dos

subtipos C e F em algumas regiões. MORGADO *et al*. (1998) e TANURI *et al.* (1999) verificaram, que no Estado do Rio de Janeiro o subtipo B, de fato, é predominante, além disso, encontraram também uma porcentagem referente ao subtipo F e uma minoria para o subtipo D. Já SOARES et al.encontraram evidências de alterações na distribuição dos subtipos pelo Brasil.

A figura 2 apresenta os dados do estudo publicado por BRINDEIRO *et al*.(2003), que analisou indivíduos portadores do HIV-1 que não aderiram a terapia ARV. A amostragem foi 535 indivíduos.

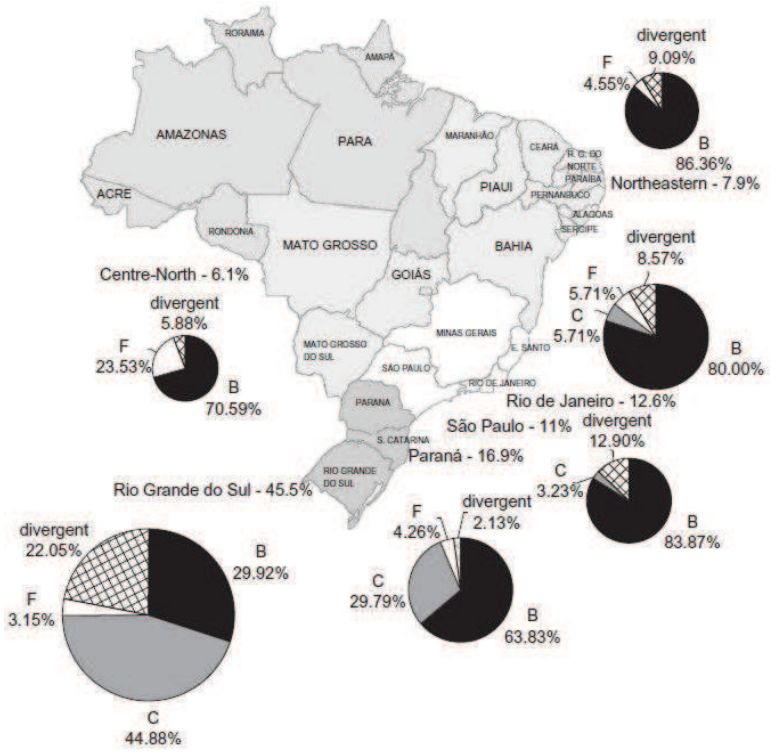

Figura 2: Mapa do Brasil mostrando a distribuição dos subtipos. Baseado na análise de PR e RT. As amostras que apresentam discordância entre PR e RT foram consideradas como genomas divergentes. As áreas dos gráficos são proporcionais a quantidade de amostras de cada local analisado.

4.1.2. Estrutura genômica

O genoma do HIV tem, aproximadamente, 9,8 Kb, com 9 genes que apresentam diversas possibilidades de processamento alternativos, o que permite a síntese de um grande número de diferentes polipeptídeos, proteínas e enzimas. Três desses genes são comuns a todos os retrovírus, são eles o *gag*, *pol* e *env*, e codificam importantes proteínas e enzimas que participam diretamente da estrutura do vírus ou de seu ciclo reprodutivo. Os demais genes são regulatórios (*tat* e *rev*) e acessórios (*nef*, *vif*, *vpr* e *vpu*). A figura 3 representa a estrutura

genômica do vírus.

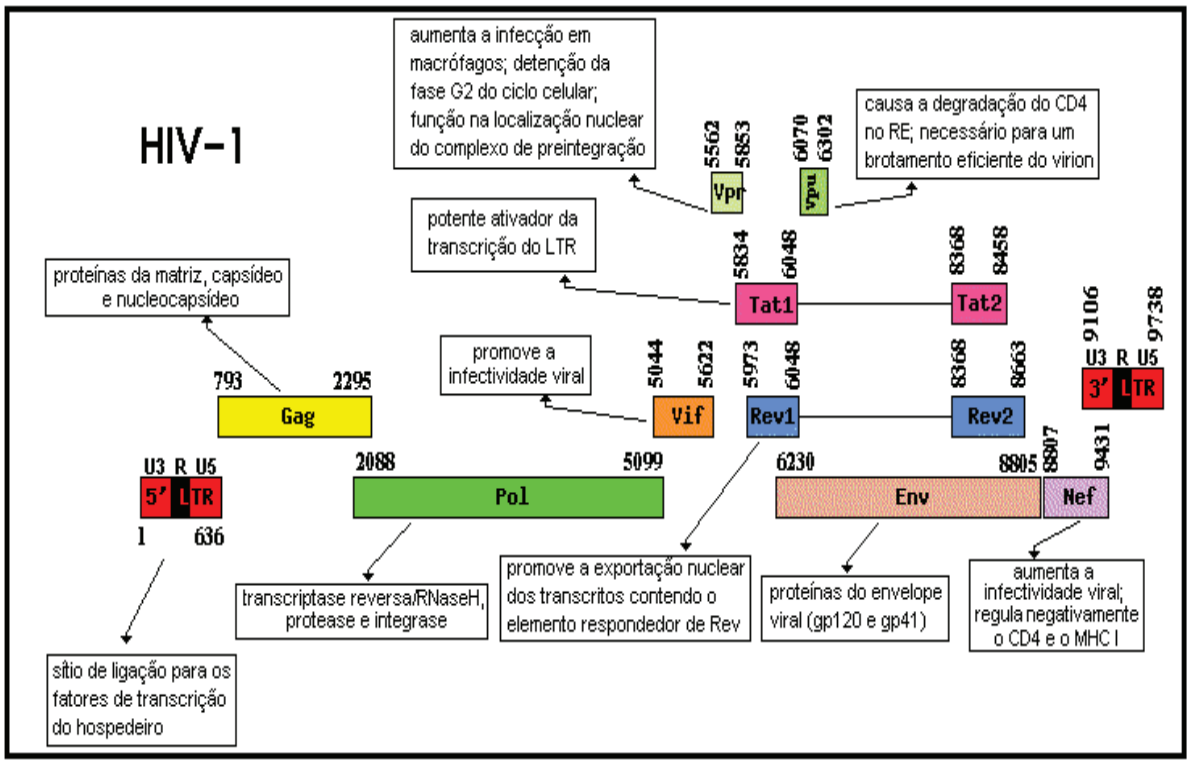

Figura 3: Representação esquemática da estrutura do genômica do HIV-1 (adpatado de www.nature.com)

O gene *gag* produz proteínas virais estruturais. Já o gene *env* codifica as glicoproteínas do envelope viral, media a entrada do vírus na célula alvo, junto a outros componentes celulares. Devido a integração das glicoproteínas do envelope com o sistema imunológico acarreta pressão seletiva, possibilitando o escape imune do vírus. O *pol* codifica poliproteínas, organizando as enzimas virais transcriptase reversa (RT), protease (PR) e integrase (IN)

.

4.1.3. Ciclo de replicação

A Figura 4 exemplifica o processo de replicação do vírus.

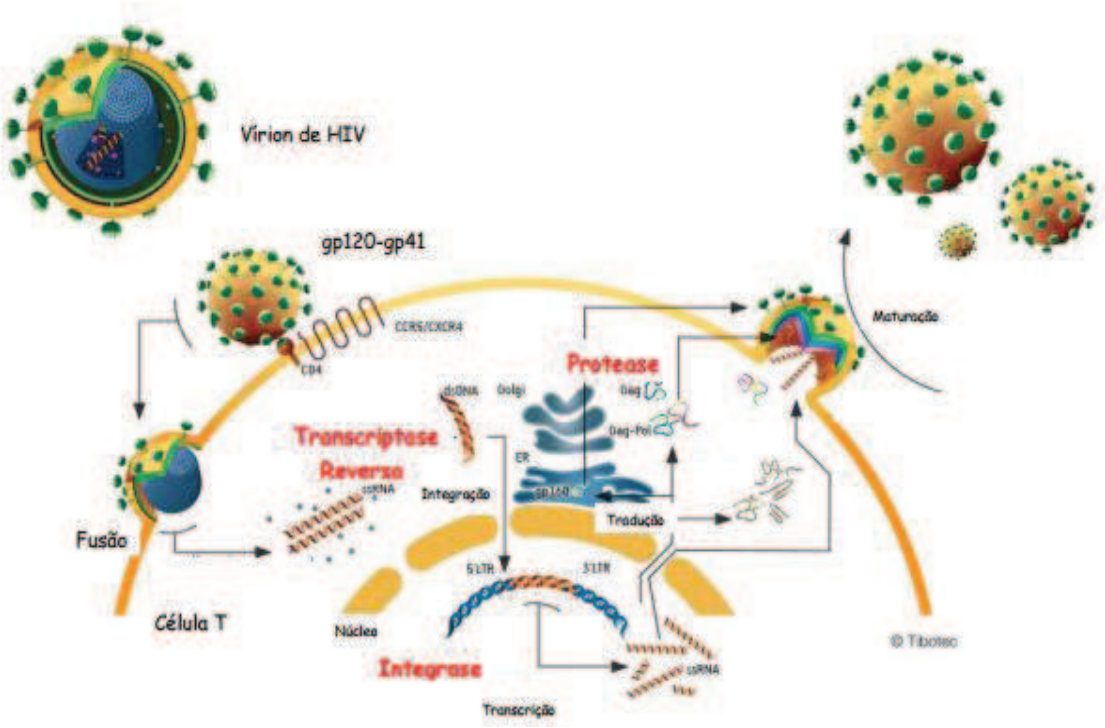

Figura 4: Ciclo de Reprodução do HIV-1(Adaptada de www.tibotec.com).

O processo de replicação do HIV-1 inicia-se a partir do momento em que o vírus se fundi a membrana celular da célula hospedeira, fazendo com que seu RNA e suas enzimas virais entrem na mesma. Neste momento, a RT do vírus sintetiza o DNA viral ao DNA celular através do RNA do HIV. É formado então um pró-vírus através da união do DNA viral com o celular por atuação da integrase. Com esse processo, a célula inicia a retrotranscrição e começa a produzir RNA e proteínas virais. Em seguida, as partículas virais passam pela maturação. A protease quebra as poliproteínas virais habilitando-as a unir o RNA em novas partículas, formando o vírions. Esta etapa encerra a replicação do HIV. Os vírus amadurecem e tornam-se capazes de infectar novas células e repetir todos os processos para se replicar.

4.1.4. Mutação

Além da velocidade de reprodução do HIV, outra característica que dificulta o tratamento contra a AIDS é sua diversidade genética, como já foi falado. O vírus da imunodeficiência chega a10^6mutações por dia, ou seja, por dia ocorrem cerca de um milhão de alterações na sequência dos aminoácidos. As mutações podem ser classificadas como primárias ou principal, secundárias ou acessórias e polimorfismo natural. Este são variantes genéticas que se replicam de forma semelhante as variantes selvagens, sem expressão fenotípica e ocorrem em pacientes que não foram submetidos a esquemas terapêuticos (WILSON & BEAN, 2000. HAUBRICH, 2004).

As mutações primárias costumam ocorrer durante tratamentos. Alteram a ligação de uma droga ao seu alvo específico e torna-se necessário o aumento da droga para inibir a enzima alvo, isto é, conferem elevado nível de resistência a uma ou várias drogas. Já as secundárias, não apresentam resistências as drogas isoladamente. Entretanto, quando unidas as primárias, contribuem na reconstrução da capacidade de replicação do vírus já contendo uma mutação principal (HIRSCH, *et al*., 2000).

4.1.5. Resistência

Como descrito, a grande taxa de mutação do vírus é uma importante característica, assim, um paciente pode ter inúmeras amostras diferentes do HIV. Este fato é fundamental para analisar os efeitos dos medicamentos, o surgimento de vírus resistentes e, por consequência, a dificuldade em combater a AIDS.

A submissão de um paciente à terapia antirretroviral causa a eliminação de uma grande parte da população viral. Porém, em decorrência da grande variedade genética do vírus, uma fração dessa população sobrevive. A replicação residual, sob o uso contínuo das drogas antirretrovirais, gera um ambiente de seleção natural. O surgimento de mutações associadas à resistência dá a esta variante viral uma vantagem seletiva. Consequentemente essa pressão possibilita a proliferação destas variantes, tornando-as predominantes. Sendo assim, a resistência é simultaneamente causa e consequência da replicação do vírus na presença das drogas antirretrovirais.

Segundo SHAFER (2002), o surgimento de resistências aos regimes terapêuticos é um fator limitante para o sucesso do combate a AIDS. Muitos pacientes sofrem falha terapêutica devido a não adesão ao tratamento e a elevada replicação de vírus resistentes aos medicamentos antirretrovirais. O acúmulo dessas mutações diminui a sensibilidade às drogas, reduzindo a eficácia do regime terapêutico aos poucos. Sendo assim, a contínua replicação do vírus submetidos às drogas aumenta a resistência às mesmas, gerando assim um ciclo vicioso de falha terapêutica, deixando o tratamento ainda mais complicado.

Quando há a ausência de medicamentos, os vírus apresentam um *fitness* reduzido, isto é, eles não se tornam mais fortes através de resistências oriundas das drogas, uma vez que não há pressão seletiva. Quando ocorre interrupção no tratamento, os vírus mutantes são substituídos por vírus selvagens (*wild-type*), ou suscetíveis aos tratamentos progressivamente. Contudo, estas populações, mesmo enfraquecidas, se mantêm no plasma, muitas vezes em populações virais minoritárias, chegando a ser indetectável nos testes de resistência, em alguns casos (VANDAMME et al., 2004).

O estudo em vírus resistentes às drogas ARV a partir do sequenciamento do gene pol, evidencia a existência de determinadas posições na cadeia peptídica que são alvos específicos para mutações que resultam na troca de aminoácido. Isto ocorre tanto na transcriptase reversa como na protease. Essas mutações na cadeia estão relacionadas à resistência aos regimes antirretrovirais e são classificadas como mutação primária ou mutação compensatória, também conhecida como secundária (SHAFER, 2002).

A denominação primária refere-se às mutações que por si só reduzem a suscetibilidade a uma droga. Já a denominação compensatória refere-se às mutações que conjuntamente com uma mutação primária diminuem a sensibilidade a uma determinada droga ou melhoram o *fitness* viral. Segundo SHAFER (2002), algumas mutações podem ser consideradas primárias relativamente a uma determinada droga e compensatória para outra.

Além dessas mutações, existem os polimorfismos genéticos, que são definidos como variações genéticas comuns (frequência maior que 1%) dos vírus isolados de indivíduos sem

terapia ARV (naive de tratamento). Ou seja, estão presentes na população viral mesmo na ausência da pressão seletiva exercida pelas drogas antirretrovirais (DIAS, 2004).

Nas figuras 5, 6, 7, 8 e 9 podemos observar as posições de mutações de alguns dos medicamentos usados no combate a AIDS. A letra superior refere-se ao aminoácido encontrado no tipo selvagem do vírus, enquanto a inferior ao aminoácido de substituição. Já as numerações em negrito referem-se ás mutações primárias, e as restantes as mutações secundárias.

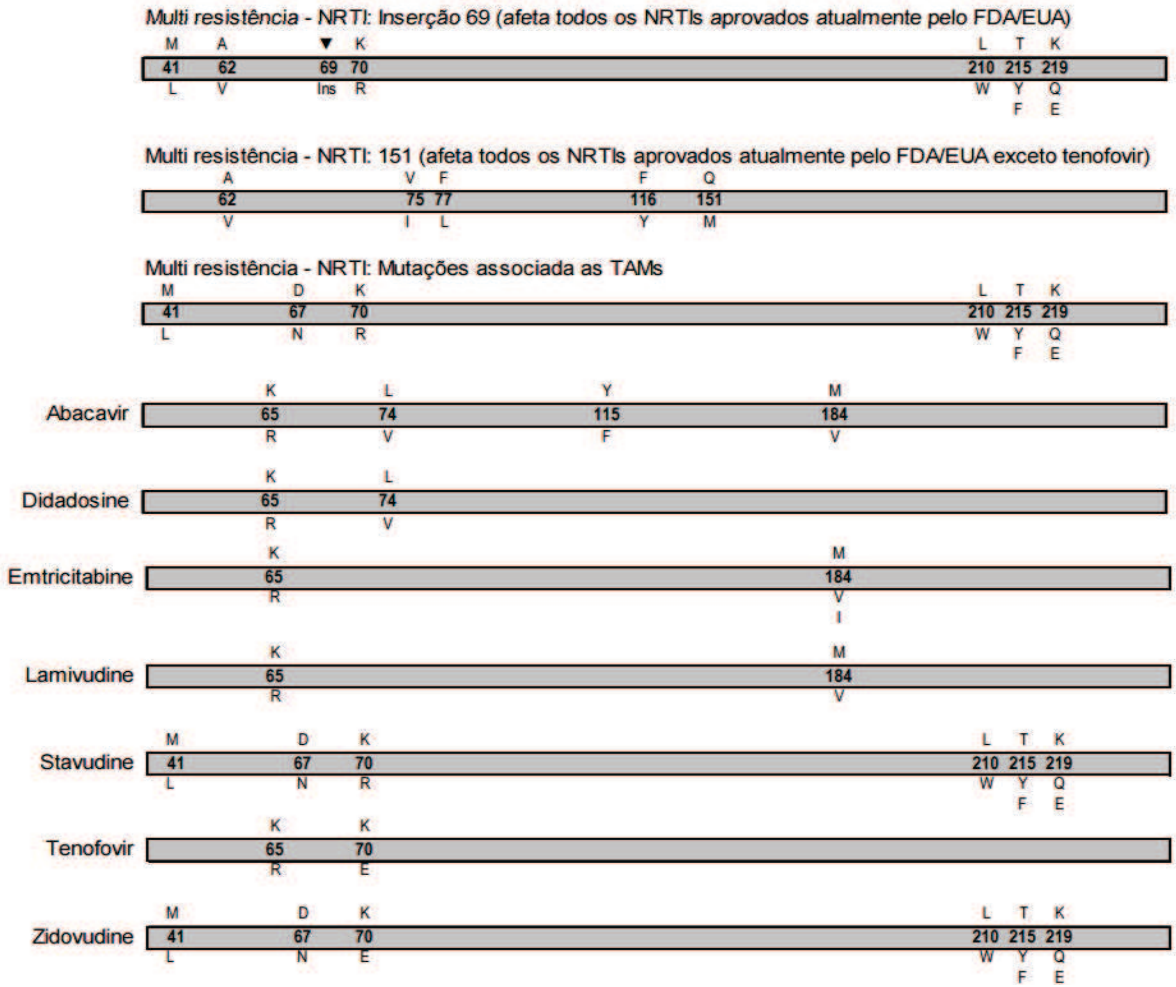

Figura 5: Resistência aos Inibidores Análogos Nucleosídeos (adaptada de JOHNSON et al., 2014).

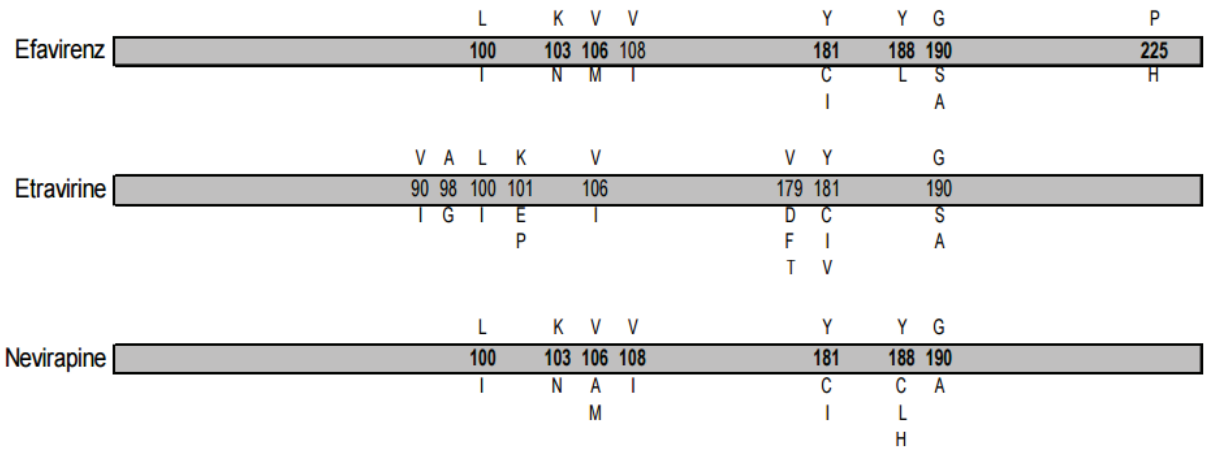

Figura 6: Resistência aos Inibidores Análogos não Nucleosídeos (adaptada de JOHNSON et al., 2014)

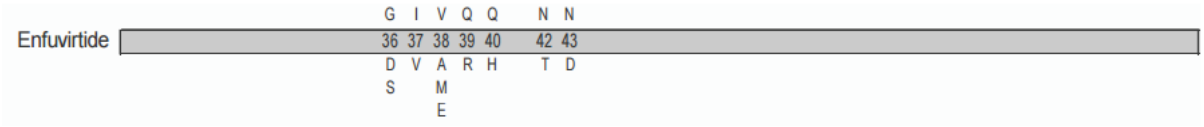

Figura 7: Resistência aos Inibidores de Fusão (adaptada de JOHNSON et al., 2014)

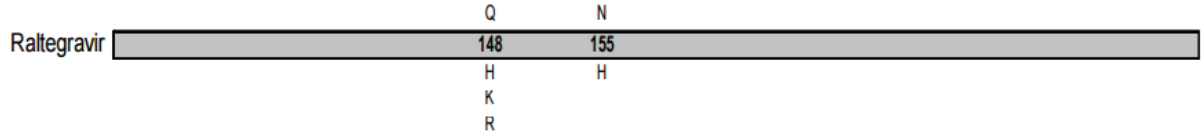

Figura 8: Resistência aos Inibidores de Integrase (adaptada de JOHNSON et al., 2014)

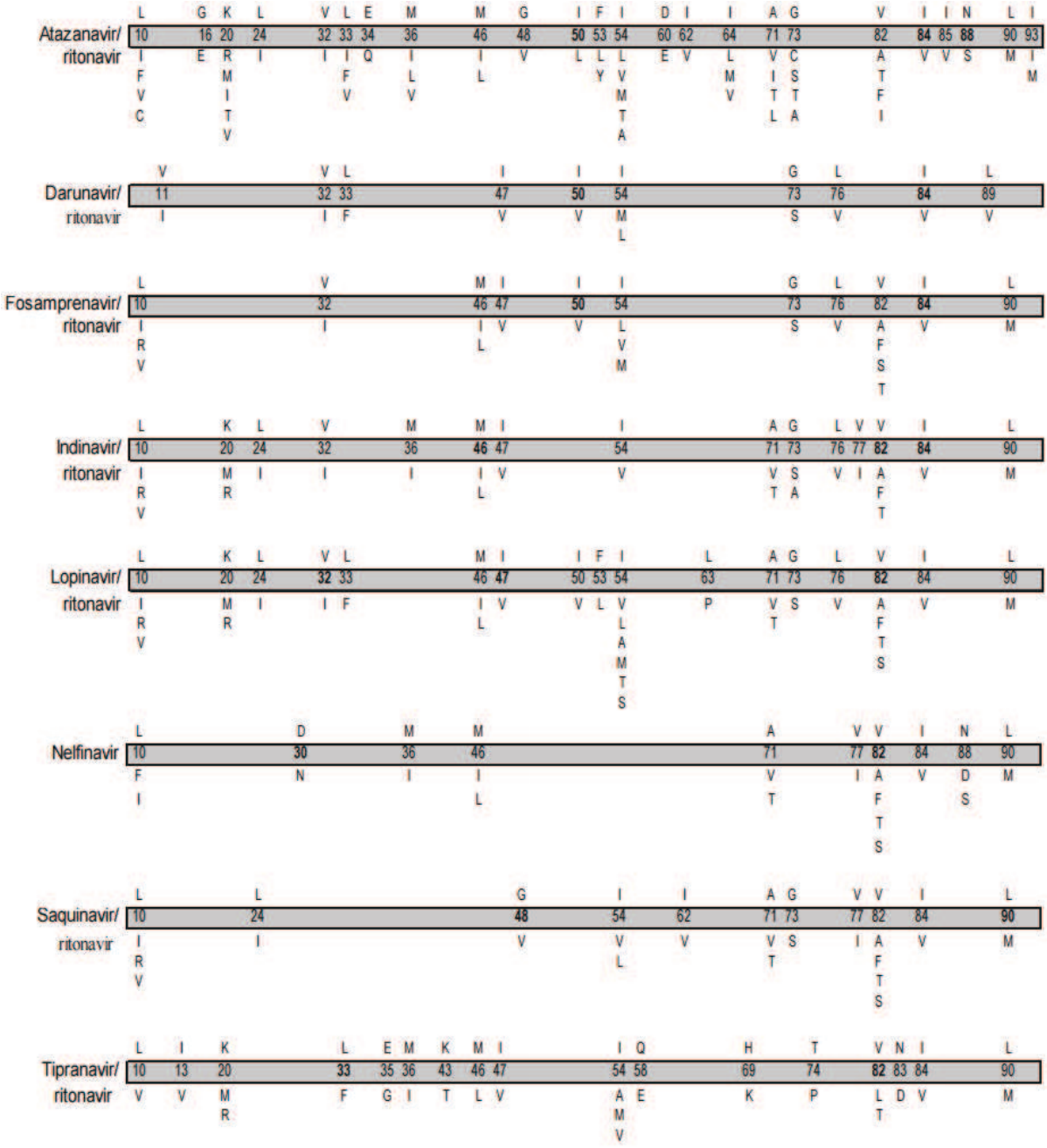

Figura 9: Resistência aos Inibidores da Protease (adaptada de JOHNSON et al., 2014).

4.1.6. Terapia antirretroviral

Devido à alta taxa de mutação do vírus seu combate ainda não é completamente eficaz, além de não ser possível atacar ao HIV-1 diretamente, uma vez que sua estrutura genômica não possui um padrão. Portanto, a solução encontrada foi utilizar drogas antirretrovirais (ARTV) para interromper seu ciclo de replicação, reduzindo assim a carga viral de modo que prolongue o tempo dos pacientes sem o surgimento de doenças oportunistas e diminuindo a transmissão do HIV (Detels *et al.*, 1998). Na atualidade, há um grande e crescente número de drogas antirretrovirais para o tratamento de pacientes infectados por HIV-1 (KATZUNG, 2006). Todas as etapas no ciclo reprodutivo do vírus são alvos potenciais para um ARTV (STROHL*et al*., 2004). Na Figura 10 é possível observar os alvos da terapia antirretroviral na ilustração de uma célula hospedeira humana.

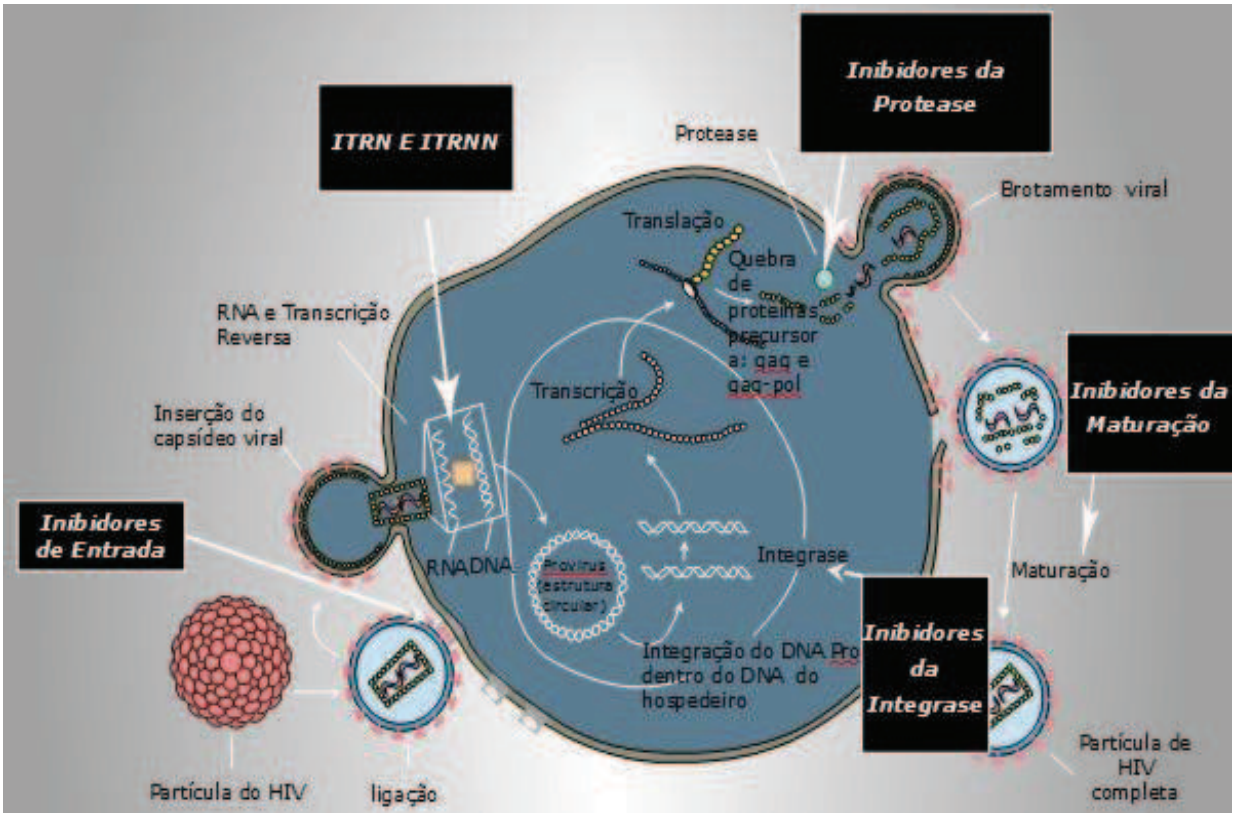

Figura 10: Adaptado do livro HIV/AIDS Handbook. 4th ed. Boston: Total Learning Concepts, 1999; Ritchie DJ. In: Powderly WG, ed. Manual of HIV Therapeutics. Philadelphia: Lippincott-Raven, 1997:33-41.

Os antirretrovirais estão divididos em cinco classes, sendo duas de inibidores de transcriptase reversa (NRTI e NNRTI), uma de inibidores de protease (IP), uma de inibidores de fusão (IF) e uma de inibidores de integrase (IN).

Para que o HIV-1 entre na célula hospedeira necessita de uma sequência de eventos como ligações proteicas e alterações estruturais, para que o vírus possa se fundir a parede celular e então, entrar na célula hospedeira. Os inibidores de fusão atuam interferindo em alguma parte desse processo (Biswas *et al*., 2007).

Uma vez que o vírus consegue completar a fusão, seu RNA e suas enzimas espalham-se pela célula e a transcriptase reversa realiza a síntese do DNA viral através do RNA do vírus. Esta etapa é de grande importância no processo de replicação do vírus, por esse motivo foi a primeira a ser estudada e ser desenvolvida terapias antirretrovirais (ARV), os inibidores de transcriptase reversa são os medicamentos mais utilizados nas terapias até hoje. Tanto os inibidores de transcriptase reversa análogos de nucleosídeos (NRTI) quanto os de transcriptase reversa não análogos de nucleosídeos (NNRTI), como os nomes já propõem atuam interferindo a transcriptase reversa. O segundo possui atividades aditivas com a maioria dos outros antirretrovirais, uma vez que não são competitivos com a TR. Os NNRTI ligam-se a TR bloqueando seu local de ativação e catalisação. Já os NRTI impedem a formação da dupla cadeia de DNA, abortando a transcrição, e consequentemente, a síntese do DNA (ROBBINS, 1998).

Os IN bloqueiam a ação da integrase, enzima que auxilia a entrada do DNA viral no DNA da célula hospedeira, onde produzirá novos virions. Sendo assim, os IN previnem esse processo.

No fim do ciclo reprodutivo do HIV-1, as partículas virais passam pelo processo de maturação. As poliproteínas virais são quebradas pela protease possibilitando a junção do RNA em novas partículas, formando os virions. Os IP bloqueiam a quebra das poliproteínas, deste modo, as partículas produzidas por células afetadas por IP possuem virions não processados e não-infecciosos (PENG *et al*., 1989).

A Tabela 1 resume os medicamentos aplicados nas terapias antirretrovirais separados por suas classes:

Tabela 1: Principais antirretrovirais (continua).

Classe	Nome genérico	Mecanismo de ação
Inibidores da Transcriptase Reversa Análogos de Nucleosídeos (ITRN)	Abacavir (ABC), Didanosina (13DI), Entricibatina, Estavudina (d4T), Lamivudina (3TC), Stavudina, Zidovudina (AZT) e Tenofovir (TDF)	Impedem a infecção das células, pois atuam sobre a transcriptase reversa, impedindo que o RNA viral se transforme em DNA complementar
Inibidores da Transcriptase Reversa Não-Análogos de Nucleosídeos (ITRNN)	Efavirenz (EFZ), Nevirapina (NVP), Etravirina	Também atuam sobre a transcriptase reversa impedindo que o RNA viral se transforme em DNA complementar

Tabela 1: Continuação.

Classe	Nome genérico	Mecanismo de ação
Inibidores de Protease (IP)(Inibidores da 'Maturação')	Fosamprenavir/r (FAPV/r), Atazanavir/r (ATV/r), Darunavir/r (DRV/r), Indinavir/r (IDV/r), Lopinavir/r (LPV/r), Nelfinavir (NFV), Saquinavir/r (SQV/r), Tripanavir/r	Atuam impedindo a clivagem da protease do polipeptídeo precursor viral e bloqueia a maturação do vírus
Inibidor da fusão (IF)	Enfuvirtida (T-20)	Impedem a entrada do material genético viral pela sua ação no mesmo local da entrada do HIV na célula que expressa receptor CD4
Inibidor de Integrase	Raltegravir	Bloqueiam a atividade da enzima integrase, responsável pela inserção do DNA do HIV ao DNA humano. Assim, inibe a replicação do vírus e sua capacidade de infectar novas células.

4.2. Redes Neurais

4.2.1. Conceito

A Rede Neural, por ser inspirada no cérebro humano, é uma máquina projetada com neurônios artificiais. Ela é uma ferramenta de inteligência artificial que se adapta e aprende a realizar tarefas ou comportamentos a partir de um conjunto de dados como exemplo (OSÓRIO e BITTENCOURT, 2000). Em outras palavras, podem-se definir redes neurais artificiais como técnicas computacionais que modelam de modo a imitar o cérebro humano (HEATON, 2010) uma tarefa particular ou uma função de interesse.

O conhecimento, em uma rede neural, é adquirido através da experiência, ou seja, a partir de um processo de aprendizagem, similar ao aprendizado do cérebro. Outra semelhança entre eles é a força de conexão entre neurônios, conhecidas como pesos sinápticos, são usados para armazenar o conhecimento adquirido.

O sistema nervoso é composto por um conjunto de neurônios. Nestes a comunicação é feita através de impulsos, de um neurônio para outro, através de ligações denominadas dendritos. A soma de todos os impulsos representa uma energia que gera um grau de ativação no neurônio que a recebeu, a partir desse processo uma resposta é gerada na forma de impulso, sendo transmitido para o próximo neurônio pelo axioma (MACHADO, 2005; AGGARWAL; SONG, 1998). O que transmite, controla a frequência, aumentando ou diminuindo a polaridade na membrana pós-sináptica. Eles são essenciais para determinar o funcionamento, comportamento e o raciocínio do ser humano. As redes naturais não propagam sinais negativos, sua ativação está relacionada a frequência com que emite pulsos, que por sua vez, são contínuos e positivos. Ao contrário das redes neurais artificiais, as redes naturais não são uniformes. A Figura 11 representa um neurônio natural, no qual as redes neurais artificiais são baseadas.

Figura 11: Neurônio natural.

A Rede Neural Artificial (RNA) é constituída por neurônios artificiais interligados onde cada ligação possui um ou mais pesos sinápticos responsáveis por armazenar as informações. Uma importante característica da rede é sua função de ativação, que tem como entrada o somatório dos pesos sinápticos, que são extremamente importantes para as redes neurais, uma vez que determinam toda a manipulação de valores da rede.

Como visto, o neurônio artificial é uma unidade de processamento fundamental para a operação de uma RNA. A figura 12 mostra o modelo de um neurônio, nela é possível identificar três elementos básicos de um modelo neural.

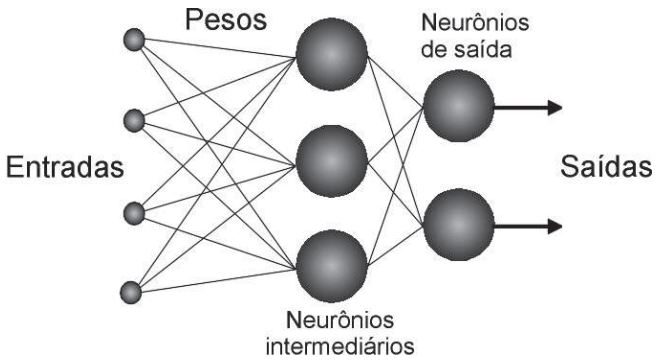

Figura 12: Modelo de um neurônio artificial.

O primeiro é o conjunto de *sinápses* ou *elos de conexão,* caracterizado por um *peso* ou *força*. O segundo é o *somatório* ou *junção aditiva,* responsável por somar os sinais de entrada, ponderado pelas respectivas sinápses do neurônio. E, por último, destaca-se a *função de ativação*, que restringe a amplitude da saída de um neurônio.

É interessante ressaltar que uma RNA é composta por uma camada de entrada e uma camada de saída e, em geral, possui camadas ocultas ou intermediárias. A Figura 13, a seguir, ilustra um modelo em rede neural artificial.

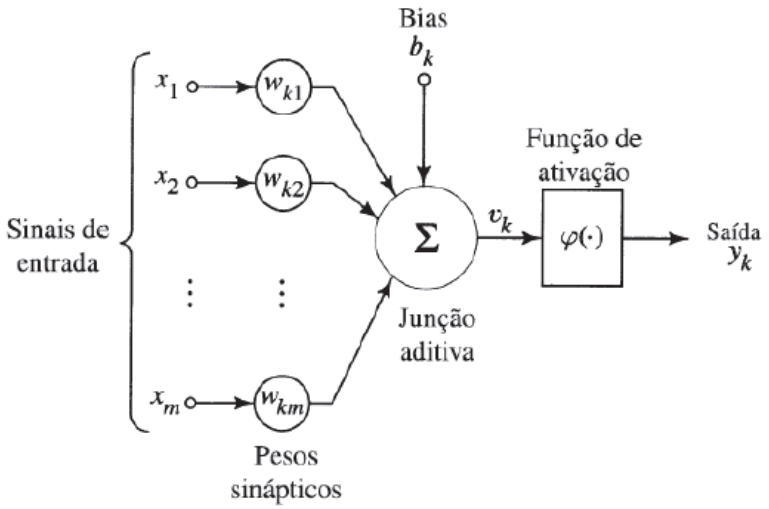

Figura 13: Estrutura de uma RNA.

Dentre as propriedades das redes neurais, a mais importante é sua habilidade em aprender de seu ambiente e com isso melhorar seu desempenho, além de ser capaz de se modificar de acordo com a necessidade de aprender o que lhe foi apresentada. Esse processo é realizado através de ajustes aplicados aos pesos da rede, o treinamento.

O aprendizado ocorre quando a rede alcança um resultado generalizado para uma classe de problemas. A rede neural pode se relacionar com o ambiente de duas maneiras distintas, aprendizado supervisionado e o não supervisionado (AGUIAR *et al*., 2007), o primeiro necessita de um supervisor enquanto o segundo não há supervisor (HAYKIN, 2001).

4.2.2. Tipos de rede

Em 1943, McCulloch e Pitts criaram o primeiro modelo de neurônio artificial. Este era uma simplificação do que se conhecia sobre os neurônios naturais, onde o impulso recebido o fazia ultrapassar o limiar através da soma dos impulsos. Conforme figura 14.

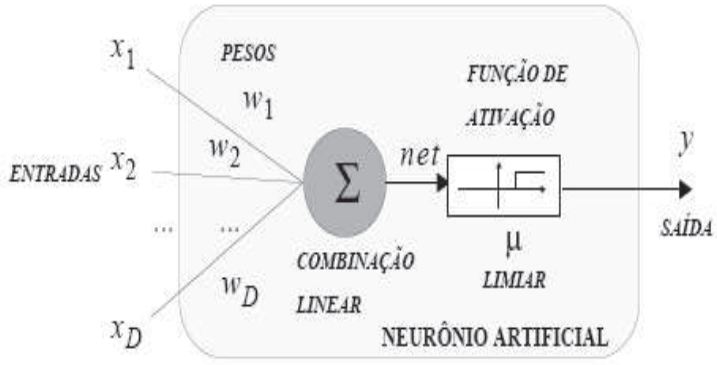

Figura 14: Modelo de neurônio de McCulloch e Pitts (1943).

- Onde x_i são as entradas.
- $w_i$são os pesos.
- y é a saída da rede.

O funcionamento da rede era bem simples, a soma dos valores $x_i w_i$ são recebidas pelo neurônio e a função de ativação decide a saída do neurônio comparando o limiar com a soma obtida. Sendo assim, a responsabilidade de gerar a saída y é atribuída a função de ativação do neurônio. Apesar de existirem diversas funções de ativação, as mais aplicadas são as *funções degrau, linear* e *sigmoides* (Equações 1, 2, 3 e 4) (BRAGA; CARVALHO; LUDEMIR, 2012, , HAYKIN, 2001, AGGARWAL; SONG, 1998).

i. Função de ativação degrau:

$$f(u_i) = \begin{cases} 1 \; se \; u_i \geq 0 \\ 0 \; se \; u_i < 0 \end{cases} \quad (1)$$

ii. Função linear:

$$f(u_i) = u_i \quad (2)$$

iii. Funções Sigmóides:

1. Função Logística:

$$f(u_i) = \frac{1}{(1 + \mathrm{e}^{(-u_i)})} \quad (3)$$

2. Função Tangente Hiperbólica

$$f(u_i) = \tanh(\frac{u_i}{2}) = \frac{1 - \exp(-u_i)}{1 + \exp(-u_i)} \quad (4)$$

A Figura 15 representa o comportamento das funções acima em um plano cartesiano, no qual o eixo x representa as entradas dos neurônios e o eixo y, as saídas.

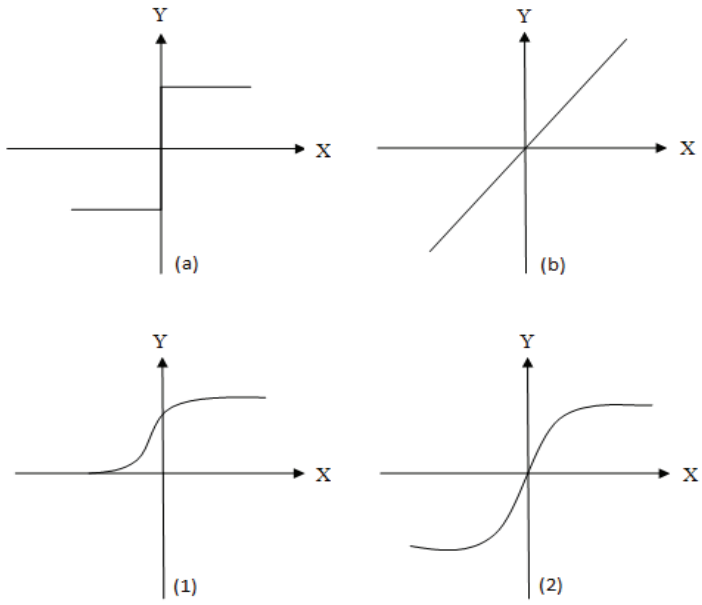

Figura 15: Gráficos dos comportamentos das funções de ativação. (a) Função Degrau, (b) Função Linear, (1) Função Logística, (2) Função Tangente Hiperbólica

4.2.3. Perceptron (Feedforward)

Uma rede feedforward possui suas camadas conectadas às outras, mas sempre em uma única direção, partindo da camada de entrada para a camada de saída. Além dessa, há as redes neurais recorrentes, são estruturas de processamento capazes de representar uma variedade de comportamentos dinâmicos. São redes que se assemelham a filtros não-lineares com resposta ao impulso infinita (NERRAND et al., 1993).

O Perceptron foi desenvolvido por Rosenblatt (1958) e é uma rede feedforward de uma única camada cujos pesos e erros podem ser treinados para obter uma saída esperada, representa em um vetor, através de uma entrada, também em vetor. Esse tipo de rede é caracterizado por um modelo de aprendizagem supervisionado, onde cada entrada é ponderada com peso w_{1j} e a soma das entradas ponderadas é a entrada da função de transferência, que tem como saída 0 ou 1, conforme a Figura 16 ilustra, a camada única se

refere à camada de saída.

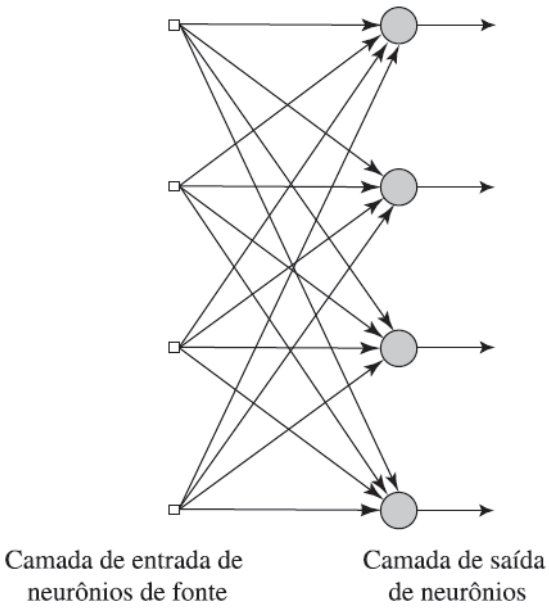

Figura 16: Modelo Perceptron.

O aprendizado do perceptron utiliza a diferença entre a resposta da rede e o vetor de saída desejado, gerando um determinado erro. A cada iteração de treinamento, a rede pode melhorar seu cálculo, obtendo um menor erro ajustando os pesos, que são aleatórios inicialmente. Sendo assim, o aprendizado é alcançado através do treino.

4.2.4. Perceptron de múltiplas camadas

As redes Perceptrons de Múltiplas Camadas (MLP) pertencem a classe das redes feedforward. As redes MLP começaram a ser estudadas por Minsky e Papert. Após os estudos de Rumelhart, Hinton e Williams em 1986 referente ao algoritmo de treinamento backpropagation, foi possível criar as redes MLP mais atualizadas.

Essa classe de rede neural distingui-se pela existência de uma ou mais camadas ocultas, que tem como função intervir entre as camadas de entrada e saída possibilitando resultados estatísticos mais elevados. Se tratando de redes de múltiplas camadas não é possível obter o erro do mesmo modo que para redes de camada única, pois deve-se considerar os erros das camadas intermediárias.

Um perceptron que possui uma camada escondida já é caracteriazado como perceptron de múltiplas camadas. Em cada camada os neurônios recebem entradas vindas da camadas anterior e é realizada uma combinação linear dessas variáveis.

O treinamento dessas redes é feito utilizando algoritmo de retropropagação ou Backpropagation, um algoritmo supervisionado que utiliza pares de entrada e saída para, por meio de correção dos erros, ajustar os pesos da rede. Esse ajuste é feito direcionando o vetor

de pesos na direção contrária ao gradiente do erro, que é calculado sendo realimentado para as camadas intermediárias. A Figura 17 ilustra o comportamento da rede.

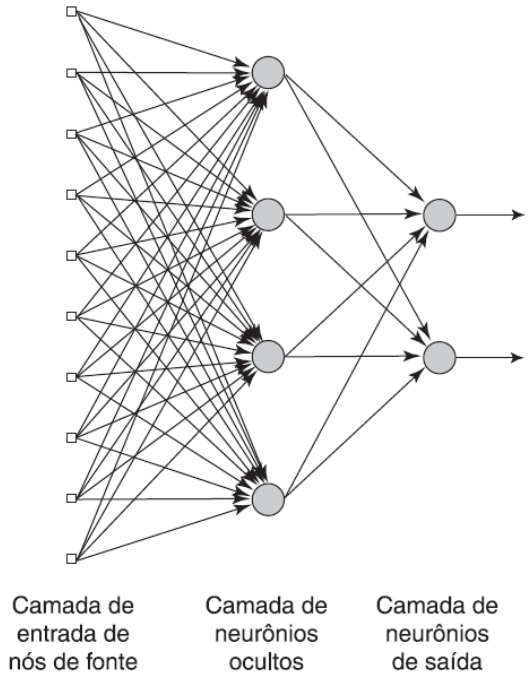

Figura 17: Modelo perceptron de múltiplas camadas.

O número de camadas da rede define a potencialidade de seu processamento, de modo que as redes multicamadas possuem potencial computacional muito maior que perceptron de uma única camada.

4.3. Teoria da Informação

A teoria da informação é um ramo da teoria da matemática estatística que analisa o armazenamento e manipulação da informação, Desta maneira, auxilia na compreensão de sistemas de telecomunicação, tomada de decisões, criptografia, codificação, ruído, correção de erros, transmissão de dados, etc. O princípio básico da teoria da informação consiste na ideia de que o conteúdo de uma mensagem sempre pode ser quantificado. Em geral, este conteúdo é medido em bits, entretanto, existem outras unidades de medidas que podem ser utilizadas.

O estudo desta teoria foi se desenvolvendo ao longo dos anos através de conceitos como o código Morse por Samuel Morse (1832-1838) e, posteriormente, as tentativas de definir uma medida de informação pelos teóricos da comunicação H. Nyquist (1924-1928), Hartley (1928) e pelo estatístico R. Fisher (1925). Contudo, o marco mundial referente a este assunto foi o artigo "The Mathematical Theory of Communication" escrito por Shannon (1948), considerado o pai da teoria da informação.

A teoria da comunicação trata de três principais conceitos, a medida da informação, a capacidade de um canal de comunicações transferir informação e a codificação, como meio de utilizar os canais com toda a sua capacidade. Além disso, diferencia a ideia de forma e conteúdo, sendo a primeira o transporte de uma matéria e a segunda a natureza do que é transportado.

Os conceitos básicos dessa teoria estão relacionados ao Teorema Fundamental da Informação (de Shannon):

> "Dada uma fonte de informação e um canal de comunicação, existe uma técnica de codificação tal que a informação pode ser transmitida por meio de canal e com uma frequência de erros arbitrariamente pequena apesar da presença do ruído.".

No artigo de Shannon, pela primeira vez, um modelo quantitativo e qualitativo apresentado como um processo estatístico, vindo a tona, então, considerações como a entropia da informação e redundância de uma fonte, informação mútua e capacidade de um canal com ruído, a lei de Shannon-Hartley para a capacidade de um canal Gaussiano e o bit, uma nova forma de enxergar a unidade fundamental da informação.

4.3.1. Medida da informação

LATHI (1998) defende que a quantidade de informação recebida está diretamente ligada à incerteza ou inversamente conectada com a probabilidade de sua ocorrência. Considerando P como a probabilidade de ocorrência de uma mensagem e I a informação obtida através dessa mensagem, é possível observar como quantificar I através da equação (5):

$$I = \log_2 \frac{1}{P} \text{ bits} \tag{5}$$

4.3.2. Entropia

No processo de desenvolvimento da teoria da informação Shannon definiu uma medida, a ser usada na melhoria de sistemas de telecomunicações, denominada entropia, em outras palavras, é uma medida de quantidade de informação de uma sequência de valores sucessivos, não havendo dependência entre os elementos. Sendo assim, a entropia é calculada utilizando a seguinte fórmula, representada pela equação (6):

$$H(X) = \sum_{j=1}^{M} P_j I_j = -\sum_{j=1}^{M} P_j \log_2 P_j \quad \text{bits/símbolo} \tag{6}$$

onde X é a média ponderada das autoinformações de cada símbolo e M são os símbolos.

4.3.3. Entropia condicional

Admiti-se que há dois acontecimentos, X e Y, com M e N possibilidades, respectivamente. A entropia condicional provê a entropia sobre um valor X uma vez que o valor de uma variável Y é conhecida. Partindo desse princípio, a entropia condicional é denotada por $H(X|Y)$ sobre a distribuição das probabilidades, conforme a equação (7):

$$H(X|Y) = -\sum_{i=1}^{M} \sum_{j=1}^{N} (Px_i) P(y_i|x_i) \log P(y_i|x_i). \tag{7}$$

4.3.4. Informação mútua

Como visto, a entropia $H(X)$ representa a incerteza sobre a entrada do canal antes de observar a saída, em contra partida, a entropia condicional, simula a incerteza referente à entrada do canal após a observação da saída. Embasado nessas definições obtém-se o conceito de informação mútua, que mede a redução da incerteza, definida pela equação (8):

$$H(X) - H(X|Y) \tag{8}$$

4.3.5. Índice de Shannon

É um dos diversos índices de biodiversidade, trata as espécies como símbolos e os

tamanhos da respectiva população como uma probabilidade. Uma de suas vantagens é que considera o peso entre as espécies raras e abundantes igualmente (MAGURRAN, 1988). Abaixo, pela equação (9), é possível observar a fórmula para calcular o índice de Shannon:

$$H^{'} = -\sum_{i=1}^{S} p_i Ln(p_i) \tag{9}$$

onde:

- H' é o índice de Shannon;
- p_i é a abundância relativa de cada espécie, calculada pela razão dos indivíduos de uma espécie (n_i) pelo número total dos indivíduos de uma comunidade (N). $\left({}^{n_i}/_{N}\right)$.
- S é o número total de espécies amostradas.

4.4. Bootstrap

4.4.1. Conceitos em reamostragem

Reamostragem é uma abordagem que pode ser paramétrica ou não paramétrica para o cálculo de distribuições empíricas, isto é, calcula a real distribuição estatística ao longo de centenas de amostras. Sendo assim, nesta abordagem, não se faz uso da distribuição de probabilidade. Uma vez que esses cálculos são realizados é possível calcular teste de normalidade dos valores, construir intervalos de confiança e testar hipóteses. Em outras palavras, defini-se reamostragem como um conjunto de métodos que se baseiam em calcular estimativas a partir de repetidas amostragens dentro de uma mesma amostra.

Para que seja possível utilizar tal ferramenta, é necessário criar múltiplas amostras a partir da amostra original e, para tanto, utilizam-se métodos de reamostragem.

4.4.2. Métodos de reamostragem

A principal diferença entre os métodos de reamostragem é relacionada à extração das amostras, que podem ser com ou sem reposição. Em ambos os casos é obtida uma amostra a partir da original para observação, contudo, quando há reposição a amostra selecionada é devolvida para que possa ser utilizada novamente, enquanto, na sem reposição, depois de observada, descarta-se a amostra.

Existem diferentes tipos de reamostragem, dentre eles destacam-se os *Testes de Aleatorização (Teste de Permutação), Validação Cruzada, Jackknife* e *Bootstrap*. Para a realização deste trabalho optou-se por utilizar o último citado.

4.4.3. Método bootstrap

Introduzido nos anos 70 por Efron, o bootstrap é uma técnica de reamostragem que possibilita a distribuição de uma função das observações por uma distribuição empírica embasa em amostras de tamanho finita e é realizada com reposição. Um dos benefícios de utilizar a esta técnica é por não ser necessário muitas suposições para estimar os parâmetros das distribuições em questão.

Para utilizar o bootstrap é preciso realizar um grande número de reamostragens, sendo assim, o método exige auxílio de programas computacionais para realizar as reamostras e os cálculos de modo mais rápido e com melhores resultados. Segundo Davison e Hinkley (1997), a repetição de um processo de análise original com muitas réplicas de dados podem ser denominado método intensivo computadorizado, o que justifica a necessidade de programas computacionais para aplicação do método.

4.4.4. Obtendo uma amostra bootstrap

Seja uma amostra original e a estatística de interesse abaixo:

1º. São geradas amostras bootstrap $x_{((1))}, x_{((2))}, \ldots, x_{((n*.))}$ com reposição de x.
2º. Deve-se calcular a estimativa da estatística de interesse:

$$\hat{\theta}_{(b)} = F[x_b], b = 1, \ldots, B \tag{10}$$

3º. É, então, calculado o erro padrão bootstrap, dado por:

$$\hat{S}_{boot} = \frac{1}{B-1}\sqrt{\sum_{b=1}^{B}\left[\hat{\theta}_b - \hat{\theta}_{(*)}\right]^2} \tag{11}$$

sendo,

$$\hat{\theta}_{(*)} = \frac{\sum_{b=1}^{B}\theta_{(b)}}{B} \tag{12}$$

4.4.5. Intervalo de confiança

Ao calcular a estimativa de uma média populacional, por exemplo, deseja-se encontrar um valor específico a ser usado para aproximar um parâmetro populacional. O intervalo de confiança possui característica similar, entretanto, trata-se de um intervalo de valores que tem a probabilidade de conter o verdadeiro valor da população. Para essa probabilidade dá-se o nome de grau de confiança, ou nível de confiança ou, ainda, coeficiente de confiança $(1-\alpha)$. Através da figura 18 é possível observar a ideia de um intervalo de confiança.

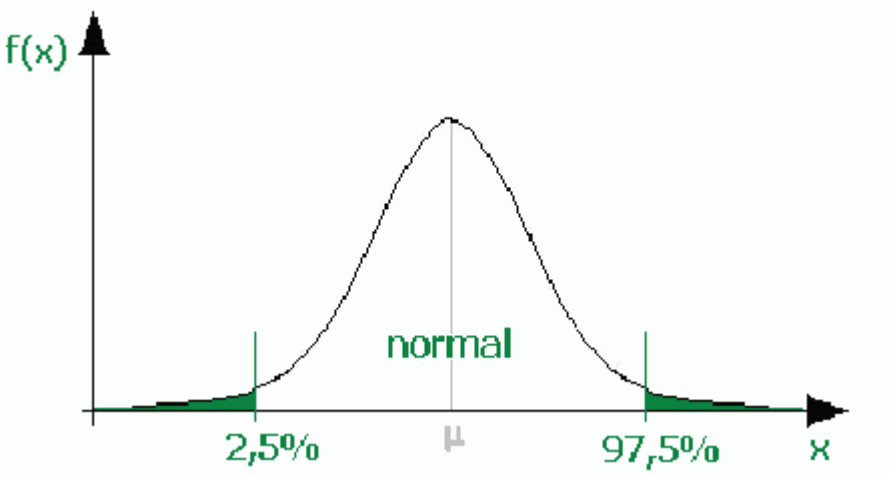

Figura 18: Distribuição normal e representação do intervalo de confiança da amostra. (Fonte: www.vestcon.com.br)

HALL (1988) destaca que se for feita uma comparação entre os intervalos de confiança, obtidos através da técnica Bootstrap e da distribuição de probabilidades conhecida do estimador, a diferença pode ser bem elevada, se as suposições necessárias para o segundo método forem inadequadas.

Existem diferentes formas de se calcular os intervalos de confiança pelo método Bootstrap, dentre eles destacam-se o método *t de Student,* o método *Percentil corrigido* e o método de C*orreção de Vício Acelerado*.

Neste trabalho, a estimativa da média e do intervalo de confiança foram baseados na distribuição de t-de Student, criada por William Gosset (1876-1937). Este método é utilizado para pequenas amostras e é, essencialmente, uma distribuição normal para todas as amostras de tamanho n. A figura 19, ilustra distribuição normal padronizada e quatro distribuições t de Student.

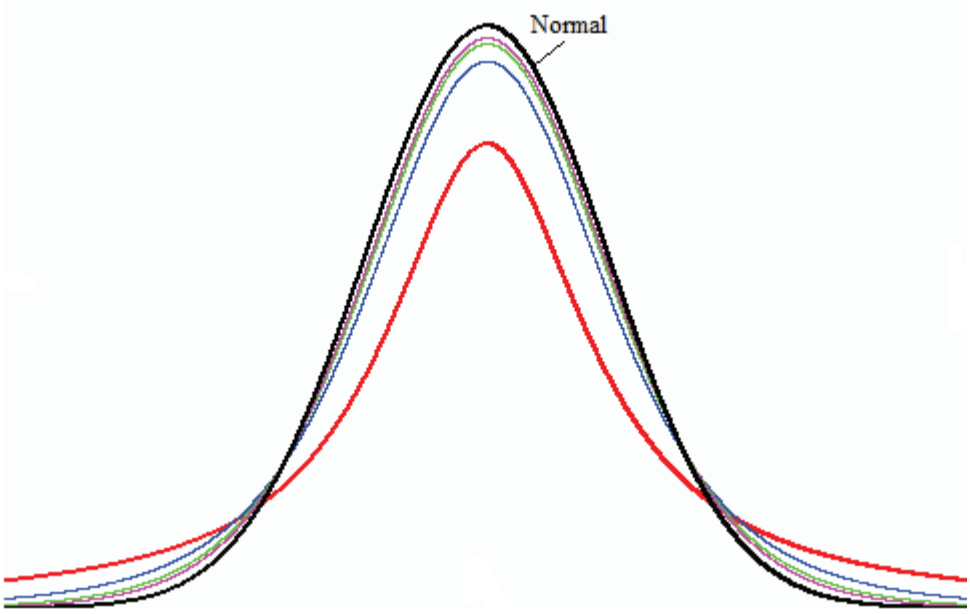

Figura 19: Modelo de distribuição normal e de t de Student (adaptada de http://www.eecis.udel.edu).

Através da distribuição, os valores críticos ($t_{\alpha/2}$) do intervalo de confiança são determinados. Onde obtém-se a fórmula 13 para o calculo da estimativa através da distribuição t de Student.

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \tag{13}$$

Com $\bar{x}$ sendo a média amostral observada e μ a média populacional (verdadeira). Sendo assim, o intervalo de confiança fica determinado conforme fórmula 14.

$$\bar{x} - E < \mu < \bar{x} + E \tag{14}$$

A figura 20 demonstra o intervalo de confiança em uma distribuição normal e na distribuição t de Student, utilizada no trabalho, uma vez que se adéqua melhor as amostras, tanto por quantidade como por alguns parâmetros necessários não serem conhecidos para uma distribuição normal.

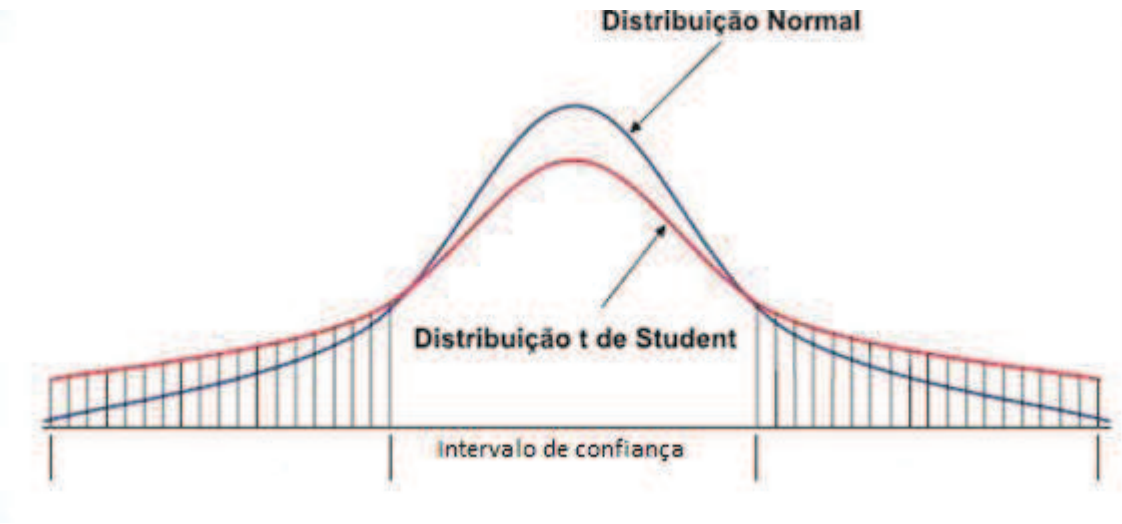

Figura 20: Intervalo de confiança aplicado na distribuição normal e t de Student.

5. Materiais e Métodos

5.1. Base de Dados e sua Triagem

O presente trabalho faz uso de dados fornecidos pelo Laboratório de Virologia Molecular da Universidade Federal do Rio de Janeiro pertencente à rede de genotipagem do Ministério da Saúde. Inicialmente a base possuía 923 dados, contendo a série de protease de cada paciente, com suas respectivas mutações, as taxas de carga viral e CD4+ dos dois últimos tratamentos, o ano em que foi diagnosticado pela primeira vez, a idade, os regimes submetidos e o tempo de duração de cada regime antes de ocorrer a falha terapêutica.

As Figuras 21-a e 21-b representam o pré-processamento realizado nos dados para a criação do modelo.

Figura 21-a: Primeira parte do fluxograma do pré-processamento.

Figura 21-b: Segunda parte do fluxograma do pré-processamento.

O fluxograma está divido de acordo com as duas etapas que consistiram o trabalho, sendo a primeira a análise dos dados e a segunda a criação do modelo.

Após o pré-processamento (eliminação de viés e divisão dos subtipos virais do HIV-1) foi realizada uma análise exploratória dos dados visando avaliar as seguintes varáveis: número de pacientes, de cada subtipo, gênero, idade e as taxas de carga viral e CD4. O resultado dessa análise pode ser observado na Tabela 2, que relaciona a frequência e o percentual de cada variável com os respectivos subtipos.

Tabela 2: Relação de dados ao fim da primeira análise.

	Subtipo	
	B	**C**
	n (%)	*n (%)*
Frequência (Total de Indivíduos)	555 (60,13)	40 (4,33)
Gênero (Feminino)	183 (32,97)	19 (47,50)
Média de CD4 no último tratamento	318,19	671,53
Média de CV no último tratamento	4,55	4,66
Tempo médio de tratamento em meses	32,12	38,02

A partir desse momento, a análise concentrou-se nos regimes do pacientes. Foram feitos dois cortes iniciais, pacientes com um e com dois regimes terapêuticos. Em cada um desses grupos foi avaliado o tipo de medicamento tomado, classificados em IP, NRTI, NRTI + IP, NRTI + NNRTI, NNRTI + IP, NRTI + NNRTI + IP, tanto para o subtipo B quanto para o subtipo C. A Tabela 3 expõe os resultados obtidos.

Tabela 3: Distribuição de frequências quanto aos regimes.

	Último Regime			
Regimes Terapêuticos	**Subtipo B**		**Subtipo C**	
	1 Regime	**2 Regimes**	**1 Regime**	**2 Regimes**
IP	-	1	-	-
NRTI	12	8	4	-
NRTI + IP	59	232	7	16
NNRTI + IP	-	4	-	-
NRTI + NNRTI	43	151	5	7
NRTI + NNRTI + IP	-	45	-	1

O passo seguinte consistiu em avaliar estatisticamente cada um dos dados considerados importantes, como idade, carga viral, CD4, mutações e tempo do regime. Embasado nas análises realizadas até esta etapa foram realizados novos agrupamentos quanto às terapias e feitas comparações entres os resultados encontrados para os subtipos B e C, que constataram que o subtipo B tem um índice de mutações muito maior, em contrapartida o subtipo C é bem mais agressivo, tendo falhas terapêuticas em menor tempo, resultados já validados pela literatura.

Através das análises realizadas foi possível observar alguns fatos interessantes para o trabalho, com base nesses resultados, foi realizado um corte em pacientes que não tivessem sido experimentados a IP até o último tratamento e paciente que não foram experimentados a nenhum inibidor de protease. Dentre os inibidores de protease foi escolhido o Nelfinavir (NFV) para o modelo a ser criado.

A escolha desse inibidor foi feita devido a características específicas do nelfinavir, como o fato de não ser conformulado com algum outro medicamento e, principalmente, por possuir posições de mutações bem definidas. A ideia de reduzir todas as possíveis combinações de medicamentos ao uso do NFV foi embasada no desejo de eliminar viés para um melhor estudo do modelo, além do tempo disponível para realização deste trabalho. A partir das observações realizadas neste trabalho é possível a obtenção do modelo levando em consideração outras saídas.

5.2. Regra de corte

Em cada um dos subtipos foi aplicada uma regra de corte na sequência dos aminoácidos da protease em três etapas:

- 1º corte (2,5%) – Foram descartadas as posições com número inferior a 2,5% resíduos de mutação, com relação ao total da amostra.
- 2º corte (5%) – Foram descartadas as posições com número inferior a 5% resíduos de mutação, com relação ao total da amostra.
- 3º corte (7,5%) – Foram descartadas as posições com número inferior a 7,5% resíduos de mutação, com relação ao total da amostra.

Após os cortes foi realizada uma análise em cada corte a fim de verificar as posições definidas como clássicas pela literatura, com o intuito de optar pela regra de corte que não houvesse perda significativa dessas posições. Sendo assim, foi utilizado o corte de 5% para ambos os subtipos.

5.3. Normalização

Uma vez que os dados necessários para o modelo foram selecionados, foi feita uma normalização das taxas de carga viral e CD4, com o intuito de alterar os valores reais de entrada para o intervalo entre 0 e 1, de modo que todos os atributos possuam valores semelhantes, sendo assim, todos os dados possuem a mesma importância, aumentando a compatibilidade dos dados com as funções de transferência (MACHADO, 2005; NETO; PELLI, 2004). Para tanto, foi realizada uma transformação linear do atributo (Equação 7).

$$x_j^{norm} = \frac{x_j - x^{min}}{x^{max} - x^{min}} \tag{13}$$

- onde x_j^{norm} é a variável normalizada.
- x_j é a variável na posição j.
- x^{min} é o valor mínimo dentre as variáveis.

- x^{max} é o valor máximo dentre as variáveis.

5.4. Codificação

A sequência de protease, onde são fornecidas as mutações são representadas por letras, onde cada uma representa um aminoácido. Novamente, com o ideal de melhorar o processamento do modelo, foi feita uma transformação na sequência, fazendo com que se tornasse uma matriz numérica, isto é, foi realizada uma codificação usando a escala de hidrofobicidade elaborada por Kyte e Doolittle (1982), também conhecida como escala KD, que foi usada por Weinert e Lopes (2003) para codificar os aminoácidos da sequência de protease. A Tabela 4 apresenta a escala KD e seus valores reais.

Tabela 4: Valores reais atribuídos a cada aminoácido conforme a escala de hidrofobicidade.

Aminoácido	Escala KD	Valor Real	Categoria
I	+4,5	0,05	Hidrofóbico
V	+4,2	0,10	Hidrofóbico
L	+3,8	0,15	Hidrofóbico
F	+2,8	0,20	Hidrofóbico
C	+2,5	0,25	Hidrofóbico
M	+1,9	0,30	Hidrofóbico
A	+1,8	0,35	Hidrofóbico
G	-0,4	0,40	Neutro
T	-0,7	0,45	Neutro
S	-0,8	0,50	Neutro
W	-0,9	0,55	Neutro
Y	-1,3	0,60	Neutro
P	-1,6	0,65	Neutro
H	-3,2	0,70	Hidrofílico
Q	-3,5	0,75	Hidrofílico
N	-3,5	0,80	Hidrofílico
E	-3,5	0,85	Hidrofílico
D	-3,5	0,90	Hidrofílico
K	-3,9	0,95	Hidrofílico
R	-4,5	1,00	Hidrofílico

5.5. Aplicação do bootstrap

Como os conjuntos de dados a serem trabalhados são, consideravelmente, pequenos, principalmente para o subtipo C, fez-se necessário utilizar técnica de reamostragem, para tanto, foi selecionada a técnica conhecida como bootstrap, entre outros motivos, por ser um método não paramétrico. A aplicação do bootstrap nos conjuntos de dados foi feita o software R.

O primeiro passo para aplicar o bootstrap foi calcular o índice de Shannon, desta

forma, cada paciente foi representado pelo seu próprio índice, isto é, por um único número, que passará a ser usado durante todo o processo de reamostragem.

Uma vez que os índices foram determinados, foi necessário estudar esses valores verificando se havia harmonia entre eles, ou seja, se houve uma distribuição normal, para tanto foi gerado um histograma e uma distribuição normal. Depois, simulou-se uma amostra do conjunto e, então, o intervalo de confiança da diversidade média foi calculado por parcela. O valor dessa amostra deve ser determinado na hora, nunca podendo ultrapassar a quantidade total de dados da amostra original.

Esses cálculos basearam-se no teste de hipóteses (teste t), que avalia a validade de uma informação sobre determinada característica da população baseada nos dados de uma amostra.

Para o subtipo B foram realizadas reamostragens até que o conjunto passasse a ter 1000 dados. Já para o subtipo C, o objetivo foi alcançar 500 dados, apenas. A diferença da quantidade de elementos para cada um dos subtipos está relacionada a mesma divergência que ocorre nos conjuntos originais, onde a quantidade de pacientes do B é, aproximadamente, quatro vezes maior que a do C, portanto, reamostrar o segundo subtipo na mesma contagem do B, não faria sentido, além de gerar um grupo muito repetitivo, podendo não registrar um resultado honesto no modelo.

5.6. Estrutura da rede

A implementação do modelo proposto também foi realizada utilizando o software R, usando os dados do subtipo B e C após a aplicação do bootstrap, sendo 20% de cada um dos conjuntos para teste. Em um primeiro momento foram realizadas diferentes implementações, variando sempre os dados de entrada, selecionando posições específicas da sequência de protease para observar quais seriam os melhores resultados, após os estudos, foi observado que quando todas as posições de mutação eram inseridas como dados de entrada juntamente com a CV e a CD4+ os resultados eram mais satisfatórios. Desse modo, foi decidido trabalhar com essa situação em ambos os subtipos.

O modelo foi criado a partir de rede de perceptron de múltiplas camadas (MLP), tendo como função de ativação a função logística. Além disso, é composto por quatro camadas, uma de entrada, duas escondidas, contendo quatro e dois neurônios, respectivamente, e uma de saída com dois neurônios. A figura 22 a seguir ilustra o modelo usado.

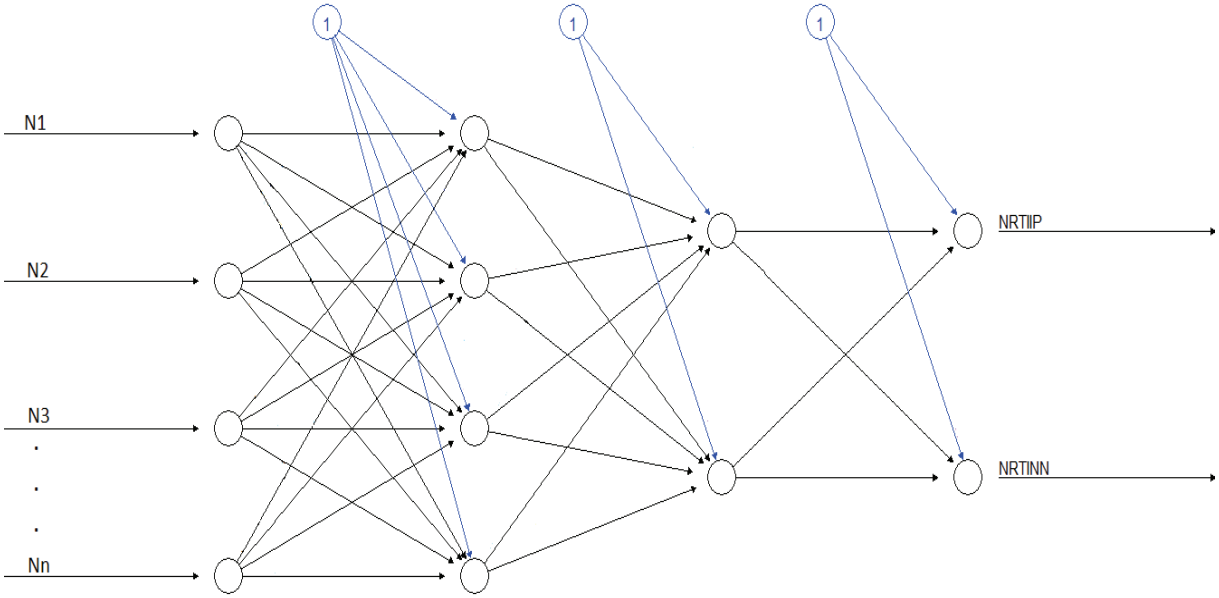

Figura 22: Modelo neural utilizado no trabalho.

Onde, N1, ... Nn são os dados de entrada, as linhas de conexão no sentido horizontal são os pesos sinápticos atribuídos pela rede e os números "1" circunscritos são os bias. Na camada de entrada encontra-se a única diferença entre os modelos usados para cada um dos subtipos, isto é, cada um dos modelos possui uma determinada quantidade de dados a serem introduzidos no modelo.

Essa quantidade específica dos subtipos foi definida após o estudo dos melhores resultados da rede com diferentes combinações de dados de entrada. Constatou-se, a partir daí, que a rede apresentava melhores resultados quando consideradas todas as posições da sequência de protease depois dos cortes. Sendo assim, para o subtipo B foram considerados 25 dados de entrada e para o C, 20 dados, ou seja, o B vai de N1 até N25, enquanto o C tem de N1 a N20.

A saída da rede fornece um resultado que pode ser compreendido como 1 e -1, no qual o primeiro determina que o paciente deve aderir a um a terapia antirretroviral unindo NRTI com Nelfinavir, de acordo com suas posições genômicas. No caso de direcionar para o -1, o modelo está determinando que o paciente seja medicado com terapias que não façam uso do inibidor de protease, restringindo-se a terapias com NRTI ou NRTI+NNRTI.

5.7. Critérios de avaliação

Para verificar o desempenho da rede em cada uma das simulações foram observadas três medidas, especificidade, sensibilidade e acurácia. O primeiro parâmetro mede a capacidade da rede em registrar corretamente os pacientes que não devem ser tratado com Nelfinavir. A sensibilidade mede a capacidade da rede em direcionar os pacientes a ser tratado com o inibidor de protease, neste caso, observa-se o quão sensível é o modelo. A acurácia é obtida através da soma da especificidade com a sensibilidade dividida pelo total de casos do conjunto teste.

Sendo assim, a rede pode retornar quatro saídas, na qual duas são corretas e duas

erradas. As corretas são representadas pelos verdadeiros positivos, relacionados a sensibilidade, e os falsos negativos, responsáveis pela especificidade, isto é, tanto os resultados da rede quanto os reais são iguais (os dois indicam o nelfinavir, é verdadeiro positivo, se ambos indicam a saída oposta, então é falso negativo). Quando o resultado da rede diverge do real, obtém-se o falso positivo ou verdadeiro negativo. As fórmulas a seguir representam os cálculos referentes a cada uma das medidas usadas:

- Especificidade:

$$E = \frac{VN}{(VN + FP)} \tag{14}$$

- Sensibilidade:

$$S = \frac{VP}{(VP + FN)} \tag{15}$$

- Acurácia:

$$A = \frac{E + S}{Total} \tag{16}$$

onde, VP é verdadeiro positivo
VN é verdadeiro negativo
FP é falso positivo
FN é falso negativo.

6. Resultados / Discussões

Neste capítulo serão apresentados os resultados das simulações realizadas para os subtipos B e C. Para cada um dos subtipos foram efetuadas 50 simulações usando como entrada todas as posições da sequência de protease selecionadas no pré-processamento mais a carga viral e a CD4+.

6.1. Subtipo B

O banco de dados do subtipo B possuía 89 pacientes, com a aplicação do Bootstrap essa amostra passou a contar com 1000 dados. A figura 23 apresenta o histograma dos índices de Shannon para o subtipo B.

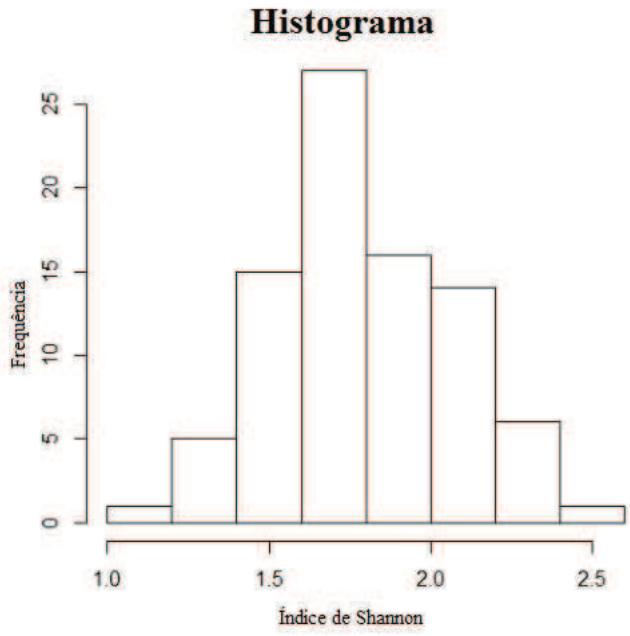

Figura 23: Relação entre os índices de Shannon e sua frequência no conjunto das 89 amostras originais do subtipo B.

O histograma acima representa distribuição dos índices de Shannon gerados durante o processo do bootstrap, onde cada amostra, isto é, cada paciente, passa a ser representado por esses índices, no qual para cada dado há um único valor numérico representativo.

Pela figura 24 é possível observar o gráfico quantil-quantil (Q-Q) da normal. Este

gráfico é uma ferramenta muito utilizada para conferir o ajuste da distribuição de frequência dos dados a uma distribuição de probabilidades. O gráfico Q-Q é usado, principalmente, para averiguar se os dados apresentam uma distribuição normal. Seu mecanismo consiste em observar se uma reta se ajusta aos pontos que representam os dados estudados.

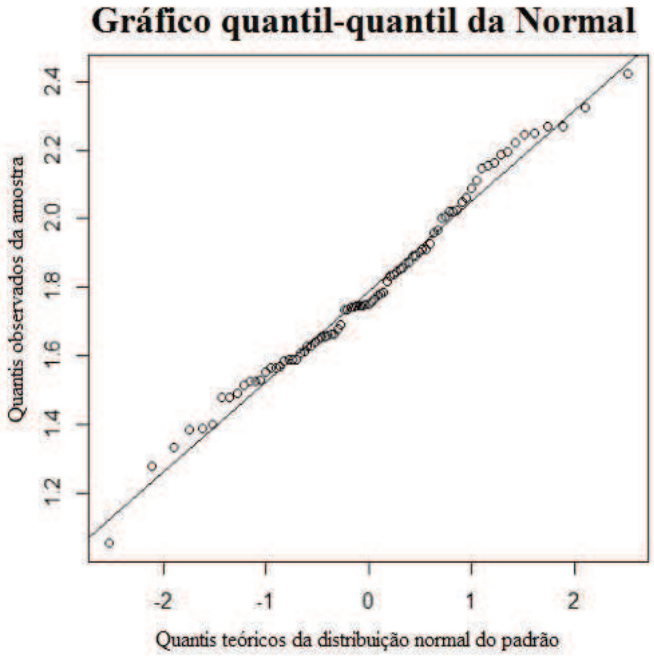

Figura 24: Gráfico Q-Q da distribuição normal das 89 amostras originais do subtipo B.

Conforme exposto acima, foram realizadas 50 simulações, com 25 dados de entrada, sendo eles as posições L10, T12, G17, L19, K20, D30, L33, E35, M36, M46, R57, Q61, I62, L63, I64, H69, I72, T74, V77, V82, N88, L90, I93 mais a carga viral e a CD4+, usando 200 pacientes para validação. A tabela 5 a seguir apresenta os resultados de cada uma das simulações.

Tabela 5: Resultados das 50 simulações do subtipo B, avaliando os parâmetros de medida de capacidade. (Continua)

Número da Simulação	Especificidade (%)	Sensibilidade (%)	Acurácia (%)
1	100,00	96,59	98,50
2	100,00	90,00	95,50
3	100,00	90,48	97,00
4	100,00	95,35	98,00
5	100,00	95,05	97,50
6	100,00	97,80	99,00
7	100,00	96,08	98,00
8	100,00	97,03	98,50

Tabela 5: Continuação.

Número da Simulação	Especificidade (%)	Sensibilidade (%)	Acurácia (%)
9	100,00	94,85	97,50
10	100,00	96,91	98,50
11	100,00	95,56	98,00
12	100,00	97,92	99,00
13	100,00	95,40	98,00
14	100,00	96,59	98,50
15	100,00	95,92	98,00
16	100,00	99,03	99,50
17	100,00	96,30	98,50
18	100,00	95,24	98,00
19	100,00	97,83	99,00
20	100,00	96,63	98,50
21	100,00	98,08	99,00
22	98,40	100,00	99,00
23	90,09	100,00	94,50
24	100,00	98,90	99,50
25	98,15	100,00	99,00
26	100,00	97,85	99,00
27	100,00	96,91	99,00
28	100,00	93,88	98,50
29	96,61	100,00	98,00
30	100,00	97,47	99,00
31	100,00	97,78	99,00
32	100,00	94,62	97,50
33	100,00	97,73	99,00
34	100,00	97,78	99,00
35	100,00	96,63	98,50
36	100,00	97,87	99,00
37	100,00	96,43	98,50
38	100,00	98,85	99,50
39	100,00	94,95	97,50
40	100,00	90,24	96,00
41	100,00	98,94	99,50
42	100,00	97,59	99,00
43	100,00	95,05	97,50
44	100,00	98,21	99,00
45	100,00	92,39	96,50
46	100,00	91,43	95,50
47	100,00	96,67	98,50
48	97,39	100,00	98,50
49	100,00	96,08	98,00
50	100,00	94,44	97,50

Observando os resultados demonstrados na tabela acima, é possível verificar que em 90% das simulações a rede obteve 100% para especificidade, enquanto a sensibilidade tem 10% de acerto total. Entretanto, mesmo com o baixo percentual de acerto, os valores de sensibilidade obtidos são bons, sendo menor 90%, com isso, os resultados da acurácia

também são satisfatórios, variando de 94,5 a 99,5 por cento, demonstrando assim um bom desempenho da rede para o subtipo B. A tabela 6 a seguir, resume os resultados das 50 simulações fazendo uso das medidas estatísticas de posição, média, mediana e moda.

Tabela 6: Resultados das 50 simulações do subtipo B agrupados em medidas estatísticas de posição.

	Média ± sd	Mediana	Moda	Intervalo [min.-máx.]
Especificidade	99,61 ± 1,53	100,00	100,00	[90,09 – 100,00]
Sensibilidade	96,47 ± 2,46	96,65	100,00	[90,00 – 100,00]
Acurácia	98,21 ± 1,09	98,50	99,00	[94,50 – 99,50]

Os três parâmetros usados para medir a capacidade da rede possuem seus ponto médios, medianos e modais em um intervalo bem próximo, em alguns casos chegam a ser iguais, demonstrando assim, que o número de simulações realizadas é satisfatório para as análises em questão. Além disso, é possível observar também que as três médias estão acima de 95%, ratificando a qualidade dos resultados e o desenvolvimento da rede.

6.2. Subtipo C

O banco de dados do subtipo C possuía 22 pacientes, com a aplicação do bootstrap essa amostra passou a ser de 500 dados. A figura 25 apresenta o histograma dos índices de Shannon para o subtipo C, referente as 22 amostras originais do conjunto.

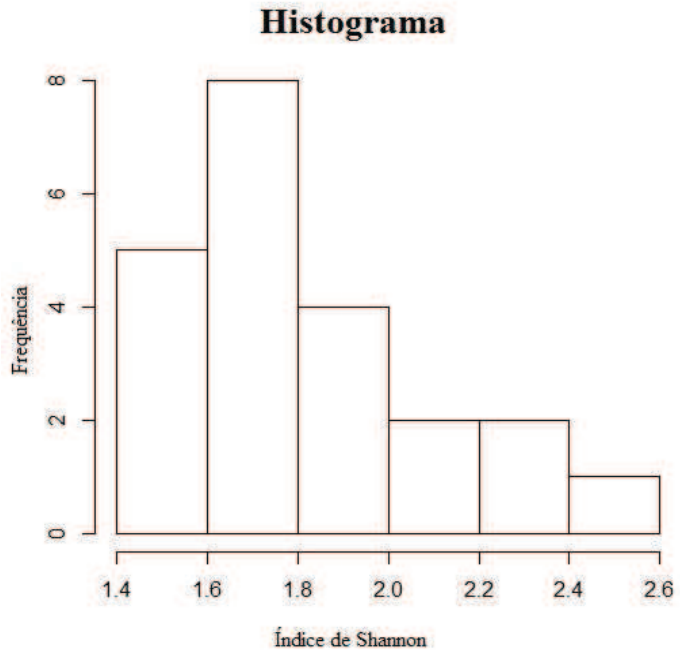

Figura 25: Relação entre os índices de Shannon e sua frequência no conjunto das 22 amostras originais do subtipo C.

Já a figura 26 possibilita a observação do gráfico Q-Q da normal.

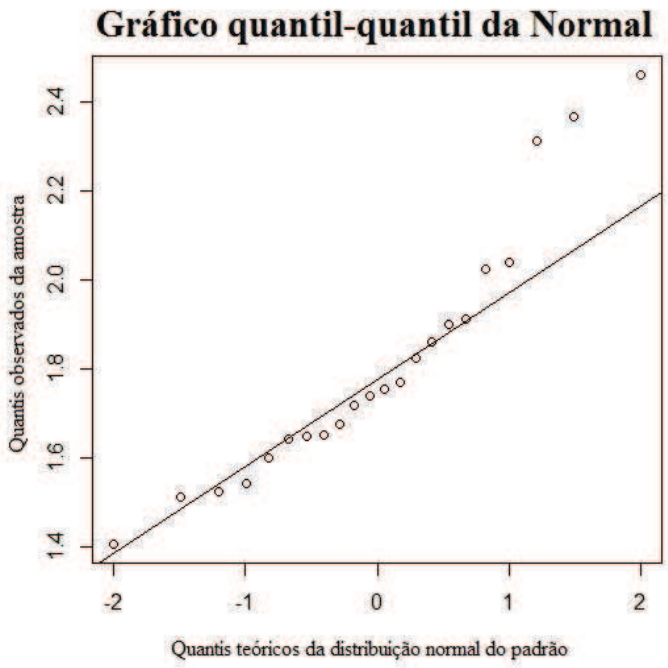

Figura 26: Gráfico Q-Q da distribuição normal das 22 amostras originais do subtipo C.

Novamente foram realizadas 50 simulações, neste caso com 20 dados de entrada, as posições T12, K14, I15, G16, K20, D30, E35, M36, R57, Q61, L63, A71, I72, T74, V82, N88, L89, L90 mais a carga viral e a CD4+, usando 100 pacientes para validação. A tabela 7 a seguir apresenta os resultados de cada uma das simulações.

Tabela 7: Resultados das 50 simulações do subtipo C, avaliando os parâmetros de medida de capacidade. (Continua)

Número da Simulação	Especificidade (%)	Sensibilidade (%)	Acurácia (%)
1	100,00	63,89	87,00
2	96,92	77,14	90,00
3	100,00	80,00	92,00
4	98,28	88,10	94,00
5	98,36	76,92	90,00
6	100,00	71,05	89,00
7	98,51	72,73	90,00
8	100,00	81,58	93,00
9	100,00	81,08	93,00
10	100,00	71,88	91,00
11	100,00	69,44	89,00
12	95,59	59,38	84,00
13	100,00	82,35	94,00
14	98,39	78,95	91,00
15	100,00	82,35	94,00
16	98,44	66,67	87,00
17	100,00	80,00	93,00
18	98,36	71,80	88,00
19	100,00	68,75	90,00
20	100,00	78,95	92,00
21	100,00	68,75	90,00
22	100,00	86,84	95,00
23	95,16	55,26	80,00
24	100,00	70,00	88,00
25	91,04	72,73	85,00
26	100,00	73,17	89,00
27	100,00	76,34	91,00
28	98,39	84,21	93,00
29	96,88	82,11	93,00
30	98,41	72,97	89,00
31	98,63	77,78	93,00
32	95,31	75,00	88,00
33	100,00	77,42	93,00

Tabela 7: Continuação.

Número da Simulação	Especificidade (%)	Sensibilidade (%)	Acurácia (%)
34	100,00	67,57	88,00
35	97,14	80,00	92,00
36	100,00	84,21	94,00
37	95,24	75,68	88,00
38	98,46	88,57	95,00
39	100,00	68,57	89,00
40	98,39	68,42	87,00
41	98,41	83,78	93,00
42	100,00	69,44	89,00
43	98,36	69,23	87,00
44	96,83	70,27	87,00
45	100,00	73,68	90,00
46	100,00	70,73	88,00
47	100,00	70,97	91,00
48	100,00	80,00	92,00
49	98,44	69,44	88,00
50	98,21	75,00	88,00

Avaliando os resultados obtidos na tabela acima, é possível verificar que em 52% dos casos a rede obteve 100% para especificidade, enquanto a sensibilidade não teve nenhum acerto total. Para esse subtipo os valores de sensibilidade obtidos não são tão bons quanto esperados, principalmente se levado em consideração os valores de especificidade, desta forma, por consequência, os números representantes da acurácia não chegam a 100%, variam entre 80 a 95 por cento. Mesmo com os baixos resultados de sensibilidade, o modelo demonstrou um bom desempenho, no contexto geral. A seguir é possível verificar a tabela 8 com as médias, medianas e modas das 50 simulações.

Tabela 8: Resultados das 50 simulações do subtipo C agrupados em medidas estatísticas de posição.

	Média $\pm$ sd	Mediana	Moda	Intervalo [min.-máx.]
Especificidade	98,72 $\pm$ 1,82	100,00	100,00	[91,04 – 100,00]
Sensibilidade	74,82 $\pm$ 7,06	74,34	80,00	[55,26 – 88,57]
Acurácia	90,08 $\pm$ 3,03	90,00	88,00 / 93,00	[80,00 – 95,00]

Assim como ocorreu para o subtipo B, a média, mediana e moda de cada uma das medidas de capacidade da rede se encontram em um intervalo próximo, igualando-se em alguns casos, demonstrando assim, que o número de simulações realizadas é conveniente para as análises em questão. Contudo, a análise das médias entre as três medidas possibilita perceber uma grande diferença entre as mesmas, em especial, para a sensibilidade.

6.3. Comparação entre os resultados obtidos pelo subtipo B e pelo subtipo C

Em ambos os modelos observa-se um melhor resultado quando se trata de especificidade. Um dos motivos desse efeito pode estar relacionado a uma melhor identificação da rede quando os pacientes não são tratados com nelfinavir.

Estudando a tabela 7 pode-se observar que a variação entre o menor e maior acerto da rede quanto à especificidade é quase quatro vezes menor que a mesma variação na sensibilidade, que já apresenta, aproximadamente, 24% a menos de acerto. Quando avaliado a tabela 5, observa-se que ocorre o mesmo padrão, isto é, a especificidade também é maior do que os outros parâmetros, entretanto, neste caso, a sensibilidade e acurácia não possuem valores percentuais tão mais baixos que o primeiro parâmetro, não chegando a 4% de diferença entre eles.

A tabela 9 relaciona as médias dos resultados dos subtipos B e C para uma análise comparativa entre eles. Entretanto, é importante ressaltar, que a comparação entre os dois está sendo feita apenas no quesito matemático, uma vez que há diferenças que podem estar contribuindo com os resultados, como o número de amostras de cada um dos subtipos.

Tabela 9: Comparação entre as médias de cada medida de capacidade dos subtipos.

	SUBTIPO B	**SUBTIPO C**	
	Média	**Média**	**Diferença entre as médias**
Especificidade	99,61	98,72	0,82
Sensibilidade	96,47	74,82	17,65
Acurácia	98,21	90,08	8,13

Ao comparar os resultados obtidos pelo modelo, para os dois subtipos, é evidente o maior percentual da acurácia do subtipo B, demonstrando assim que a rede forneceu um melhor resultado para este subtipo. Contudo, vale ressaltar que o conjunto de dados do subtipo C é quatro vezes menor que o subtipo B, ainda nas amostras originais, e, após a aplicação do bootstrap, essa diferença é duas vezes menor. Apesar dessa diferença nos conjuntos de dados, as acurácias dos dois subtipos não se diferenciam tanto, sendo de 8,13%. Conforme exposto, em ambas as simulações, a melhor média encontrada é referente a especificidade, este resultado pode ser compreendido através do estudo dos dados originais, isto é, antes das reamostrangens.

O número total de pacientes do subtipo B são 89, sendo 50 pacientes não submetidos ao Nelfinavir e 39 que são. Já no C, há 22 pacientes, com 14 deles não sendo usuários do inibidor de protease e, apenas, 8 experimentados a esta classe de inibidor. A partir desses dados já é possível notar maior demanda de pacientes não usuários de IP, entretanto, esta demanda pode se tornar ainda maior ao ser realizada a técnica do bootstrap, que quando reamostra tem maior probabilidade de selecionar um paciente que não tenha sido experimentado ao Nelfinavir, aumentando ainda mais a quantidade dessa característica no conjunto de dados. Justificando assim, o maior acerto da rede na especificidade, uma vez que a rede pôde treinar mais esses dados.

Outro fator a ser considerado na comparação entre os resultados de cada subtipo é a prevalência do subtipo B, isto é, este é um subtipo com maior ocorrência, fato este que faz com que seu estudo seja mais consolidado do que o C, que vem se destacando recentemente. Os médicos, por exemplo, já estão mais familiarizados com suas características e

medicamentos específicos, fazendo com que se torne mais comum o acerto do médico, conjunto real deste trabalho, influenciando em um melhor resultado para o subtipo B, conforme as tabelas de resultados.

Outro fator que chamou atenção neste trabalho está relacionado à análise do gráfico Q-Q de cada um dos subtipos estudados. Quando avaliado a representação do subtipo B é notável a adequação da reta aos pontos, ou seja, os pontos, praticamente, sobrepõem a reta e estão bem agrupados, de modo que poucos estão mais afastados, e ainda esses, não estão tão distantes. Deste modo, comprova-se uma distribuição harmônica.

Quando a representação gráfica do subtipo C quanto a sua distribuição normal é analisada, ocorre o oposto ao observado no B. Os pontos estão bem espaçados entre si e poucos se adéquam a reta normal. Esse comportamento pode está relacionado ao pouco número de dados originais, mas é algo que precisa ser estudado, pois pode estar influenciando a rede a não obter resultados melhores para o subtipo C e possibilitando melhores resultados para o B.

7. Conclusões

O presente trabalho teve como objetivo tratar a problemática do tratamento aos portadores do HIV-1. Para tanto foi elaborado um modelo a ser usado como ferramenta auxiliar na decisão dessas terapias antirretrovirais em combate ao HIV-1 considerando o histórico genômico de cada paciente infectado pelo vírus. Tal modelo usou como base apenas os pacientes com subtipos B e C e restringiu-se ao medicamento nelfinavir para validação do modelo.

Conforme dito, o modelo demonstrou melhor resultados quando incorporada todas as posições de mutações estabelecidas pela regra de corte, mesmo quando comparado à simulação usando apenas posições consideradas clássicas. Esse resultado sugere a existência de posições não clássicas que influenciam em um melhor posicionamento terapêutico, demonstrando assim, a necessidade de constante estudo quanto às características do vírus e das terapias de combate ao mesmo, uma vez que suas constantes mutações afetam sua estrutura genômica continuamente. Além disso, essas influências não esperadas podem dificultar ainda mais a decisão do profissional da saúde em orientar um paciente portador do HIV-1 quanto à terapia a ser diagnosticada.

De acordo com os resultados encontrados, pode-se concluir que a regra de corte realizada no pré-processamento foi eficaz, bem como todas as etapas realizadas durante o processo, uma vez que os resultados do modelo são considerados bons, tendo alta acurácia. Contudo, deve-se ressalvar que o número de amostras é pequeno, principalmente do subtipo C, e mesmo com aplicação do bootstrap com o ideal de aumentar esse número, a necessidade de mais amostras originais é fundamental para o aprimoramento do modelo.

A escolha de realizar o estudo baseado no inibidor de protease também ajudou nas simulações, uma vez que, como dito, possui características bem específicas, como posições bem definidas de mutação. Desta forma, o bom desempenho do modelo garante resultados tão bons ou melhores, para outros medicamentos de comportamento similar, ressalvando a importância de constantes estudos referentes aos medicamentos que compõem os regimes terapêuticos, a fim de encontrar soluções para melhorar a qualidade e expectativa de vida dos pacientes portadores do HIV-1.

8. Trabalhos Futuros

Uma das dificuldades encontradas durante a realização deste trabalho foi a baixa quantidade de dados, principalmente relacionado ao subtipo C, sendo assim, em trabalhos futuros visa-se inserir novas amostras de pacientes portadores do vírus, inclusive amostras mais recentes, e considerar novos medicamentos, de modo que o modelo alcance todas as combinações possíveis de terapias gradativamente. Além disso, há o interesse de estudar as outras sequências genômicas, como a da transcriptase reversa, uma vez que muitas das drogas inseridas na terapia antirretroviral agem nessa área.

Outro interesse é considerar outros subtipos, como o F, que já demonstrou ser um dos que prevalecem em algumas regiões do Brasil, e por ser pouco conhecido, não possui um tratamento especial, além de não se saber quais são as posições específicas para esse subtipo e quais os melhores medicamentos para ele, sendo então, uma importante contribuição para os estudos referentes ao HIV-1, principalmente, no Brasil. Além de ampliar os estudos referentes ao subtipo C, e, também, analisar o modelo em outras arquiteturas de redes, que utilizam a teoria da ressonância adaptativa (ART).

9. Referências Bibliográficas

ABRANTES, S. A., **Apontamentos de teoria da informação**. Capítulo 1. Departamento de Engenharia Electrotécnica e de Computadores. Universidade do Porto, 2003.

AGGARWAL, R. e SONG, Y., **Artificial Neural Networks in Power Systems** – Part 3. Power Engineering Journal, Hearts, Inglaterra, v. 12, n. 6, p. 279-287, dec./1998.

BALASUBRAMANIE, P.; FLORENCE, M. L. **Application of Radial Basis Network Model for HIV/AIDs Regimen Specifications**. Journal of Computing. Vol. 1, 2009.

BISWAS, P.; TAMBUSSI, G.; LAZZARIN, A. **Access denied? The status of coreceptor inhibition to counter HIV entry. Expert Opin Pharmacother, London**; v.8, n.7, p.923-33, 2007.

BRAGA, A. P.; CARVALHO, A. C. P. L. F.; LUDEMIR, T. B. **Redes Neurais Artificiais - Teoria e Aplicações**. Ed. 2, p. 225, 2012.

CARDOSO, L. P. V., **Genotipagem para resistência primária e secundária em pacientes infectados pelo HIV-1 do Estado de Goiás**. 2009. 174f. Tese (Doutorado em Medicina Tropical e Saúde Pública) – Instituto de Patologia Tropical e Saúde Pública, Universida Federal de Goiás, Goiás. 2009.

CARVALHO, G.S., **Pessoas vivendo com HIV/AIDS: Vivências do tratamento anti-retroviral**.2008. 98f. Dissertação (Mestrado em Saúde Coletiva) – Centro de Ciências da Saúde, Universidade Estadual de Londrina, Londrina. 2008.

COSTA, G. G.O., **Um Procedimento Inferencial para Análise Fatorial Utilizando as Técnicas Bootstrap e Jackknife: Construção de Intervalos de Confiança e Testes de Hipóteses.** 2006. 196f. Tese (Doutorado em Engenharia elétrica) – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro. 2006.

COSTA, J. C. G. D., **Investigação de técnicas de simulação na modelagem de rede neural artificial aplicada a transplante renal**. 2005. 126f. Dissertação (Mestrado em Engenharia Biomédica) – COPPE/UFRJ, Universidade Federal do Rio de Janeiro, Rio de Janeiro. 2005.

DAVISON, A. C.; HINKLEY, D. V. **Bootstrap methods and their application.** Cambridge: Cambridge University Press, 1997.

DIAS, R., 2004. ***— Guia para o manuseio de testes de resistência antiretroviral no paciente infectado pelo HIV-1***‖, ABBOTT Laboratórios do Brasil.

EFRON, B. e TIBSHIRANI, **Na Introduction to the Bootstrap**, Chapman & Hall, 1993.

FERREIRA, A.S. **Resistência primária aos antirretrovirais e mapeamento genético do HIV-1 no Estado do Mato Grosso**. 99 p. Dissertação (Mestrado em Ciências da Saúde).- Universidade Federal de Goias, Goiânia. 2011

HAYKIN, S. S., **Redes Neurais:** Princípios e prática. 2ª edição. Brasil: Bookman Companhia, 2001, 900p.

HALL, P. **Theoretical comparison of bootstrap confidence intervals**. Annals of Statistics, v. 16, n. 3, p. 927–953, Sep. 1988

HAUBRICH, R.H. **Resistance and replication capacity assays: clinical utility and interpretation**. Topics in HIV Medicine, 12: 52-56, 2004.

HEATON, J. **Programming Neural Networks with Encog 2 in Java**. p. 481, 2010.

HIRSCH, M.S. *et al*, **Antiretroviral drug resistance testing in adult HIV-1 infection: recommendations of an International AIDS Society-USA Panel**. The Journal of The American Medical Association, 283: 2417-2426, 2000.

JOHNSON, V. A., *et al.*, **2014 Update of the drug resistance mutations HIV-1.** Topics in Antiviral Medicine – Estados Unidos, v. 22, Issue,3. Jun./Jul. 2014.

KALMAR, E. M. N., **Avaliação da resistência do HIV-1 às drogas anti-retrovirais em 150 pacientes em interrupção terapêutica por mais de seis meses**. 2007. 110f. Tese (Doutorado em Medicina). – Universidade de São Paulo, São Paulo, 2007.

KATZUNG, B. G. **Farmacologia: básica & clínica**. 9. ed. Rio de Janeiro: Guanabara Koogan, 2006.

LATHI, B. P., 1998. **Modern Digital and Analog Communication Systems**. Oxford.

MACÊDO, O., **Caracterização molecular da resistência genotípica secundária aos antirretrovirais em pacientes com AIDS e prevalência de subtipos do HIV-1 nos Estados do Pará e Amazonas, Brasil: 2002 a 2006**. 2010. 109f. Dissertação (Mestrado em Biologia de Agentes Infecciosos e Parasitários) – Instituto de Ciências Biológicas, Universidade Federal do Pará, Pará. 2010.

MACHADO, A. **Neuroanomia Fucional. Minas Gera**is: Atheneu LTDA, 2005.

MAGURRAN, A.E. 1988. **Ecological diversity and its measurement. New Jersey: Princeton University Press**, 179 p.

MINISTÉRIO DA SAÚDE DO BRASIL. **Recomendações para terapia antirretroviral em adultos e adolescentes infectados pelo HIV – 2013/2014.** (Programa Nacional de DST/AIDS), 2014.

MINISTÉRIO DA SAÚDE DST/AIDS. **Boletim Epidemiológico de AIDS/DST 2014.** Disponível em: http://www.aids.gov.br, acessado em outubro de 2014.

NERRAND, O., ROUSSEL-RAGOT, P., PERSONNAZ, L., DREYFUS, G. **Neural Networks and Nonlinear Adaptive Filtering**: Unifying Concepts and New Algorithms. Neural Computation, vol. 5, no. 2, pp. 165-199, 1993.

OLIVEROS, M. P. R., **Evolução das mutações de resistência aos inibidores de Protease em pacientes infectados pelo HIV-1 subtipo F**. 2009. 129f. Tese (Doutorado em Ciências). Faculdade de Medicina - Universidade de São Paulo, São Paulo – SP. 2009.

OSÓRIO F., BITTENCOURT J. R., 2000. **— Sistemas Inteligentes Baseados Em Rnas Aplicados Ao Processamento De Imagens‖**. In: Workshop De Inteligência Artificial, Santa Cruz Do Sul: Unisc.

PENG C. *et al.,* **Role of human immunodeficiency vírus type 1-specific protease in core protein maturation and viral infectivity**. J Virol 63:2550-2556. 1989.

PINTO, M. E. e STRUCHINER, C. J., **A diversidade do HIV-1: uma ferramenta para o estudo da epidemia**. Caderno Saúde Pública. Rio de Janeiro. p. 473. Mar./2006. Disponível em < http://www.scielo.br/>. Acesso em: 30 out. 2014.

PORTAL DA EDUCAÇÃO. **Terapia Antirretroviral.** Disponível em: http://www.portaleducacao.com.br/farmacia/artigos/7829/terapia-anti-retroviral#ixzz386Kl4hT2, acessado em outubro de 2014

SABINO, E. C., **Subtipos de HIV-1 no Brasil**. *Epidemiologia*: contextos e pluralidade [online]. Rio de Janeiro: Editora FIOCRUZ, 1998. 172 p. EpidemioLógica series, n°4. Disponível em < http://www.scielo.br/>. Acesso em: 15 out. 2014.

SHAFER, R. W. **Genotypic testing for human immunodeficiency virus type 1 drug resistance**. Clin. Microbiol. Rev., Washington; v.15, n.2, p.247-77, 2002.

SILVEIRA, A. A., **Mapeamento genético do HIV-1 e análise de resistências associadas aos antirretrovirais em pacientes do Centro-Oeste brasileiro**. 2011. 86f. Tese (Doutorado em Medicina Tropical e Saúde Pública) – Universidade Federal de Goiás. Goiás. 2011.

SOUZA, C.R. **Redes neurais sem-peso aplicadas na categorização de subtipos do HIV-1.** 101 p. Dissertação (Programa de Pós-graduação em Engenharia de Sistemas e Computação - Mestre em Engenharia de Sistemas e Computação). Universidade Federal do Rio De Janeiro (UFRJ), 2010. Rio de Janeiro - RJ.

STROHL, W. A.; ROUSE, H.; FISHER, B. D. **Microbiologia ilustrada**. Porto Alegre: Artmed, 2004

RAZVAN, A. *et al.*, **A new Fuzzy ARTMAP approach for predicting biological activity of potential HIV-1 protease inhibitors**. In: The IEEE International conference on Bioinformatics and Biomedicine 2007 – Silicon Valley, Estados Unidos. 2007.

RIZZO, A. L. T.; CYMROT, R. **Estudo e aplicações da técnica bootstrap.** In: II Jornada de iniciação científica. Universidade Presbiteriana Mackenzie, São Paulo. 2006.

RIZZO, A. L. T.; CYMROT, R. **Utilização da técnica de reamostragem bootstrap em aplicação na Engenharia de Produção.** Universidade Presbiteriana Mackenzie.In: X Encontro Latino Americano de Iniciação Científica e VI Encontro Latino Americano de Pós-graduação – Universidade do Vale, São Paulo. 2006. p. 488.

ROBBINS B. L., *et al.,***Anti-human immunideficiency vírus activity and cellular metabolism of a potential prodrug of the acyclic nucleoside phosphonate 9-R adenine, Bis PMPA**. Antimicrob Agents Chemother 42:612-617. 1998.

ROMEU, G. A., **Avaliação da adesão a terapia antirretroviral de paciente portadores de HIV**. Revista Brasileira de Farmácia Hospitalar e Serviços de Saúde, São Paulo, v. 3, n. 1, p. 37-41, jan./mar. 2012.

ROSSI, A. L. D. e BRUNETTO, M. A. O. C., **Métodos de codificação de proteínas para uso com redes neurais artificiais**. In: Congresso Brasileiro de Informática em Saúde, 7, 2006. Florianópolis. *Anais*...Florianópolis, 2006.

SILVA, M. M. G., **Características das gestantes infectadas pelo HIV, de acordo com o momento do seu diagnóstico**. 2007. 99f. Dissertação (Mestrado em Medicina: Epidemiologia) – Universidade Federal do Rio Grande do Sul, Rio Grande do Sul. 2007.

SILVA, R. M., **Identificação de mutações do HIV-1 em pacientes com falha terapêutica ao Nelfinavir usando o modelo computacional híbrido**. In: Congresso Brasileiro de Informática em Saúde, 12, 2010. Porto de Galinhas. *Anais*...Porto de Galinhas, 2010.

SILVA, R. M. **Algoritmo genético e kernel discriminante de Fisher aplicado a identificação de mutações de resistência do HIV-1 aos inibidores antirretrovirais da Protease.** 126 p. Tese (Programa de Pós-graduação em Engenharia Biomédica - Doutor em Engenharia Biomédica). Universidade Federal do Rio de Janeiro (UFRJ), 2009. Rio de Janeiro - RJ.

SIMON, D. *et al*, **Prevalência de subtipos do HIV-1 em amostra de pacientes de um centro urbano no Sul**. Revista Saúde Pública, São Paulo, v. 44, n. 6, out/2010. Disponível em < http://www.scielo.br/>. Acesso em: 23 nov. 2014.

SOUZA, C. R., **Redes Neurais sem-peso aplicadas na categorização de subtipos do HIV-1**. 2011. 91f. Dissertação (Mestrado em Engenharia de Sistemas e Computação) – Universidade Federal do Rio de Janeiro, Rio de Janeiro. 2011.

UNAIDS, 2014. **AIDS em números.** Disponível em: http://www.unaids.org.br/documentos/Aids%20by%20the%20numbersPORT.pdf, Acessado em 30 de setembro de 2014.

UNAIDS, 2014. **AIDS epidemic update.** Disponível em: http://www.unaids.org/en/HIV_data/Epidemiology/default.asp, Acessado em 30 de setembro de 2014.

VANDAMME, A., SONNERBORG, A., AIT-KHALED, M., *et al.*, 2004. **— Updated European recommendations for the clinical use of HIV drug resistance testing‖**, Antiviral Therapy, v.9(6), pp.829-848.

WILSON, J.W., BEAN, P. **A Physician's Primer to Antiretroviral Drug Resistance Testing**. The AIDS Reader, 10: 469-478, 2000.