COPPE/UFRJ

ALGORITMO GENÉTICO E KERNEL DISCRIMINANTE DE FISHER APLICADO A IDENTIFICAÇÃO DE MUTAÇÕES DE RESISTÊNCIA DO HIV-1 AOS INIBIDORES ANTIRETROVIRAIS DA PROTEASE.

Robson Mariano da Silva

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

Orientador(es): Flávio Fonseca Nobre

Rodrigo de Moraes Brindeiro

Rio de Janeiro

Janeiro 2009

ALGORITMO GENÉTICO E KERNEL DISCRIMINANTE DE FISHER APLICADO A IDENTIFICAÇÃO DE MUTAÇÕES DE RESISTÊNCIA DO HIV-1 AOS INIBIDORES ANTIRETROVIRAIS DA PROTEASE.

Robson Mariano da Silva

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

Aprovada por:

_____
Prof. Flávio Fonseca Nobre, Ph.D.

_____
Prof. Rodrigo de Moraes Brindeiro, D.Sc.

_____
Prof. Marcio Nogueira de Souza, D.Sc.

_____
Prof. Amílcar Tanuri, D.Sc.

_____
Prof. José Carlos Couto Fernandez, D.Sc.

_____
Prof. André Carlos Ponce de Leon Ferreira de Carvalho, Ph.D.

RIO DE JANEIRO, RJ - BRASIL
JANEIRO DE 2009

Mariano, Robson da Silva

Algoritmo Genético e Kernel Discriminante de Fisher Aplicado a Identificação de Mutações de Resistência do HIV-1 aos Inibidores Antiretrovirais da Protease. / Robson Mariano da Silva. – Rio de Janeiro: UFRJ/COPPE, 2009.

XIV, 112 p.: il.; 29,7 cm.

Orientadores: Flávio Fonseca Nobre

Rodrigo de Moraes Brindeiro

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia Biomédica, 2009.

Referencias Bibliográficas: p. 97-105.

1. Resistência HIV-1. 2. Algoritmo Genético. 3. kernel Discriminante de Fisher. I. Nobre, Flávio Fonseca, *et al*. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Biomédica. III. Titulo.

"A minha esposa, Sandra e aos meus filhos, Luís Felipe e João Vítor. Seus gestos de amor, carinho e afeto sustentam minha força e determinação na realização de sonhos."

Agradecimentos

A minha família Sandra, Luís Felipe e João Vítor, pelo incentivo e compreensão nos momentos de ausência.

Aos meus pais por tudo que representam em minha vida.

Ao Prof. Flávio Fonseca Nobre por seu apoio e orientação segura, tranqüila e sempre paciente.

Ao Prof. Rodrigo de Moraes Brindeiro por sua orientação segura na área de Virologia do HIV-1.

Aos amigos Marcelo Ribeiro-Alves (LESS/PEB) e Mônica B. Arruda (Laboratório de Virologia Molecular/UFRJ), por suas valiosas contribuições na elaboração dessa tese.

Aos amigos do LESS/PEB que tive o prazer de conviver nesse período, bem como aos amigos do LAVIMOAN/UFRJ.

Ao DEMAT (PICDT/UFRRJ) por possibilitar meu afastamento visando o aprimoramento do seu corpo docente.

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

ALGORITMO GENÉTICO E KERNEL DISCRIMINANTE DE FISHER APLICADO A IDENTIFICAÇÃO DE MUTAÇÕES DE RESISTÊNCIA DO HIV-1 AOS INIBIDORES ANTIRETROVIRAIS DA PROTEASE.

Robson Mariano da Silva

Janeiro/2009

Orientadores: Flávio Fonseca Nobre
Rodrigo de Moraes Brindeiro
Programa: Engenharia Biomédica

O acúmulo de mutações genéticas de resistência do HIV-1 às drogas antiretrovirais, em variantes virais, é uma causa importante no surgimento de falha terapêutica em pacientes sob tratamento. Nos últimos anos, diversas metodologias foram desenvolvidas visando avaliar fenotipicamente a resistência do HIV-1 aos medicamentos antiretrovirais, bem como avaliar genotipicamente o perfil mutacional do vírus. Entretanto muitas dessas metodologias têm limitações na interpretação de novas mutações, possivelmente associadas à resistência. O objetivo dessa tese é propor um modelo computacional baseado na utilização de algoritmo genético (AG) e no classificador de *kernel* de Fisher Discriminante (KDF), de modo a poder identificar possíveis novas mutações de resistência nos genes do HIV-1. O modelo aqui analisado é aquele do gene da aspartil-protease do HIV-1 de subtipos B e C, em pacientes com falha terapêutica e utilizando os inibidores de protease SQV, NFV e LPV. O conjunto de dados utilizado consiste de 1092 seqüências do gene da protease provenientes de isolados séricos de pacientes portadores do HIV-1, resistentes à terapia antiretroviral, obtidos junto ao Laboratório de Virologia Molecular (UFRJ, Brasil). Os resultados do modelo proposto mostraram-se promissores quanto ao AG e ao classificador KDF, na seleção de mutações de resistência.

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

GENETIC ALGORITHM AND KERNEL DISCRIMINANT APPLIED TO IDENTIFICATION OF MUTATION IN THE HIV-1 RESISTANCE TO THE ANTIRETROVIRAL PROTEASE INHIBITORS.

Robson Mariano da Silva

January/2009

Advisors: Flávio Fonseca Nobre
Rodrigo de Moraes Brindeiro
Department: Biomedical Engineering

Accumulation of HIV-1 resistance mutations to antiretroviral drugs, in resistant viral variants, has been an important cause of emergence of, ARV therapy failure in patients under treatment. In recent years, several methodologies were developed to phenotypically assess the HIV-1 resistance to antiretroviral drugs, as well as genotypically depict the mutational profiles of the resistant virus. However many of these methods have limitations in the elucidation of new mutations associated with resistance. The aim of this thesis is to propose a computational model (AG/KDF) based on the use of genetic algorithm (GA) and the classifier of Kernel Fisher Discriminant (KFD) in order to identify possible new resistance mutations in the therapy targets' genes of HIV-1. The model proposed herein was tested for the aspartyl-protease gene of HIV-1 from subtypes B and C, in order to predict the resistance mutation profile of isolates in patients under treatment failure and using the protease inhibitors SQV, NFV and LPV, currently used as parts of the ARV "cocktail" to control epidemic of AIDS in Brazil. Data set used consists of 1092 gene sequences of the HIV-1 protease from virus isolates of the Laboratory of Molecular Virology resistance cohort (UFRJ/BRAZIL). The results of the model proposed, were promising on the use of GA and application of the classifier KFD in the selection of possible mutations of resistance.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O vírus da imunodeficiência humana do tipo 1 (HIV-1) é um retrovírus pertencente à família *Retroviridae* do gênero *Lentivirus* e seu conteúdo genético está disposto em uma fita simples de RNA (disposta de forma diplóide -2n- na partícula viral, ou vírion). O HIV-1 caracteriza-se por sua enorme variabilidade genética e antigênica, apresentando uma taxa estimada de mutação de 1% ao ano, possibilitando assim que distintas variantes virais convivam no mesmo indivíduo infectado (a quasispécie), enquanto que subpopulações virais geneticamente agrupáveis e distintas entre os grupos, denominados de subtipos, estejam distribuídas em diferentes partes do mundo (MORGADO, 2000).

Dentre os subtipos do HIV-1, o subtipo B é o mais prevalente no Brasil e nos países do primeiro mundo, subtipo esse que acumula o maior número de informações genotípicas e pesquisa científica, mas o subtipo C é o mais prevalente no mundo (China, Índia,Africa Subsahariana e sul do Brasil).

Segundo o último boletim epidemiológico da UNAIDS (2008), estima-se que cerca de 33,0 milhões de pessoas estejam infectadas com o HIV em todo o mundo, sendo 30,8 milhões de adultos, 15,4 milhões de mulheres e 2,4 milhões de crianças infectadas com menos de 15 anos. Aproximadamente, no ano de 2007, ocorreram 2,5 milhões de novas infecções, com taxa aproximada de 7 mil novos casos de infecções ao dia. O número de óbitos decorrente da Síndrome de Imunodeficiência Adquirida

(SIDA, em inglês AIDS), no ano de 2007, foi de 2,1 milhões de pessoas, destas 330 mil eram crianças com menos de 15 anos, o que corresponde a uma taxa aproximada de 15,7% do número total de crianças infectadas com o HIV.

No Brasil, estima-se que existam 620 mil pessoas infectadas pelo HIV-1, o que corresponde um terço da população infectada pelo vírus na América Latina, com taxa de prevalência estimada de 0,5 (0,3-1,6)% na população adulta (15 a 49 anos) (UNAIDS, 2008). Deste total, 235 mil tem conhecimento de sua sorologia e 180 mil encontram-se em tratamento com os medicamentos antiretrovirais (MINISTÉRIO DA SAÚDE, 2007).

O Governo Brasileiro adotou, desde 1991, uma política que visa garantir o acesso universal à terapia com antiretrovirais (ARV) para indivíduos portadores do HIV-1, segundo critérios definidos por comitês técnicos assessores. Esta política tem causado um grande impacto na epidemia de HIV/AIDS, reduzindo a morbidade e a mortalidade.

Entretanto, o aparecimento de novas cepas do HIV-1, selecionadas pelo acúmulo de resistência às drogas antiretrovirais disponíveis é um problema global para o sucesso do tratamento da AIDS, representando assim um importante problema de Saúde Pública.

No caso do HIV-1, a resistência é uma conseqüência direta da diversidade do vírus, da não adesão ao tratamento, de problemas farmacobiológicos com os ARV (má absorção, eliminação) e do surgimento de resistência genética viral às drogas antiretrovirais (MINISTÉRIO DA SAÚDE, 2007). O acúmulo de mutações de resistência e a replicação continuada do vírus fazem com que a suscetibilidade às drogas diminua, reduzindo progressivamente a potência dos componentes do esquema terapêutico. Torna-se então necessária a utilização de testes laboratoriais de avaliação de resistência do HIV-1 à terapia antiretroviral.

Os testes de resistência permitem verificar a presença de mutações, tornando-se um importante instrumento de gerenciamento nas infecções por HIV-1. Estes testes se baseiam na análise do genoma viral, visando identificar mutações associadas à resistência (teste genotípico) ou na medida direta *in vitro* da suscetibilidade do vírus aos ARV (teste fenotípico). Entretanto, a interpretação de resistência aos medicamentos antiretrovirais, utilizando somente informações do genoma viral é complexa e, muitas vezes, exige análise técnica.

Atualmente, diversos estudos vêm sendo realizados visando a previsão e/ou a determinação de possíveis novas mutações que levam à resistência à terapia antiretroviral. Dentre estes estudos pode ser citado o trabalho de SEVIN *et al*. (2000), que utilizam análise de aglomerados e análise discriminante linear para investigar resistência de mutações antiretrovirais para os inibidores de protease saquinavir (SQV) e indinavir (IDV) -os resultados dessas análises foram semelhantes. Em particular, ambas as análises foram capazes de identificar a associação de mutações nas posições dos aminoácidos 10, 63, 71 e 90 com a resistência *in vitro* para o SQV e IDV.

Em WANG e LARDER (2003) encontramos a aplicação de redes neurais artificiais para a previsão de resistência ao inibidor de protease lopinavir (LPV). Para tal, os autores desenvolveram dois modelos de redes neurais. No primeiro utilizou-se, 11 posições de mutações da seqüência da protease descrita na literatura que promovem resistência para o lopinavir; e o segundo foi baseado nas 28 posições de mutação do gene da protease resultante da análise de prevalência no conjunto de dados. Para avaliar as performances dos modelos, utilizou-se o coeficiente de determinação r^2. Os resultados revelaram que o modelo com 28 mutações apresentou resultados mais precisos quando comparados com o modelo de 11 posições ($r^2 = 0,88$ contra $r^2 = 0,84$) na previsão de resistência para o inibidor em estudo, em um conjunto de teste de 117 casos.

DEFORCE *et al*. (2007) propõem a aplicação de redes neurais bayesianas, objetivando visualizar as relações entre o tratamento, mutações de resistência e a presença de polimorfismos para os inibidores de protease indinavir (IDV), saquinavir (SQV) e nelfinavir (NFV). Os resultados obtidos permitiram identificar as posições de mutação 30N, 88S e 90M para o NFV, 90M para SQV e 82A/T para IDV como as principais mutações de resistência.

Em DIRIENZO *et al*. (2003) um método não paramétrico é utilizado para avaliar a resistência ao amprenavir (APV), onde o método proposto é dividido em três etapas: a primeira consiste na construção do modelo de modo a permitir a previsão do fenótipo a partir da resposta da seqüência genotípica; a segunda na identificação de padrões específicos da seqüência de aminoácidos que mais impacta na previsão do fenótipo e a terceira na avaliação das combinações dos códons que apresentam padrões semelhantes ou não na ocorrência de mutação de resistência. Os resultados obtidos permitiram identificar oito códons (32, 46, 54, 71, 82, 84, 88 e 90) na região do genoma da protease do HIV capazes de caracterizar a resistência para o amprenavir.

1.1 Objetivos

1.1.1 Objetivo Geral

Nessa tese abordaremos o problema da resistência à terapia antiretroviral do HIV-1. O objetivo desse trabalho é propor um modelo computacional híbrido baseado na utilização de algoritmos genéticos (AGs) e no classificador *Kernel* Discriminante de Fisher (KDF), intitulado (AG/KDF), de modo a poder identificar possíveis novas mutações de resistência no gene da protease do genoma do HIV-1, bem como predizer a resistência em pacientes em falha terapêutica no Brasil, para os inibidores de protease (Saquinavir, Nelfinavir e Lopinavir). O banco de dados de sequências de protease do HIV utilizado é composto, na sua quase totalidade, de sequências dos subtipos B e C do HIV-1, o que nos permite traçar um paralelo entre os resultados obtidos para estes diferentes subtipos.

1.1.2 Objetivo(s) Específico(s)

Avaliar a aplicação da hidrofobicidade na codificação de seqüência de aminoácido da protease em dados de resistência do HIV-1.

Verificar a aplicabilidade do emprego de algoritmo genético na seleção de variáveis em dados de resistência a terapia antiretroviral do HIV-1.

Avaliar a capacidade discriminatória do modelo computacional entre os inibidores de protease e os respectivos subtipos B e C do HIV-1, provenientes tanto de grupos de indivíduos infectados com estes dois subtipos e em falha terapêutica usando inibidores de protease SQV, NFV e LVP, quanto de indivíduos naïve de tratamento com inibidores de protease ou ARV de forma geral.

1.2 Estrutura do Trabalho

O capítulo 2 descreve as bases teóricas da biologia molecular e do HIV-1 (seções 2.1 e 2.2), a terapia antiretroviral utilizada em pacientes infectados com o vírus (seção 2.3), os mecanismos de resistência à terapia antiretroviral (seção 2.4) e os métodos de classificação e seleção de variáveis (seção 2.5).

Na descrição das bases teóricas do HIV-1 (seção 2.2) são apresentados conceitos como: a estrutura da partícula viral; ciclo de replicação do HIV-1, estrutura genômica, diversidade genética do HIV-1 e sua distribuição mundial.

Na descrição dos princípios da terapia antiretroviral (seção 2.3) são citados o objetivo e combinação de medicamentos; os inibidores de transcriptase reversa nucleosídeo e não nucleosídeo, os inibidores de protease e os inibidores de fusão.

Na descrição dos princípios de resistência antiretrovirais (seção 2.4), são apresentados, respectivamente, os mecanismos de resistência do HIV-1, os tipos de resistências e as mutações de resistência descritas para os inibidores da transcriptase reversa, protease e fusão. Bem como os testes de fenotipagem e genotipagem, utilizados na avaliação de resistência dos vírus mutantes aos fármacos.

No capítulo 3, são apresentados os fundamentos teóricos dos AGs. Inicialmente é fornecida uma visão geral do processo de busca evolutiva onde é descrito o AG canônico, assim como seus principais operadores binários (seleção, mutação, recombinação e re-inserção). E em seguida (seção 3.5) é apresentado uma revisão bibliográfica da aplicação dos AGs na seleção de variáveis.

No capítulo 4, é apresentado o classificador de Kernel de Fisher Discriminante (KDF). Primeiramente são apresentadas as bases teóricas do classificador linear de Fisher. São apresentadas linhas gerais da teoria da função de kernel, importantes para a compreensão do classificador KDF. A seguir são apresentadas as bases teóricas do KDF, seguido dos principais trabalhos onde KDF foram empregados na classificação.

No capítulo 5 são apresentados o conjunto de dados e o algoritmo do modelo computacional proposto nessa tese, denominado AG/KDF, que associa um AG com o classificador KDF. Os resultados obtidos pelo modelo são apresentados no capítulo 6 e discutidos no capítulo 7.

No capítulo 8 apresentamos as conclusões da tese, recomendações finais e algumas propostas de trabalhos futuros na área de resistência a antiretrovirais do HIV-1.

Capítulo 2

Fundamentos Teóricos

2.1 Biologia Molecular

O código genético, na forma de unidades conhecidas como genes, reside no ácido desoxirribonucléico (DNA) no interior das células. O DNA contém apenas quatro diferentes bases: adenina, timina, citosina e guanina (abreviadas A, T, G e C), mas podem ser organizadas em qualquer seqüência. A ordem seqüencial dessas bases, em qualquer gene, determina a mensagem contida no mesmo, da mesma forma que letras do alfabeto podem ser combinadas de diferentes maneiras, formando novas palavras e orações (figura 2.1).

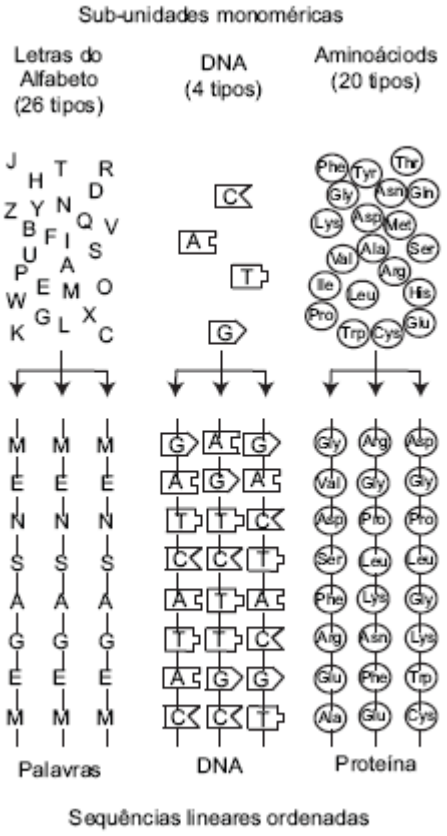

Figura 2.1: Analogia entre a informação molecular e aquela contida na linguagem escrita. Para um segmento de 8 unidades, o número possível de seqüências é igual a 26^8 ou 2,1 x 10^{11}, para as letras do alfabeto; 4^8, ou 65536, para as bases nucleosídicas do DNA; e, 20^8, ou 2,56 x 10^{10}, para os aminoácidos (adaptado de NELSON e COX, 2000).

Os genes guardam informações críticas a toda a vida. Enquanto todas as bases que compõem um gene são copiadas, nem todas as informações são mantidas. Isto se dá porque em um gene há tanto porções de base codificantes, quanto não-codificantes. Por exemplo, em uma partição teórica do gene (figura 2.2), seções codificantes, chamadas *exons*, fornecem as instruções genéticas que são copiadas para direcionar a construção de proteínas. Estas seções são preservadas, mas outras seções não-codificantes do gene, denominadas *introns*, são rapidamente removidas e degradadas. Próximo a cada gene está situada a seqüência promotora do DNA, que é capaz de "ligar" ou "desligar" o gene. Há ainda regiões indutoras, capazes de "acelerar" a atividade gênica e regiões repressoras, capazes de "frear" a atividade gênica. Os cromossomos também apresentam regiões não-codificantes localizadas fora dos genes. Estas contêm grandes porções de

seqüências repetidas. Algumas destas seqüências estão envolvidas na regulação da expressão gênica, e outras simplesmente atuam como espaçadores.

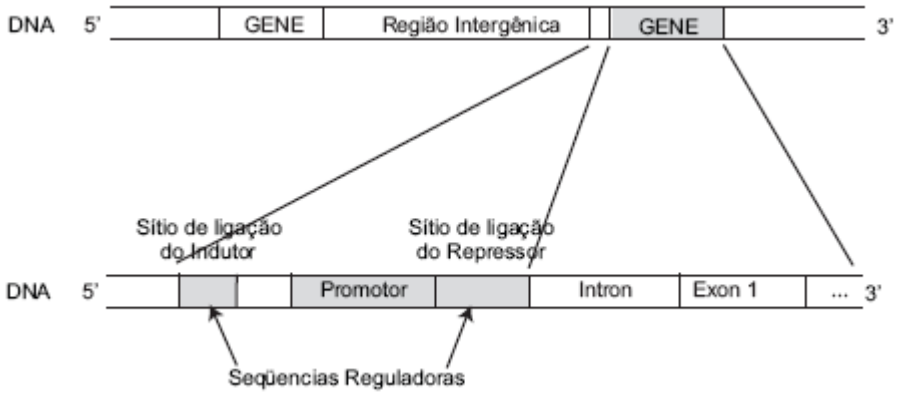

Figura 2.2: Visão simplificada da estrutura gênica. Dentro da seqüência linear do DNA há genes e regiões intergênicas. Há estruturas importantes na porção gênica do DNA, *introns*, ou regiões não-codificantes; e, *exons*, ou regiões codificantes; assim como as regiões intergênicas: região promotora, região indutora e região repressora, responsáveis, respectivamentes, pela ligação da RNA polimerase, pela transcrição gênica, sua indução (estabilidade da ligação entre a RNA polimerase e a seqüência de DNA) ou repressão (instabilidade de ligação, ou deslocamento da RNA polimerase), (adaptado de NELSON e COX, 2000).

Quando um gene é "ligado", ele eventualmente gera uma proteína, mas não diretamente. Primeiro, os genes codificam uma molécula intermediária denominada RNA mensageiro, processo conhecido com transcrição (figura 2.3). Há três tipos de RNA, ou ácido ribonucléico, são eles o mRNA, ou RNA mensageiro; o rRNA, ou RNA ribossomial; e, tRNA, ou RNA de transporte/transferência. Para transferir uma informação gênica do DNA para o mRNA, o pareamento de bases é usado. Para a replicação do DNA (figura 2.3) o pareamento de base se dá entre as bases adenina (A) e timina (T) ou entre citosina (C) e guanina (G). No processo de transcrição há uma alteração: uma base adenina (A) no DNA pareia com uma nova base uracil (U) no mRNA (figura2.4). Esta diferença auxilia a distinguir o mRNA do DNA. O mRNA, então, evade o núcleo pelo citoplasma através de organelas denominadas ribossomos. Nessas organelas, o mRNA direciona a formação de seqüências de amonoácidos, que dobram-se em uma única proteína. Antes de deixar o núcleo, o mRNA sofre um pré-processamento/maturação, onde o mRNA maduro, contém apenas *exons* que serão usados na construção da proteína, processo este denominado tradução (figura 2.3).

A tradução de seqüências de base do DNA em proteínas depende da disposição de tríades de nucleotídeos no mRNA. Cada tríade de mRNA, denominada *códon*, codifica para um único aminoácido (figura 2.5). O DNA referencia para um mRNA particular apenas quatro diferentes bases nucleotídicas em um gene, proporcionando assim 64 (4 x 4 x 4) combinações de códons disponíveis para codificar os 20 aminoácidos conhecidos (figura 2.5). A maioria dos aminoácidos é codificada por mais de um códon, entretanto, cada tríade está associada a apenas um aminoácido, característica conhecida como degradação do código genético.

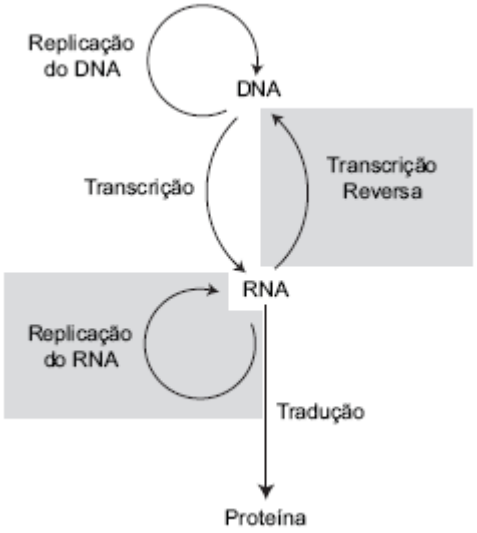

Figura 2.3: Dogma central da biologia molecular ampliado. A conservação da informação do DNA e sua decodificação em unidades funcionais (proteínas ou polipeptídeos) se dá pela seqüência de processos biomoleculares chamados de replicação, transcrição e tradução. O primeiro é responsável pela duplicação da fita de DNA, quando da divisão celular; o segundo é responsável pela geração de fitas-simples de RNA funcional (mRNA, tRNA e rRNA); e, o terceiro responsável pela geração de seqüências polipeptídicas (adaptado de NELSON e COX, 2000).

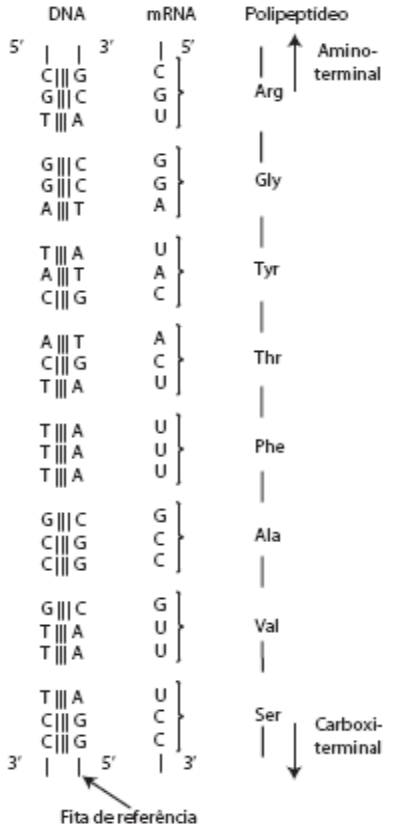

Figura 2.4: Decodificação da mensagem contida no DNA ao mRNA e, deste, ao polipeptídeo (adptado de NELSON e COX, 2000).

Segunda letra do codon

Primeira letra do codon (5′ término)	U		C		A		G	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
	UUA	Leu	UCA	Ser	UAA	Fim	UGA	Fim
	UUG	Leu	UCG	Ser	UAG	Fim	UGG	Trp
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Figura 2.5: Código genético humano. Durante o processo de tradução, t-RNAs específicos reconhecem a tríade de oligonucleotídeos (códon) da fita de mRNA, na estrutura ribossômica, promovendo a ligação peptídica entre os aminoácidos a eles associados (adaptado de NELSON e COX, 2000).

As mutações são alterações na seqüência de base normal do DNA, podendo ocorrer tanto nas regiões codificantes, quanto não-codificantes. As mutações podem ser silenciosas e não terem efeito sobre a proteína resultante. Isto se dá quando a mutação ocorre na região não codificante do DNA. Mesmo alterações no pareamento de bases da região codificante podem ser silenciosas, já que há redundância do código genético. Por exemplo, uma mutação em um códon pode ocorrer, e ainda assim codificar para o mesmo aminoácido que codificaria anteriormente.

Os genes podem sofrer mutações por vias distintas. O tipo mais simples de mutação envolve a alteração de uma única base ao longo da seqüência de bases de um gene particular, sendo denominada mutação pontual. Em outros casos, uma ou mais bases podem ser adicionadas (inserção) ou removidas (remoção).

As mutações pontuais podem ter efeitos variados na proteína resultante (figura 2.6). Uma mutação pontual *missense* substitui um nucleotídeo por outro diferente, mas deixa o resto do código intacto. O impacto dessas mutações pontuais depende do aminoácido específico, que é trocado na seqüência resultante. As *mutações nonsense* são mutações pontuais que alteram o códon do aminoácido para um dos três códons de "parada", resultando numa terminação precoce da tradução protéica. As mutações *nonsense* podem ser provocadas pela substituição de um único par de bases, ou pela mutação do *frameshift*. No núcleo, uma fita de mRNA copia o DNA de fita simples de forma exata. Este codifica precisamente para uma proteína, não deixando espaço separando as tríades (códons). O conjunto de tríades conectadas denomina-se *reading frame* ou "quadro de leitura". Uma mutação *frameshift* é causada pela adição ou perda de um ou mais nucleotídeos (figura 2.7). Esta mutação altera o conteúdo de todos os códons a partir do nucleotídeo modificado, ou seja, todos os "quadros de leitura" subseqüentes. Este tipo de mutação resulta, geralmente, em proteínas menores e não funcionais, já que não raramente geram códons de "parada" na seqüência que a segue. Se o número de pares de base ausente for múltiplo de 3, indicando a ausência de múltiplos aminoácidos, a proteína resultante pode ser alterada de forma drástica, e sua função dependerá da extensão dessa alteração (NELSON e COX, 2000).

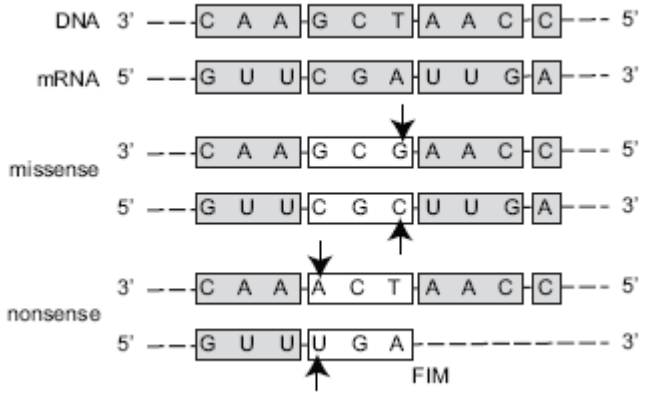

Figura 2.6: Exemplos de mutaçõs pontuais, *missense* e *nonsense*, da seqüência de DNA. A mutação (seta) da base timina (T) por uma base guanina (G) na molécula de DNA, leva à substituição do códon CGA, pelo códon CGC, ambos correspondentes ao aminoácido arginina (Arg) (*mutação missense*). Outra mutação (seta), agora da base guanina (G) pela base adenina (A), na seqüência de DNA, leva à substituição do códon CGA, pelo códon UGA, ou seja, um códon de sinalização de parada de transcrição (*mutação nonsense*) (adaptado de NELSON e COX, 2000).

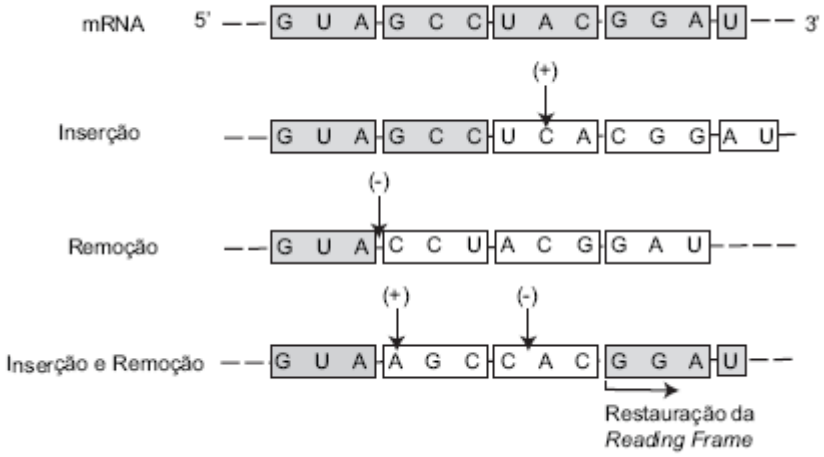

Figura 2.7: Mutações por inserção ou remoção de bases nucleotídicas. A inserção da base nucleotídica guanina (G), na molécula de DNA, que leva à adição (seta) da base citosina (C) na molécula de mRNA, assim como a remoção da base citosina (C) na molécula de DNA, que leva à subtração (seta) da base guanina (G) na molécula de mRNA, provocam a alteração do quadro de leitura (*frameshift*), alterando drasticamente a proteína resultante da transcrição. Já a dupla mutação, que gera a adição (seta) e subtração (seta), respectivamente, das bases adenina (A) e guanina (G) na molécula de mRNA, levam à restauração do quadro de leitura após a modificação de apenas dois aminoácidos (Ala-Tyr por Ser-His), gerando, possivelmente uma proteína funcional (adaptado de NELSON e COX, 2000).

2.2 O HIV-1

Os vírus estão envolvidos em uma grande variedade de doenças crônicas e degenerativas, sendo responsáveis por mais de 60% das doenças causadas no homem (GALO, 2002). O combate contra as infecções virais é difícil, pois sua replicação é um processo intracelular, estando intimamente relacionada ao metabolismo das células infectadas.

Atualmente um dos vírus mais estudados é o vírus da imunodeficiência humana (HIV-1 e HIV-2), capaz de parasitar o sistema imunológico do homem, causando a destruição de linfócitos T $CD4^+$. Segundo HAHN *et al.* (2000) estes tipos de vírus entraram na população humana através de múltiplas infecções zoonóticas, a partir de primatas não-humanos infectados com o vírus da imunodeficiência de símios (SIV).

Uma evidência preponderante para a transmissão entre espécies é a relação entre o HIV-2 de indivíduos da África Ocidental e o SIV_{sm} ("*sooty mangabey*"), primata da espécie *Cercocebus atys*, cujo habitat natural é a região costeira da África Ocidental. Isolados deste vírus mostram uma homologia de cerca de 80% nas seqüências de aminoácidos com o HIV-2, que é endêmico da mesma região geográfica (HIRSCH *et al*., 1989). Além disso, análises filogenéticas indicam que diferentes isolados de HIV-2 são mais similares a isolados de SIV_{sm} do que entre si, o que sugere recentes e contínuas transmissões entre espécies (GAO *et al*., 1992).

Desde 1992 já eram conhecidos fortes indícios da origem do HIV-2 (GAO *et al*., 1992), porém até o ano de 1999 a origem do HIV-1 permanecia incerta. Então, GAO *et al*., analisaram filogeneticamente todas as cepas de SIV conhecidas até o momento, e identificaram duas linhagens principais e altamente divergentes, que infectam duas subespécies de chimpanzés, uma da África Central, a *Pan troglodytes troglodytes*, e outra da África Oriental, a *Pan troglodytes schweinfurthii*. Apenas a linhagem de SIV que infecta o chimpanzé *Pan troglodytes troglodytes*, mostrou estar relacionada com o HIV-1, e parece ter originado todas as suas linhagens. Outros dados que corroboram esta afirmação são: (1) os dois vírus compartilham a mesma estrutura genômica; (2) vírus de chimpanzés e variantes do HIV-1 se agrupam próximos nas árvores filogenéticas; (3) chimpanzés e representantes de todos os grupos de HIV-1 compartilham a mesma região geográfica da África, onde podem ser encontrados todos os grupos do HIV-1 (M, O e N) e seus subtipos.

2.2.1 – Estrutura Viral

Assim como todos os retrovírus, o HIV é um vírus envelopado que possui um genoma diplóide constituído por uma fita simples de RNA com polaridade positiva (COFFIN, 1996). A partícula viral madura (figura 2.8) é constituída por um envelope externo, uma matriz e um nucleocapsídeo.

O envelope viral é constituído por uma camada bilipídica derivada da membrana citoplasmática da célula hospedeira durante a maturação do vírus. Duas glicoproteínas, codificadas pelo vírus, estão inseridas nesta camada, associadas não covalentemente em heterodímeros. São elas, a proteína de superfície *gp120* e a proteína transmembrana *gp41*, ambas responsáveis pela ligação ao receptor celular e entrada do vírus na célula hospedeira. O principal antígeno do vírus é a proteína *gp120*, que durante a entrada do vírus na célula, interage com o receptor celular CD4, localizado na superfície dos linfócitos T. Este receptor é parte integrante do receptor de células T (*"T cell receptor"* TCR) dos linfócitos T *helper* (DALGLEISH *et al*., 1984, MADDON *et al*.,1986), podendo estar presente também na superfície de células da linhagem monocitária/macrofágica e macrofágicas especializadas, como células da microglia, dendríticas e células de Langerhans. Além da *gp120* e *gp41*, o envelope viral também contém diversas proteínas do hospedeiro adquiridas durante a maturação do vírus, incluindo MHC (complexo de histocompatibilidade principal) de classe I e II, CD44, entre outras (LUCIW, 1996).

Internamente, no envelope encontra-se uma vasta matriz protéica fortemente associada à membrana e formada pela proteína viral *p17* (MA). O capsídeo da partícula viral madura apresenta formato cônico, típico dos lentivírus, e é formado pelo monômero estrutural constituído pela proteína *p24* (CA), também um importante antígeno viral. Encontra-se dentro do capsídeo um complexo nucleocapsídico, formado pelas proteínas do nucleocapsídeo *p7* e *p6* (NC), além das enzimas virais transcriptase reversa (*p66/p51*), protease (*p12*) e integrase (*p32*).

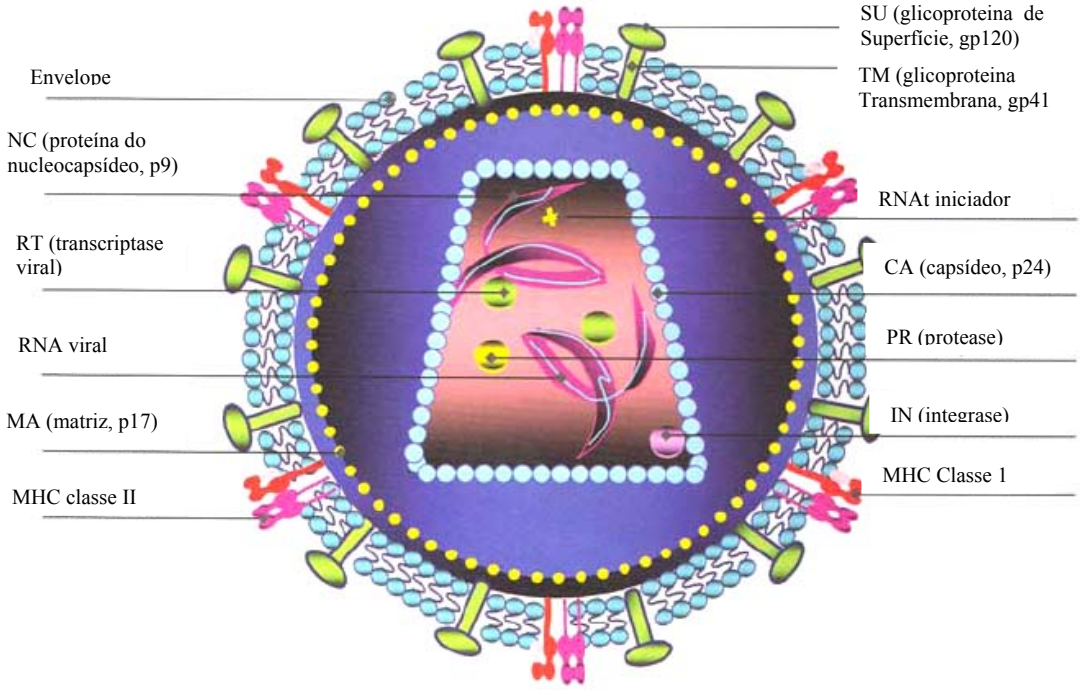

Figura 2.8: Estrutura morfológica do vírus da imunodeficiência humana do tipo 1 (HIV-1), (adptado de SANTOS, ROMANOS e WIGG, 2002).

Como todos os outros retrovírus, o HIV possui uma organização genômica complexa, com aproximadamente 9,8 Kb, com nove genes que apresentam diversas possibilidades de processamentos alternativos (figura 2.9), fato este que permite a síntese de um elevado número de diferentes polipeptídeos, proteínas e enzimas.

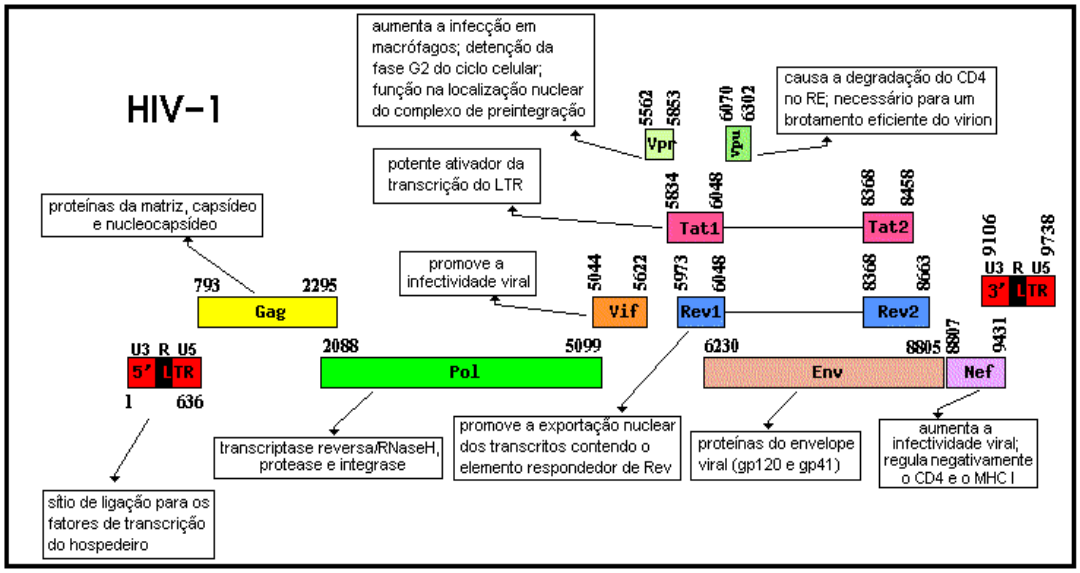

Figura 2.9: Representação esquemática da estrutura do genômica do HIV-1 (adpatado de www.nature.com)

Pode se observar na figura 2.9, que o HIV-1 é formado por 3 genes estruturais essenciais para a sua replicação: a região específica do grupo antígeno (*gag*), que codifica as proteínas estruturais internas *p17*, *p24*, *p7* e *p6* durante o processo de maturação do vírus; a região polimerase (*pol*), que codifica a protease (*p11*, PR), transcriptase reversa (*p66/p51*, RT), integrase (*p31*, IN) e a região envelope (*env*), responsável pela codificação das proteínas do envoltório, *gp120* e *gp41*.

O genoma do HIV-1 codifica ainda seis outras proteínas acessórias, sendo duas: a proteína transativadora (*tat*) e a reguladora da expressão viral (*rev*) responsáveis pela regulação da expressão gênica e as demais: proteína viral R (*vpr*), proteína viral U (*vpu*), proteína da infectividade viral (*vif*) e fator de regulação negativa (*nef*), responsáveis, respectivamente, por: facilitar a entrada do complexo pós integração, maturação e liberação da partícula viral das células infectadas, controlar a produção das proteínas e regular o sucesso do ciclo infeccioso viral. Na forma de provírus, o genoma viral possui em cada uma de suas extremidades uma longa seqüência repetida ou LTR (*Long Terminal Repeats*), denominadas de LTR5' e LTR3', que possibilita a integração no genoma da célula hospedeira.

2.2.2 Ciclo de Replicação do HIV-1

O ciclo infeccioso do HIV-1 pode ser dividido em três estágios distintos: (1) fusão viral e transcrição reversa do genoma viral, (2) integração do cDNA viral no genoma celular e (3) expressão gênica pelo provírus e formação de novas partículas (figura 2.10).

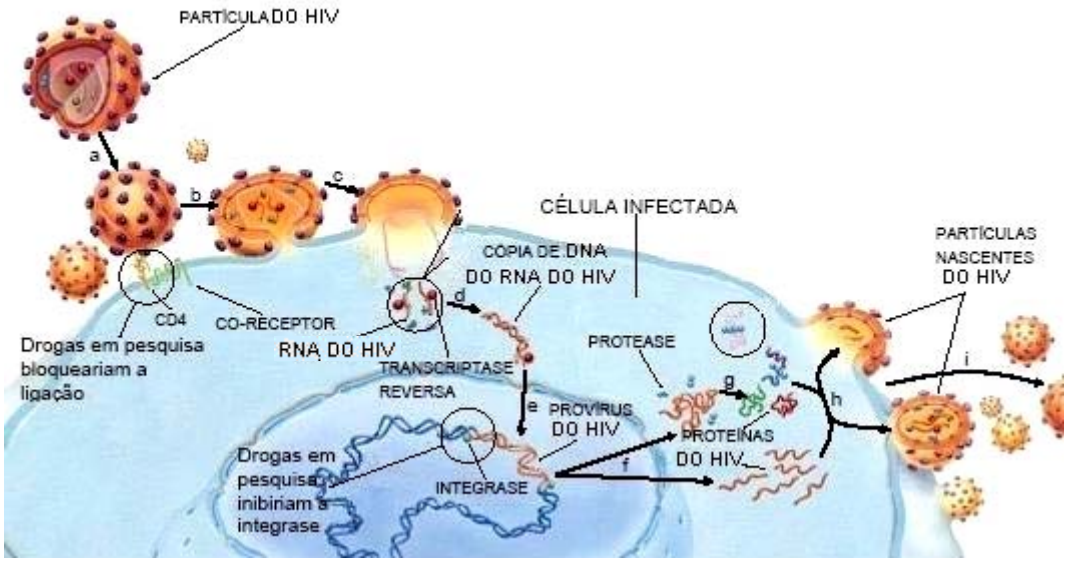

Figura 2.10: Esquema do ciclo de replicação do HIV que começa quando o vírus se liga a superfície celular (a) funde-se à membrana celular (b) e libera seu conteúdo dentro do citoplasma da célula hospedeira (c). Depois, a enzima de transcriptase reversa do HIV copia o material genético viral a partir do RNA em DNA de filamento duplo (d), o qual será unido ao DNA celular pela enzima de integrase do HIV (e). Usando-se do DNA integrado, ou provírus, como uma matriz, a célula produz RNA e proteínas virais (f). Uma terceira enzima, a protease do HIV, parte as novas proteínas (g), habilitando-as a juntar o RNA em novas partículas virais (h) que brotarão da célula (i) e infectarão outras (j) (adaptado de www.sciam.com).

Inicialmente, ocorre a interação do vírus com a membrana da célula hospedeira via associação da proteína viral *gp120* com a molécula CD4, uma proteína tipo imunoglobulina (Ig) expressa na superfície de células T e macrófagos primários (CHAN e KIM, 1998). Após a ligação à membrana celular, a proteína *gp120* dissocia-se da proteínas *gp41*, que passa por modificações conformacionais que promovem a fusão vírus-célula, permitindo a entrada do capsídeo na célula.

Terminada a fusão, o capsídeo do virion é então desencapado em um processo que consiste na liberação no citoplasma do conteúdo do capsídeo, o RNA genômico e enzimas virais, o que se faz necessário para a etapa posterior, a transcrição reversa. A transcriptase reversa (RT) promove a síntese de uma cópia de DNA de fita dupla, catalisando as reações de polimerização de DNA dependente de RNA e dependente de DNA, além de clivar a porção de RNA do híbrido RNA-DNA formado durante este estágio. Em seguida o complexo nucleoproteico (enzimas e DNA) formado é transportado para o núcleo da célula hospedeira em um processo mediado pela proteína *Vpr*. A ação da enzima viral integrase (IN), permite uma integração estável do cDNA do genoma viral no DNA cromossômico da célula hospedeira no que resulta na formação de um pró-vírus, completando assim a fase pré-integrativa.

Uma vez integrado no DNA hospedeiro, o pró-vírus comporta-se como um gene celular residente. O conjunto de RNAs transcritos são então transportados para o citoplasma, onde serão traduzidos, ou constituirão novas partículas virais.

Os virions são inicialmente montados próximo à membrana celular na forma de partículas imatura, compostas de um envelope glicoprotéico, RNA genômico e poliproteínas virais (GONDA *et al*., 1986). Durante ou após o "brotamento", as partículas virais passam por uma modificação morfológica denominada maturação. A maturação consiste na elisão das poliproteínas *gag* e *gag-pol* pela protease viral (PR), produzindo enzimas e proteínas estruturais do capsídeo. O processamento das poliproteínas no vírion completa o ciclo de replicação do HIV, tornando os vírions maduros capazes de infectar um linfócito adjacente.

2.2.3 A Diversidade Genética do HIV-1

O HIV-1 se caracteriza por uma enorme diversidade genética e antigênica (figura 2.11). Na região que codifica as glicoproteínas do envelope (gene *env*), por exemplo, estima-se que a amplitude dessa diversidade possa ser superior a 10% em um único paciente e podendo essa amplitude de variabilidade chegar a até 50% entre cepas de diferentes grupos (PINTO e STRUCHINER, 2006).

O primeiro esforço na tentativa de organizar esta grande diversidade das seqüências do HIV-1 foi subdividi-las em cepas Européias/Americanas e Africanas, pois as seqüências derivadas de isolados virais da Europa e da América do Norte formavam um aglomerado distinto em árvores filogenéticas, enquanto que as cepas Africanas se separavam em diferentes linhagens. Com a descoberta de novos espécimes, esta separação tornou-se inadequada e foi a princípio proposta uma classificação em subtipos com base na análise dos genes *env* e *gag*. Atualmente a classificação adotada se baseia na análise do genoma completo de amostras de HIV-1, colhidas em diferentes regiões geográficas. Fato este que possibilitou classificar a diversidade viral do HIV-1, em: grupo M (*major*), constituído por nove subtipos nomeados (A-D, F-H, J e K), sendo relativamente eqüidistantes, com exceção do subtipo B e D que são mais próximos (figura 2.11). Um segundo grupo com características diferentes ao grupo M foi identificado na República dos Camarões, sendo denominado de grupo O (*outliers*). Por último foi descrito um terceiro grupo classificado de N (*non M and non O*). Além destes, 16 formas recombinantes (CRF) circulam na epidemia, sendo as mais freqüentes as CRF02_AG e CRF01_AE .

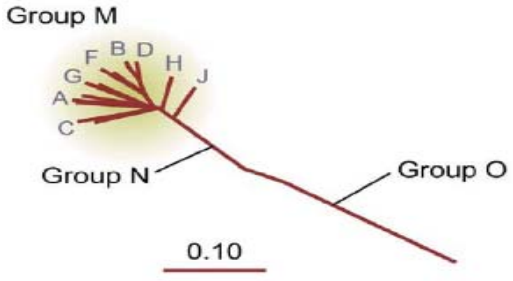

Figura 2.11: Classificação da diversidade do HIV-1. A barra representa uma diferença de 10% na seqüência do nucleotídeo (adaptado de www.vircolab.com)

Segundo PEETERS (2000), os subtipos têm sido fortes marcadores epidemiológicos para delinear o perfil da epidemia do HIV-1, tendo sua origem pautada em acidentes epidemiológicos. A figura 2.12, mostra a prevalência e distribuição geográfica global dos subtipos e as formas recombinantes circulantes (CRFs) do HIV-1.

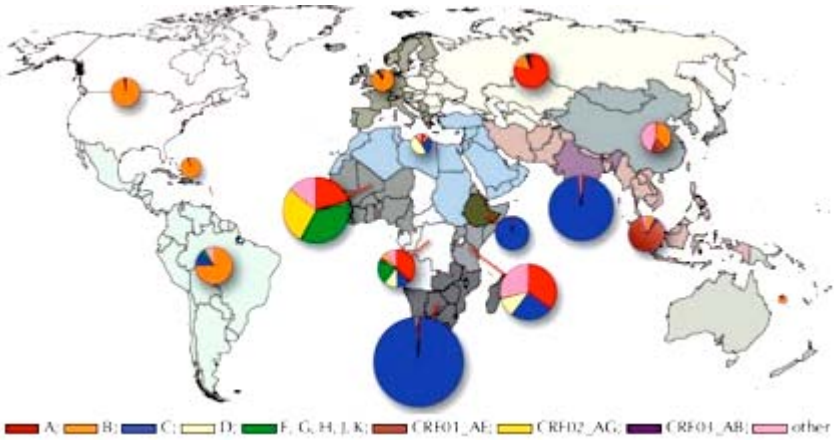

Figura 2.12: Distribuição geográfica dos subtipos e formas recombinantes circulantes (CRFs) do HIV-1. Adaptado de *Los Alamos* HIV *Sequence Database* [http://hiv-web.lanl.gov].

A distribuição das variantes do HIV-1 na população humana é heterogênea. O subtipo A predomina na África, o subtipo B predomina na Europa e na América do Norte. O subtipo C predomina na África, sendo responsável pelos epicentros da epidemia do HIV-1, devido à dispersão incontrolada no Botswana, Zimbabwe, Malawi, Zâmbia, Namíbia, Lesoto, África do Sul, Índia, Nepal e China (SPIRA *et al*., 2003). O subtipo F foi documentado na América do Sul e na Europa. O subtipo G foi relatado na África e na Europa. O subtipo D está geralmente limitado a África, Europa e América do Norte.

No Brasil, os primeiros estudos realizados com o intuito de identificar a prevalência de subtipos, detectaram a presença dos subtipos B e F, onde o subtipo B mostrava ser o mais freqüente. No estudo realizado por MORGADO *et al*. (1998), analisaram a região C2V3 do *env* de 131 pacientes do Rio de Janeiro (RJ) e identificaram que 80,9% das amostras eram de subtipo B, 15,3% de subtipo F e uma única amostra de (0,8%) de subtipo D. Estes dados foram confirmados com os achados de TANURI *et al*. (1999), que seqüenciaram as regiões *gag* p24, *env* C2V3, e *env* gp41 de 43 doadores de sangue do Rio de Janeiro. Análises filogenéticas mostraram que

76,70% destas amostras pertenciam ao subtipo B, 14,00% ao subtipo F e 9,30% eram mosaicos dos subtipos B e F ou B e D.

Através de um estudo de coorte em 14 pacientes no Rio de Janeiro, tratados com terapia antiretroviral potente (três ou mais drogas) – HAART para a região da transcriptase reversa, CARIDE *et al*. (2000), demonstraram que 35,60% dos pacientes pertenciam ao subtipo não-B (28,50% subtipo F e 7,10% subtipo A), e 64,40% ao subtipo B.

O estudo realizado por SOARES *et al*. (2003) mostra evidências que a distribuição da infecção de acordo com os subtipos no Brasil parece estar sofrendo alterações, este estudo se baseou nas regiões da protease e da transcriptase reversa do HIV-1 de 105 indivíduos soro positivos sem experiência de antiretrovirais, de diversos estados brasileiros (Paraná, Rio de Janeiro, São Paulo, Rio Grande do Sul, Mato Grosso e outros), onde 55,20% dos pacientes eram subtipo B (n = 58), 28,60% subtipo C (n = 30), 6,70% subtipo F (n = 7) e 9,50% possuíam genótipos recombinantes entre dois ou mais subtipos (n = 10). Paralelamente, BRINDEIRO *et al*. (2003), publicaram um estudo que também analisa indivíduos soro positivos sem experiência de antiretrovirais, no entanto o estudo abrange um número maior de sítios de coleta e com isto uma amostragem superior (n = 535). Do total de amostras amplificadas, foi encontrado 64,90% e 62,50% de subtipos B, 22,80% e 29,50% de subtipos C e 11,80% e 8,00% de subtipo F, respectivamente para os genes da protease e transcriptase reversa viral. Quando as duas regiões genômicas foram analisadas simultaneamente, 42 seqüências apresentaram evidências de divergências de subtipos, representando juntas 14,5% do total de amostras analisadas em ambas as regiões.

Assim, baseado na análise do polimorfismo genético do HIV-1 em diferentes regiões do Brasil, foi possível identificar um predomínio do subtipo B na maioria das regiões, bem como a presença importante dos subtipos C e F, cujas freqüências variam entre regiões.

2.3 Terapia Antiretroviral

O objetivo principal da terapia antiretroviral é retardar o desenvolvimento da imunodeficiência e/ou estabelecer, quando possível, a imunidade, aumentando o tempo e melhorando a qualidade de vida da pessoa infectada. No entanto, as terapias atuais, nem sempre conseguem manter uma supressão viral duradoura na maioria dos pacientes.

Quando uma combinação terapêutica potente é administrada com eficiência em um paciente portador do HIV-1, os níveis de RNA viral plasmático e de células infectadas no tecido linfóide são rapidamente diminuídos. Tal comportamento é atribuído à morte de linfócitos T $CD4^+$ infectados ativados e a prevenção de novas infecções. Após esta morte acentuada dos linfócitos T $CD4^+$ infectados, ocorre uma segunda etapa mais lenta e mais heterogênea entre os indivíduos, associada à eliminação de macrófagos infectados ou a virions ligados a células dendríticas nos linfonodos, que também pode estar relacionada a linfócitos T $CD4^+$ cronicamente infectados com uma maior meia-vida e com uma menor taxa de replicação viral (PERELSON, 1997).

A impossibilidade de se reduzir o RNA viral a um valor abaixo de 50 cópias por mililitro de plasma (o limite de detecção dos atuais ensaios de carga viral) indica uma supressão inadequada (falha) e risco do crescimento de vírus resistentes.

Segundo o CDC (2006), o tratamento deve ser oferecido a todos os pacientes HIV-positivos na fase aguda, ou dentro dos seis meses da soroconversão, além de a todos os pacientes com sintomas descritos para infecção pelo HIV. A recomendação para iniciar a terapia em pacientes assintomáticos depende de fatores virológicos e imunológicos. Em geral, o tratamento deve ser oferecido a pacientes com menos de 500 linfócitos T $CD4^+$ /mm^3 ou carga viral acima de 10.000 cópias/ml pelo método de bDNA ou 20.000 cópias/ml pelo método da reação em cadeia da polimerase com transcrição reversa (RT-PCR).

O resultado da terapia deve ser avaliado primariamente através da carga viral do paciente. Espera-se que após a transformação logarítmica da carga viral, a mesma tenha pelo menos um decréscimo de um 1*log* (em adultos) em oito semanas e que chegue a níveis indetectáveis (<50cópias de RNA viral/ml de sangue) em 4 a 6 meses depois do início da terapia. Os pacientes que não apresentarem este quadro são considerados em falha terapêutica, e as causas podem estar associadas aos seguintes fatores: pouca

aderência; concentrações sub-ótimas da droga (por má absorção ou extrusão celular); potência inadequada da droga; ou resistência viral.

O MINISTÉRIO DA SAUDE (2003) utilizou no Brasil, as recomendações do CDC como base para elaboração das recomendações para terapia antiretroviral em pacientes adultos e adolescentes, sendo que, algumas modificações foram incorporadas em relação a quando e como iniciar a terapia. Desta forma, o tratamento antiretroviral é indicado para todos pacientes infectados pelo HIV, sintomáticos ou assintomáticos, que apresentam contagem de linfócitos T CD4+ abaixo de 200/mm^3. No caso de paciente assintomático, ou seja, paciente que apresenta contagem de linfócitos T CD4+ entre 200 e 350/mm^3, o início da terapia antiretroviral deve ser considerado conforme a evolução dos parâmetros imunológicos (contagem de linfócitos T CD4+), virológicos (carga viral) e outras características do paciente (motivação, capacidade de adesão, co-morbidades).

O uso de esquemas antiretrovirais HAART está recomendado para todos pacientes em início de tratamento. Esquemas de terapia dupla (dois análogos de nucleosídeos) não estão mais indicados para início de tratamento e para os casos de co-infecção HIV-tuberculose, sendo, mantidos somente como opção de quimioprofilaxia em algumas situações de exposição ocupacional. Os pacientes que já apresentaram falhas terapêuticas em esquemas de terapia dupla, ou potente, é recomendado um esquema mega potente – MEGAHAART (dois análogos nucleosídeos mais um (1) análogo não nucleosídeo e um (1) inibidor de protease, ou dois análogos nucleosídeos mais dois (2) inibidores de protease afora o ritonavir).

2.3.1 As Drogas Antiretrovirais

O uso de drogas antiretrovirais constitui a forma de tratamento da infecção pelo HIV-1 onde se têm registrado avanços mais significativos.

O ciclo de replicação do HIV-1 apresenta diversos eventos exclusivamente relacionados a componentes virais, que podem ser utilizados como alvos para intervenção quimioterápica (PEÇANHA *et al*., 2002). Segundo SOUZA e ALMEIDA (2003), atualmente os compostos disponíveis anti-HIV, atuam na inibição no sítio de ligação das seguintes enzimas: Transcriptase Reversa (RT), Inibidores de Protease (IPs),

Inibidores de Fusão (IFs) e Inibidores de Integrase (inibidores de transferência de fita, "strand transfer inhibitors", ITs).

2.3.1.1 Inibidores da Trancriptase Reversa

Dois grupos de inibidores da RT têm sido extensivamente investigados: os inibidores análogos de nucleosídeos da transcriptase reversa (NRTI) e os inibidores análogos não nucleosídeos da transcriptase reversa (NNRTI). Os NRTI catalisam um dos processos mais característicos dos retrovírus, ou seja, a transcrição reversa de seu RNA em cDNA dupla fita. Este processo é essencial para a replicação viral, e por isso a transcriptase reversa (figura 2.13) foi o primeiro alvo no desenvolvimento da terapia anti-HIV (POCH *et al*., 1989).

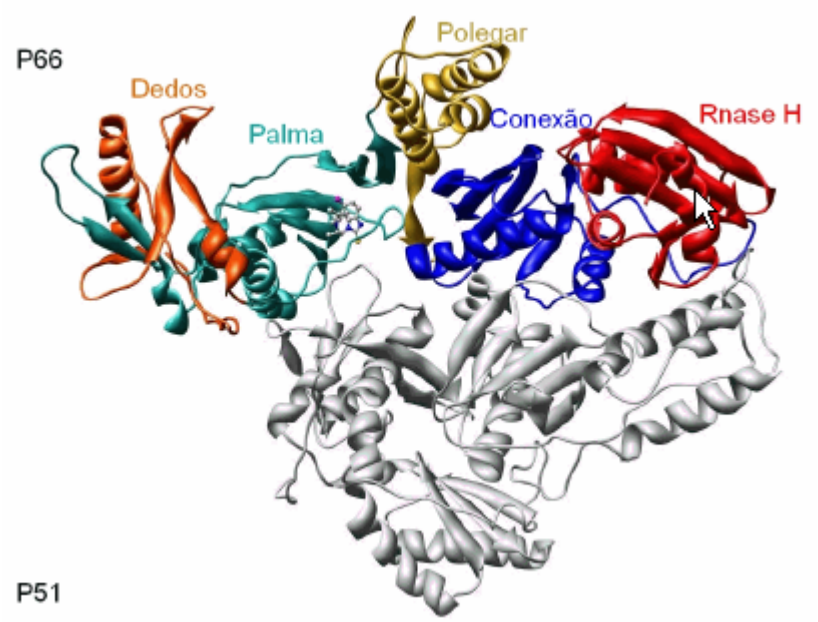

Figura 2.13: Estrutura cristalina da enzima RT do HIV-1 mostrando que a mesma é constituída de uma cadeia de 560 resíduos de aminoácidos (p66) e uma segunda cadeia compreendendo os 440 resíduos iniciais da *p66* (*p51*). A subunidade *p66* pode ser dividida em cinco domínios comumente conhecidos como dedos, palma e polegar, em função da semelhança com a mão direita entre aberta, além dos domínios de conexão e RNase H (adaptado de www.vircolab.com).

Os NRTIs necessitam sofrer uma ou duas fosforilações para se tornarem ativos. Eles atuam ligando-se ao sítio ativo da polimerase e se incorporando dentro do filamento de DNA provocando, então, a interrupção da síntese do DNA viral. Atualmente encontra-se disponível no mercado farmacêutico, sete análogos nucleosídeos (NRTI): Zidovudina (AZT), Stavudina (d4T), Emtricitabine (FTC), Lamivudina (3TC), Didanosina (ddl), Abacavir (ABC) e Tenofovir (TFV).

Segundo MERLUZZI *et al*. (1990), os inibidores NNRTIs são muito menos tóxicos que os NRTI e estão estruturalmente relacionados às benzodiazepinas (TIBO e BIRG-857) e a um derivado da piridinona. Essa classe de inibidores não competitivos, ligam-se a uma posição adjacente ao sítio ativo enzima, causando uma mudança conformacional da molécula a qual reduz sua atividade. Existem atualmente no mercado três inibidores (NNRTIs) disponíveis para o tratamento de pacientes portadores do HIV-1: Nevirapina (NVP), Efavirenz (EFV) e Etravirine (TMC-125).

2.3.1.2 Inibidores de Protease.

A protease do HIV-1, responsável pelo processo pós-traducional das poliproteínas virais *gag* e *gag-pol*, transformando-as nas proteínas estruturais e enzimas presentes no vírion (partículas virais nascentes), é uma protease aspártica constituída por dois monômeros, não covalentes associados e idênticos entre si, de 99 aminoácidos. A enzima contém uma região flexível denominada *abas* que se fecha sobre o sítio ativo após a ligação com o substrato (figura 2.14).

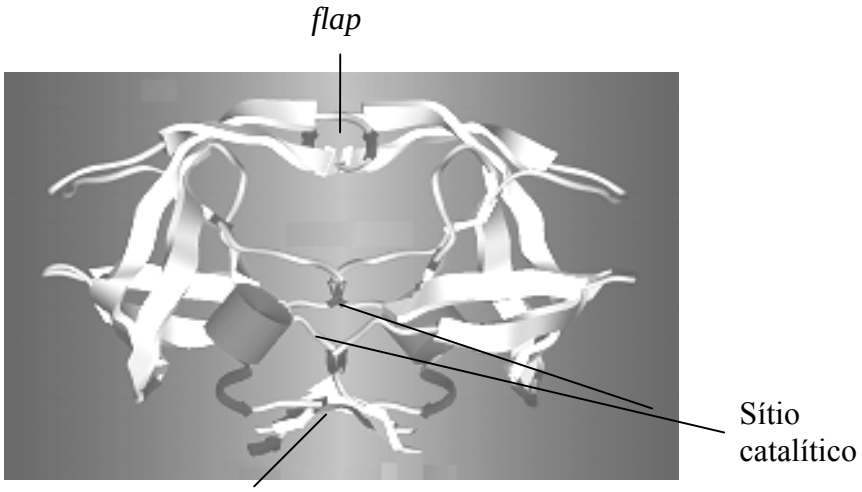

Figura 2.14: Representação estrutural da protease do HIV-1 (adptado de Hhttp://mcl1.ncifcrf.gov/hivdb/index.htmlH)

Seu sítio catalítico apresenta estrutura semelhante ao de outras proteases aspárticas, apresentando nas posições 25-27 uma tríade catalítica *Asp* (ácido aspártico) - *Thr*(Threonine) - *Gly*(glycina) (TURNER e SUMMERS, 1999). O sítio hidrofóbico de elisão (*hydrophobic substrate cleft*) reconhece e fragmenta nove diferentes seqüências para produzir a matriz, o capsídeo, o nucleocapsídeo e as proteínas *p6* da poliproteina *gag*, além da protease, transcriptase reversa e integrase a partir da poliproteína *gag-pol* (WLODAWER e GUSTCHINA, 2000).

Os Inibidores de Protease (IPs) pertencem a uma classe de compostos não peptídicos que se ligam ao sítio ativo da enzima por competição com o substrato natural *gag* (e *gag-pol*), bloqueando a produção de vírus infecciosos de células infectadas por meio da inibição da fragmentação das poliproteínas precursoras necessárias para produzir virions. As partículas virais produzidas por células submetidas a tratamento com IPs, possuem precursores não processados e não são infecciosas. Segundo JOHNSON *et al*. (2008), existem atualmente oito inibidores peptídicos da HIV-protease (IPs) aprovados pela FDA (*Food and Drug Administration*) e são baseados em seqüências reconhecidas e fragmentadas em proteínas do HIV-1. Saquinavir/ritonavir (SQV/r), Indinavir/ritonavir (IDV/r), Darunavir/ritonavir (DRV/r), Nelfinavir (NFV), Fosamprenavir (FPV/r), Lopinavir/ritonavir (LPV/r), Atazanavir/ritonavir (ATV/r) e Tripanavir/ritonavir (TPV/r).

2.3.1.3 Inibidores de Acoplamento e Fusão

Os Inibidores de Fusão (IFs) e acoplamento representam uma nova abordagem na estratégica de combate à capacidade de replicação do HIV-1 no organismo. Para que o vírus complete o seu ciclo reprodutivo, o mesmo necessita se fundir com o linfócito T, onde deposita as informações genéticas, dando origem a novos vírus. Os IFs e de acoplamento foram concebidos de modo a bloquear a interação da *gp120* com o CD4, a interação da *gp120* com os co-receptores (CCR5 e CXCR4) ou inibir as interações da célula com a *gp41* (figura 2.15); ou seja, visam impedir que o vírus consiga penetrar nos linfócitos ou monócitos, não possibilitando que o mesmo inicie o processo infeccioso (DOMS, 2004). Atualmente existe um inibidor de fusão e outro de acoplamento aprovados pelo FDA: Enfuvirtide e Maraviroc, respectivamente.

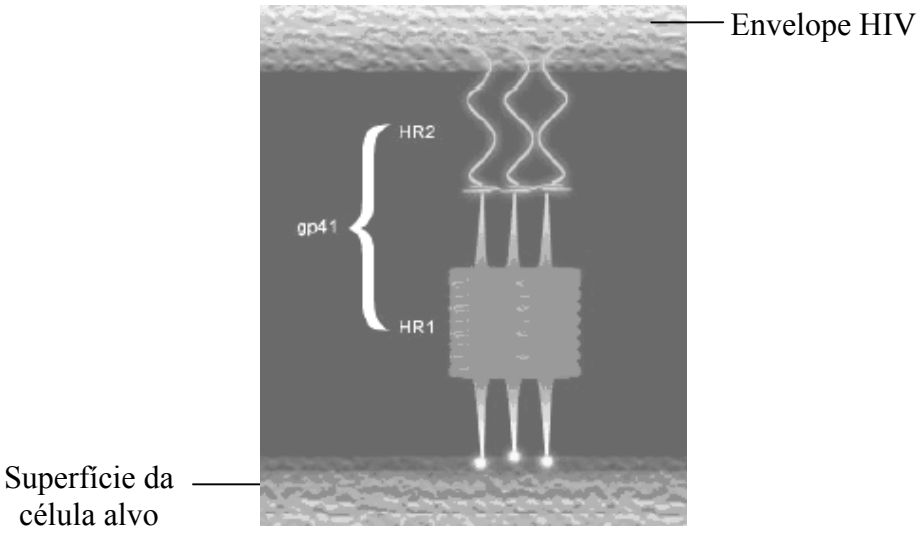

Figura 2.15: Mostra a estrutura da *gp41*, incluindo as regiões denominadas HR1 (Heptad Repeat 1) e HR2 (Heptad Repeat 2). Na *gp41* a dobra da HR1 juntamente com a HR2 é crítica para a fusão da membrana dos linfócitos (adaptado de www.vircolab.com).

2.3.1.3 Inibidores de Integrase

A enzima integrase codificada pela terminação 3' do gene viral *pol*, é fundamental por inserir o DNA proviral no cromossomo do hospedeiro, sendo fundamental no processo re replicação viral. É composta de uma cadeia polipeptítica simples que se dobra em três domínios funcionais (figura 2.16): o domínio N-terminal (resíduos de aminoácidos 1-50), onde se encontram dois resíduos de histidina e dois de cisteína que promovem a ligação com Zn(II); o domínio "cerne" (resíduos de amino ácidos 50-212), contendo os sítios catalíticos para a endonuclease e polinucleotidil transferase, a tríade ácida Asp 64, Asp 116 e Glu 52, onde se ligam Mn(II) ou Mg(II), chamada de "motriz DDE"; e o domínio C-terminal (aminoácidos 213-288), contém resíduos de amino ácidos básicos e liga-se ao DNA (ADESOKAN *et al*., 2004).

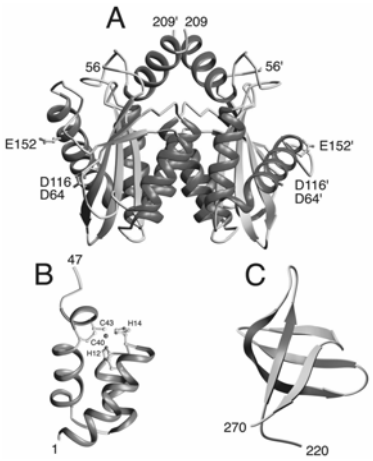

Figura 2.16: Representa a estrutura da enzima integrase do HIV-1 e seus respectivos domínios funcionais (adaptado de CRAIGE, 2001).

2.4 Resistência aos Antiretrovirais

O surgimento da resistência às drogas antiretrovirais é o fator limitante para o sucesso da terapia da AIDS. A falha terapêutica de muitos pacientes portadores do HIV-1 está relacionada a não adesão ao tratamento e a replicação elevada de vírus mutantes resistentes aos alvos moleculares dos fármacos administrados na terapia (SHAFER, 2002). A resistência é simultaneamente causa e conseqüência da replicação do vírus na presença das drogas anti-HIV. A replicação residual, sob pressão seletiva das drogas antiretrovirais, cria um ambiente de seleção que, face ao surgimento de mutações associadas à resistência, numa pequena população viral, dá a esta variante viral uma vantagem seletiva. Esta pressão seletiva possibilita a resistência destas variantes, tornando-as predominantes.

O acúmulo de mutações de resistência diminui a suscetibilidade às drogas, reduzindo progressivamente a potência dos componentes do esquema terapêutico. A replicação contínua do vírus na presença das drogas aumenta a resistência às mesmas. Sendo assim, esquemas terapêuticos impotentes, adesão sub-ótima, absorção limitada e compartimentos corporais tratados de forma não efetiva podem permitir o surgimento de vírus resistentes; fato este que origina um ciclo vicioso de falha terapêutica, deixando o tratamento ainda mais difícil (RICHMAN *et al*., 2003).

Frequentemente, na ausência da droga, os vírus mutantes apresentam um *fitness* reduzido, a ausência de uma pressão seletiva, motivada, por exemplo, pela interrupção do tratamento, levará a substituição dos vírus mutantes pelo vírus selvagem (*wild-type*),

ou susceptível às drogas, de modo progressivo. No entanto, observa-se que estas populações virais persistem no plasma, mesmo que em populações virais minoritárias, chegando a não ser às vezes detectável nos testes de resistência (VANDAMME *et al*., 2004).

2.4.1 Mecanismos de Resistência

A análise por sequenciamento do gene *pol* presente em isolados resistentes às drogas antiretrovirais, demonstram tanto na Transcriptase Reversa (RT) como na Protease viral (PR) a presença de determinadas posições na cadeia peptídica que são alvos específicos para mutações que resultam na troca de aminoácido. Estas inúmeras mutações tanto na RT quanto na PR, estão associadas à resistência às drogas antiretrovirais, podendo ser estas mutações classificadas em mutação primária ou mutação compensatória (secundária). A denominação primária refere-se às mutações que por si só reduzem a susceptibilidade a uma droga. Já a denominação compensatória refere-se às mutações que conjuntamente com uma mutação primária diminuem a susceptibilidade a uma determinada droga ou melhoram o *fitness* viral. Algumas mutações podem ser consideradas primárias relativamente a uma determinada droga e compensatória para outra (SHAFER, 2002).

Além das mutações descritas anteriormente, outro grupo de mutações merece atenção ao se estudar o HIV-1: os polimorfismos genéticos, que são definidos como variações genéticas comuns (freqüência maior que 1%) dos vírus isolados de indivíduos sem terapia ARV (naïve de tratamento), i.e., estão presentes na população viral mesmo na ausência da pressão seletiva exercida pelas drogas antiretrovirais (DIAS, 2004).

2.4.1.1 Resistência aos Inibidores Análogos Nucleosídeos (NRTIs)

O mecanismo bioquímico de resistência aos NRTIs é mediado por mutações que possibilitam à enzima RT reconhecer as NRTIs durante a síntese, impedindo a sua adição à cadeia de DNA que está a formar. As mutações associadas à resistência aos inibidores NRTIs são mostradas na (figura 2.17).

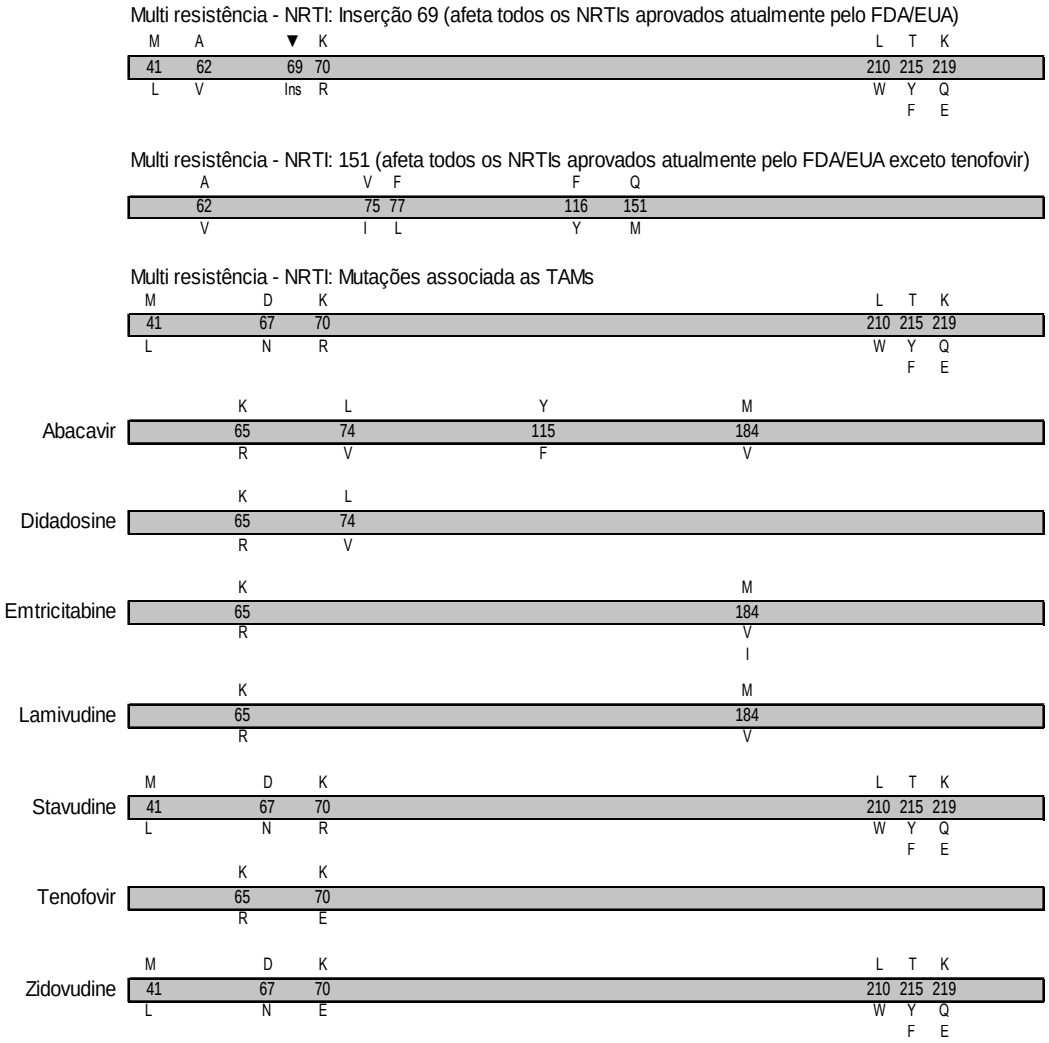

Figura 2.17: Mutações mais freqüentes encontradas no gene da transcriptase reversa associada à resistência a inibidores análogos nucleosídeos (NRTIs). Na figura as linhas verticais indicam TAMs (*Thimidine-Associated Mutations*), a letra superior refere-se ao aminoácido *wild type*, a letra inferior ao aminoácido de substituição, a numeração em negrito indica as mutações primárias e a restante as mutações compensatórias (adpatada de JOHNSON *et* al., 2008).

2.4.1.2 Resistência aos Inibidores Análogos Não Nucleosídeos (NNRTIs)

As mutações responsáveis pela resistência aos NNRTIs (figura 2.18), localizam-se no sítio hidrofóbico, onde estes inibidores se ligam. Uma única mutação neste local pode acarretar um elevado nível de resistência a um ou mais NNRTIs. Geralmente, a resistência surge rapidamente quando estes são administrados como monoterapia ou na presença de uma supressão incompleta do vírus, sugerindo que a resistência pode ser

originada pela seleção de população de vírus mutante já existente no indivíduo. Algumas mutações associadas à resistência a NNRTIs podem diminuir a capacidade replicativa do vírus. Dois mecanismos foram propostos de modo a explicar tal fato: alterações na conformação na bolsa de ligação ao dNTP e a alterações de atividades da Rnase H (SHAFER, 2002).

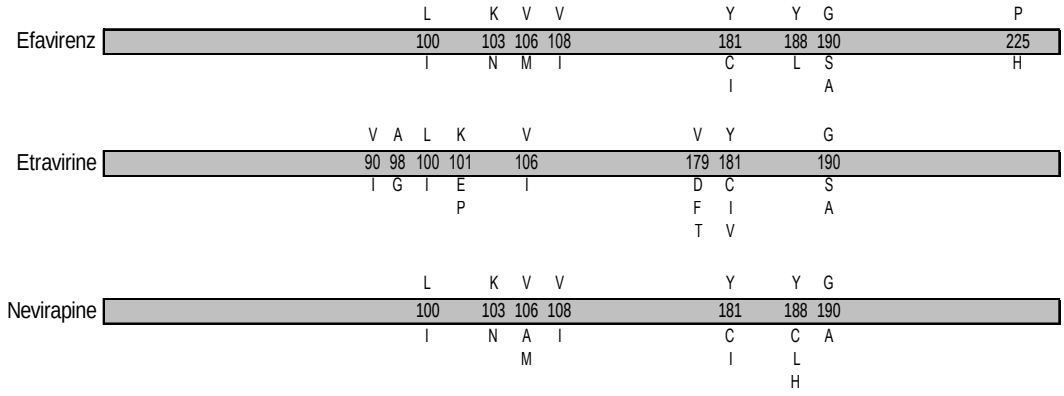

Figura 2.18: Mutações mais freqüentes encontradas no gene da transcriptase reversa (RT) associadas à resistência aos inibidores análogos não nucleosídeos (NNRTIs). Na figura, a letra superior refere-se ao aminoácido *will type*, a letra inferior ao aminoácido de substituição, a numeração em negrito representa as mutações primárias e a restante as mutações compensatórias (adpatada de JOHNSON *et al*., 2008).

2.4.1.3 Resistência aos Inibidores da Protease (IPs)

O mecanismo estrutural de resistência, aos inibidores da protease viral, influencia na habilidade do inibidor em se ligar efetivamente na enzima, já que estas mutações também afetam os resíduos da região "*flap*" e os da interface do dímero no qual consiste a enzima protease; enquanto que mutações compensatórias ocorrem principalmente nas redondezas do sítio ativo, recuperando a plasticidade e atividade enzimática (geralmente perdidas pela presença das mutações primárias). Normalmente, a resistência a IPs (figura 2.19) desenvolve-se gradualmente com a acumulação de múltiplas mutações primárias e compensatórias.

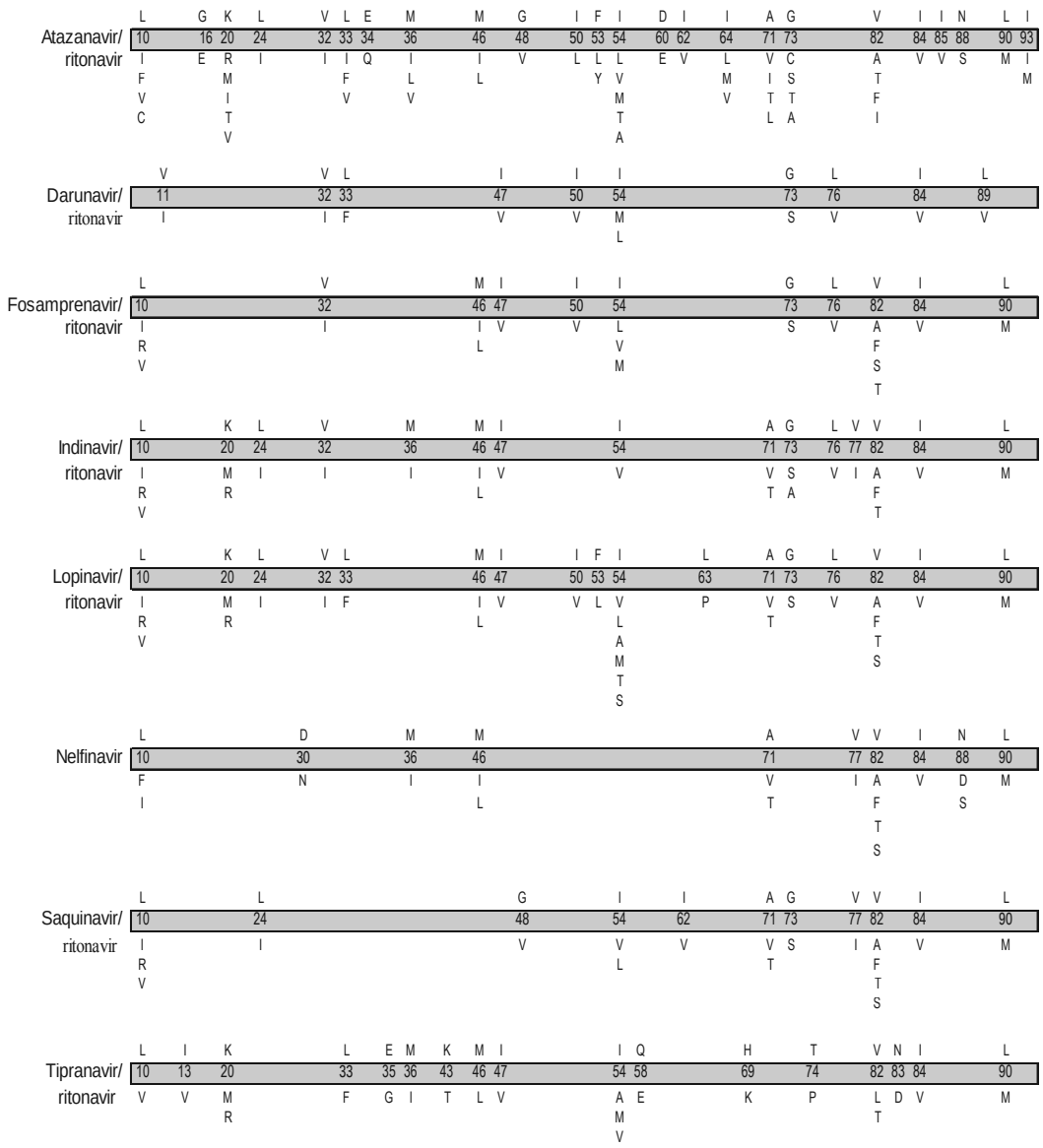

Figura 2.19: Mutações mais freqüentes encontradas no gene da protease associadas à resistência aos inibidores de protease. Na figura, a letra superior refere-se ao aminoácido *wild type*, a letra inferior ao aminoácido de substituição, a numeração em negrito as mutações primárias e a restante as mutações compensatórias (adpatada de JOHNSON *et* al., 2008).

2.4.1.4 Resistência aos Inibidores de Fusão

A resistência ao enfuvirtide (T20), inibidor de fusão aprovado pelo FDA, está associada primeiramente a mutações na região HR1 ("Heptad Repeat 1") do gene que codifica a *gp41* (figura 2.20). Além destas, mutações ou polimorfismos em outras regiões do envelope, como, por exemplo, na região HR2, pode também afetar a susceptibilidade a esta droga (JOHNSON *et al*., 2008).

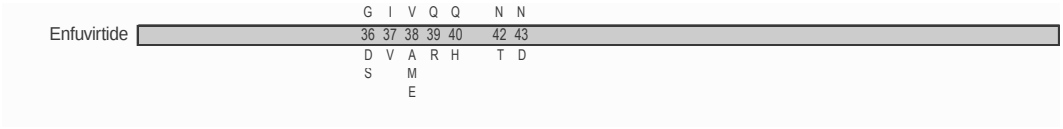

Figura 2.20: Mutações mais freqüentes encontradas no gene que codifica a *gp41* associada ao inibidor de fusão enfuvirtide. Na figura, a letra superior refere-se ao aminoácido *wild type*, a letra inferior ao aminoácido de substituição (adpatada de JOHNSON *et* al., 2008).

No estudo realizado por CARMONA *et al.* (2005), foram detectadas mutações naturais de resistência ao enfuvirtide, bem como polimorfismos na região HR1 da *gp41*, em pacientes que não haviam sido tratados com T20.

2.4.1.5 Resistência aos Inibidores de Integrase

Segundo JOHNSON *et al*. (2008), as mutações de resistência aos inibidores de integrase aparecem com freqüência como polimorfismos naturais (figura 2.21). Entretanto para o Raltegravir (RAL) a falha terapêutica está associada a duas mutações Q148H/K/R ou N155H.

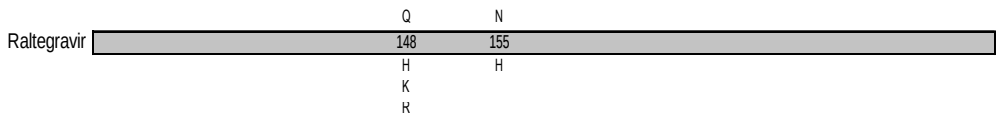

Figura 2.21: Mutações no gene da integrase associada com a resistência primária para o inibidor de integrase Raltegravir. Na figura, a letra superior refere-se ao aminoácido *wild type*, a letra inferior ao aminoácido de substituição, a numeração em negrito as mutações primárias para este inibidor (adpatada de JOHNSON *et* al., 2008).

2.4.2 Testes para Avaliação da Resistência aos Antiretrovirais.

O acúmulo de resistência às drogas antiretrovirais, em decorrência à falha terapêutica, leva a desafios de como melhorar o tratamento dos pacientes portadores do HIV-1, do mesmo modo que a saúde pública lida com a resistência aos antibióticos. Nos últimos anos têm sido desenvolvidas metodologias de modo a permitir avaliar fenotipicamente a susceptibilidade/resistência do HIV-1 a partir do perfil mutacional dos dados de genotipagem da seqüência do gene *pol*.

Estudos prospectivos têm demonstrado que pacientes cujos médicos têm acesso a dados sobre resistência, em particular dados genotípicos, respondem melhor à terapia do que pacientes cujos médicos não têm acesso aos mesmos dados (BAXTER, *et al*., 2000).

Dois tipos de teste estão sendo utilizados para verificar a presença de mutações de resistência em pacientes infectados pelo HIV-1, os testes de fenotipagem e genotipagem. Os testes fenotípicos medem a suscetibilidade as drogas (*in vitro*), utilizando uma porção da região *gag-pol*, amplificada por reação em cadeia da plomerase (PCR), a partir do RNA viral plasmático, estes testes têm a vantagem de avaliar com maior fidelidade à sensibilidade do vírus a um determinado fármaco. O material amplificado pode ser incorporado num vírus recombinante que não possui a região *pol*. Um fragmento deste vírus é utilizado de modo a infectar uma linhagem de células, levando à geração de vírus recombinantes. Estes vírus, contendo os genes de um HIV-1 resistente ou não, terão a sua suscetibilidade a drogas determinada em diferentes concentrações de fármaco (PETROPOULOS *et al*., 2000). A susceptibilidade é determinada pela concentração do fármaco necessário para inibir a replicação viral em 50% (IC50) ou em 90% (IC90).

Os testes de genotipagem, que identificam mutações no HIV-1 que são conhecidas por conferir resistência fenotípica aos fármacos, por meio de seqüenciamento automático de DNA, são mais comumente utilizados quando comparado com os testes de fenotipagem. Tal fato pode ser explicado em função da maior disponibilidade de testes genotípicos, menor custo e tempo de duração menor (BAXTER, *et al*., 2000).

Os testes de genotipagem envolvem extração do RNA viral, transcrição reversa para obtenção de um cDNA e amplificação por PCR de um fragmento do genoma com

mais de 1 Kb e podem ser realizados clonando-se ou não o material amplificado. Geralmente a clonagem é empregada nas pesquisas que visam encontrar respostas acerca da evolução da resistência do HIV-1 aos fármacos antiretrovirais. Já a genotipagem sem clonagem envolve sequenciamento direto do produto de PCR, sendo utilizada para finalidades clínicas, pois é mais rápida e exige menos recursos financeiros (SHAFER, 2002).

A interpretação dos resultados de genotipagem e fenotipagem do HIV-1 pode ser difícil, principalmente em função da existência de resistência-cruzada entre composto da mesma classe de fármacos. Entretanto, o teste de genotipagem tem se mostrado clinicamente útil em ensaios clínicos randomizados (BAXTER *et al*., 2000).

A resistência a drogas não é a única causa da falha terapêutica. A não aderência, o uso de regimes terapêuticos pouco potentes e fatores farmacocinéticos que diminuem os níveis de uma ou mais drogas no regime, também contribuem para a falha terapêutica. A incapacidade em se detectar mutações resistentes ao HIV-1 minoritárias também corrobora na limitação dos testes atualmente utilizados no estudo da resistência aos antiretrovirais, visto que nenhum destes testes é capaz de detectar as variantes resistentes minoritárias presentes abaixo dos 20% - 30% da população viral total (PETROPOULOS *et al*.,2000).

Neste capítulo foram introduzidos os fundamentos teóricos da Biologia Molecular, da partícula viral do HIV-1 e da terapia antiretroviral. Finalmente, foram descritos os testes para avaliação de resistência aos antiretrovirais, verificando as vantagens e desvantagens de cada um. No próximo capítulo serão apresentados os Algoritmos Genéticos, enfatizando sua aplicação na seleção de variáveis.

Capítulo 3

Algoritmos Genéticos

3.1 Introdução

A evolução pode ser compreendida como um processo de otimização, que não visa a perfeição como objetivo final, mas que é capaz de descobrir soluções altamente precisas e funcionais para um problema imposto por um ambiente a um organismo (MAYR, 1987). A otimização é à busca da melhor solução para um dado problema. Consiste em tentar várias soluções e utilizar a informação obtida nesse processo de modo a encontrar soluções cada vez mais eficientes. As técnicas de otimização e busca apresentam geralmente um espaço de busca, no qual estão localizadas todas as possíveis soluções do problema e uma função objetivo (aptidão), que é utilizada para avaliar as possíveis soluções, atribuindo a cada uma delas um valor.

Em termos matemáticos, a otimização consiste em procurar a solução ótima da função objetivo, ou seja, a solução que corresponde ao valor do ponto de máximo ou mínimo da função objetivo. Entretanto, existem funções que apresentam várias soluções ótimas locais, ou seja, soluções que otimizam o valor da função objetivo. Tais pontos são denominados ótimos locais, pois a função atinge, nesses pontos, valores ótimos em relação à vizinhança desses pontos. Apesar da melhor solução para esse problema estar no ponto em que a função atinge seu valor ótimo, denominado ótimo global. Em muitas funções, as técnicas de otimização (por exemplo, método do gradiente) não são capazes

de localizar o ponto de ótimo global em uma função que apresenta múltiplos pontos ótimos. Nesses casos, Algoritmos Genéticos (AGs) geralmente são capazes de encontrá-los. Os AGs descritos e analisados por HOLLAND (1975) e GOLDBERG (1989) são métodos de busca estocásticos, que simulam a evolução biológica natural não exigindo conhecimento a *priori* sobre como encontrar uma solução ótima para um problema, mas sim como reconhece-la como ótima.

A execução de um algoritmo genético pode ser vista como um conjunto de procedimentos. Inicialmente deve-se criar uma população em geral de forma aleatória, contendo um número *n* de indivíduos. Após a seleção dos indivíduos proporcionalmente a sua aptidão, os operadores genéticos probabilísticos são utilizados para a modificação dos indivíduos e criação de uma nova população. Este ciclo é repetido até que o critério de parada seja alcançado. O processo, partindo de uma população atual para a próxima população, constitui uma geração na execução do AG (figura 3.1)

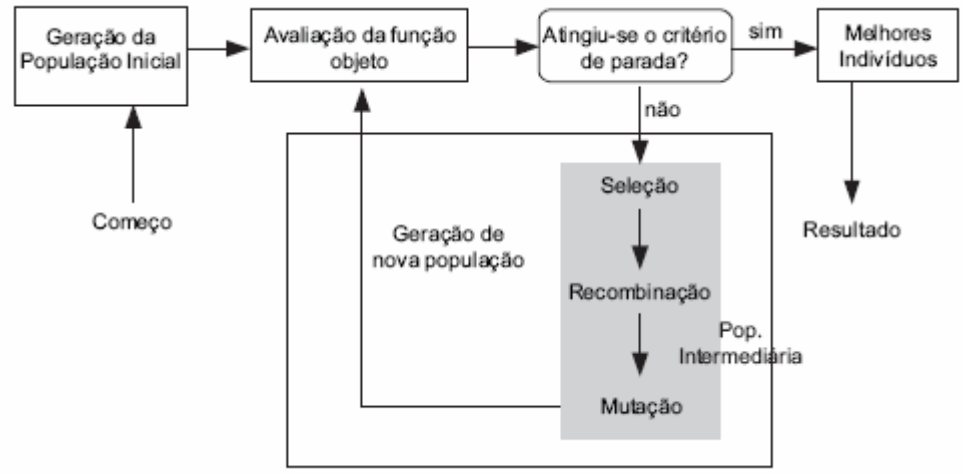

Figura 3.1: Estrutura de um Algoritmo Genético de uma única população, proposta por GOLDBERG (1989).

De modo a avaliar a aptidão de cada indivíduo na resolução de um problema, dois métodos podem ser aplicados. O primeiro baseado na ordenação e o segundo proporcional à aptidão (REEVES & ROWE, 2002). No primeiro, os indivíduos da população atual são ordenados de acordo com a sua avaliação e a probabilidade de seleção de cada indivíduo é uma função dessa ordem, ou seja, considerando-se N_{ind} o número de indivíduos na população atual, *Pos* a posição de um determinado indivíduo nessa população (o indivíduo de menor valor da função objeto tem *Pos* = 1, e o de

maior valor $Pos = N_{ind}$) e SP a pressão seletiva, a aptidão (Fit) atribuída a cada indivíduo é determinada como:

$$Fit(Pos) = 2 - SP + 2(SP - 1)\frac{Pos - 1}{N_{ind} - 1}, \tag{3.1}$$

no caso linear, são admissíveis, valores de SP entre 1 e 2. Para o caso não-linear se introduz uma distribuição não-linear, podendo este ser calculado por:

$$Fit(Pos) = \frac{N_{ind} X^{Pos-1}}{\sum_{i=1}^{N_{ind}} X^{i-1}}, \tag{3.2}$$

onde X é obtido pelas raízes reais do polinômio:

$$(SP - 1)X^{N_{ind}-1} + SPX^{N_{ind}-2} + \ldots + SPX + SP = 0, \tag{3.3}$$

admitindo valores de SP entre 1 e $(N_{ind} - 2)$.

3.2 População

A população é constituída por um conjunto de indivíduos que representam possíveis soluções para um determinado problema. Seu tamanho afeta diretamente o desempenho global e a eficiência dos resultados. Populações muito pequenas tendem a perder a diversidade genética rapidamente e podem não obter uma boa solução, visto que a busca realizada pelo AG cobre uma pequena parte do espaço de soluções do problema. No caso de uma população muito grande o custo computacional do AG tenderá a ser muito alto, implicando em um processo muito lento; principalmente se o cálculo da função de aptidão for complexo, o que frequentemente acontece na resolução de problemas complexos (não lineares).

Os indivíduos de uma população podem ser representados por diversas metodologias: vetores de números inteiros ou reais e seqüências de símbolos, (ESHELMAN, 2000). Entretanto a representação típica consiste em uma seqüência de *bits* de comprimento fixo. Onde a concatenação dessas seqüências é denominada de *cromossomos* e o valor em cada *bit* nesta seqüência é denominado *alelo*.

3.3 Métodos de Seleção

Segundo BENTLEY (2002), a seleção é o componente do processo evolucionário que determinará os indivíduos "vencedores" e "perdedores" na luta pela sobrevivência, desempenhando um papel fundamental na evolução, podendo ser executada por meio de um dos algoritmos: seleção proporcional, seleção truncada e seleção por torneio.

A seleção proporcional consiste em selecionar os indivíduos com probabilidade proporcional à sua aptidão. Conceitualmente, este método é equivalente a dividir uma roleta em *n* partes, sendo *n* o número de indivíduos da população. Cada uma das partes da roleta é proporcional à aptidão do indivíduo associado àquela parte. A roleta é girada *n* vezes, e a cada uma delas o indivíduo indicado pelo ponteiro é selecionado e inserido na nova população. O processo é repetido até que o número desejado de indivíduos seja obtido.

Na seleção truncada apenas os melhores indivíduos, em função de suas aptidões, são selecionados da população atual. Esses indivíduos selecionados produzem descendentes de forma aleatória e uniforme. O parâmetro do operador de seleção truncada é o ponto de corte para a seleção, ponto esse que indica a proporção da população a ser selecionada como população intermediária, e assume valores entre 50% e 10%. Indivíduos abaixo do ponto de corte são excluídos da população intermediária (BLICKLE e THIELE, 1995).

Na seleção por torneio um subconjunto de indivíduos é selecionado aleatoriamente na população atual, e o melhor indivíduo desse subconjunto é selecionado para a população intermediária. Este processo é repetido com freqüência igual ao número de indivíduos a ser escolhido para a população intermediária (BANZHAF *et al*., 1998).

3.4 Operadores Genéticos

O princípio básico dos operadores genéticos é transformar a população, possibilitando modificações nos indivíduos da população de modo a possibilitar a diversificação e perpetuação das características adaptativas adquiridas em gerações anteriores.

Todos os AGs funcionam associando o operador de seleção a um mecanismo (operador) que combine informação dos pares de indivíduos da população intermediária (recombinação-*crossover*) e a outro que produza variação (mutação) nos indivíduos recombinados.

O operador genético *crossover* permite a troca de informação entre soluções diferentes. Combinando as características de dois indivíduos pais para formar dois indivíduos filhos por meio da troca de segmentos correspondentes dos indivíduos pais. Dentre os tipos de operadores de *crossover* existentes, podemos referenciar o *crossover* de ponto único, o *crossover* de dois pontos e *crossover* uniforme.

No *crossover* de ponto único (figura 3.2), uma única posição de troca $k \in \{1,2,...,N_{\text{var}} -1\}$, onde N_{var} é o número de bits de um indivíduo, é selecionada de forma aleatória e uniforme, dada uma razão de recombinação, que varia normalmente no intervalo [0,6,1,0]. A seguir, os *bits* a partir desse ponto são trocados entre os indivíduos, produzindo dois indivíduos.

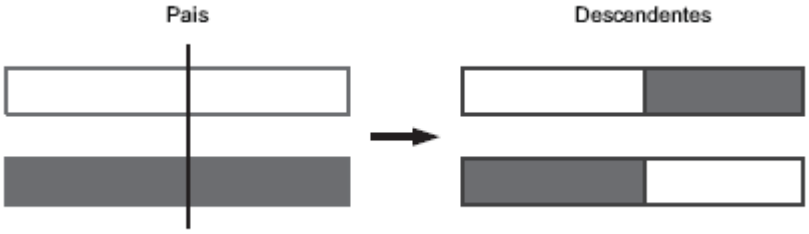

Figura 3.2: *Crossover* de ponto único

Já o *crossover* de dois pontos, como próprio nome diz, duas posições de troca são determinadas de forma aleatória e uniforme e os bits entre esses pontos são trocados entre os indivíduos, gerando também dois indivíduos.

O operador *crossover* uniforme leva o número de pontos de *crossover* ao extremo, utilizando uma decisão aleatória de fazer ou não a troca de informação *bit*-a-*bit* entre os indivíduos selecionados (SYSWERD, 1989). Para cada posição, o indivíduo que contribui com sua variável (0 ou 1) ao indivíduo recombinando é associado de modo uniforme e aleatória, com a seguinte probabilidade:

$$Var_{i\,recomb} = Var_{i_{Ind}1}\,\alpha_i + Var_{i_{Ind}2}\,(1-\alpha_i), \text{ para } i \in \{1,2,...,N_{\text{var}}\}, \qquad (3.4)$$

onde $\alpha_i = \{0,1\}$ é escolhido de forma uniforme e aleatória.

Uma vez selecionados os indivíduos da população intermediária, e recombinados pelos operadores apresentados, o AG impõe aos indivíduos dessa população intermediária um mecanismo de variação, de modo a tornar mais diverso o universo de busca e, conseqüentemente, menor a probabilidade de convergência precoce para pontos de mínimos locais. O mecanismo de variação mais conhecido é a mutação.

Este operador modifica de forma aleatória alguma característica do indivíduo sobre o qual é aplicado (figura 3.3). Estas modificações são necessárias para a introdução e manutenção da diversidade genética da população, ou seja, o operador de mutação assegura que a probabilidade de se atingir qualquer ponto do espaço de busca seja diferente de zero. Segundo GOLBERG (1989), normalmente este operador é aplicado com probabilidade muito pequena, de modo a evitar que a busca realizada pelo AG se torne aleatória.

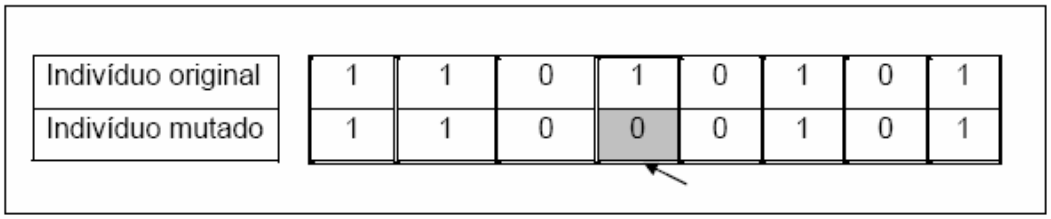

Figura 3.3: Exemplo de operador de mutação

Apesar de muitos AGs usarem conjuntamente a mutação e a recombinação (*crossover*), para muitos problemas de otimização, a utilização de AGs utilizando mutação na ausência de *crossover* pode ser bastante eficaz (MATHIAS e WHITLET, 1994).

Uma vez que os indivíduos da população intermediária foram produzidos por seleção, recombinação e mutação dos indivíduos da população atual cabe agora decidir quais indivíduos comporão a população atual da nova geração. Isso se dá por operadores conhecidos como operadores re-inserção, que atuam inserindo ou removendo indivíduos das populações atuais e intermediárias, de modo a compor a população dessa nova geração.

Há diferentes formas de re-inserção. Pode-se, por exemplo, gerar um número de indivíduos na população intermediária igual ao número de indivíduos na população atual e substituí-los todos (re-inserção simples); pode-se gerar um número menor de indivíduos na população intermediária e nela inserir indivíduos da população atual, selecionados de forma uniforme e aleatória (re-inserção aleatória); pode-se gerar um número menor de indivíduos na população intermediária e nela simplesmente inserir

indivíduos da população atual de maior aptidão (re-inserção elitista); ou ainda, pode-se gerar um número de indivíduos na população intermediária maior do que o da população atual, e então inserir apenas os indivíduos da população intermediária de maior aptidão (re-inserção baseada na aptidão).

A re-inserção elitista, quando combinada com a re-inserção baseada na aptidão, previne a perda de informação, ou seja, previne que indivíduos muito bons sejam substituídos sem que tenham gerado descendentes ainda melhores (figura 3.4). A cada geração, um dado número de indivíduos menos apto é substituído pelo mesmo número de indivíduos da população intermediária de melhor aptidão.

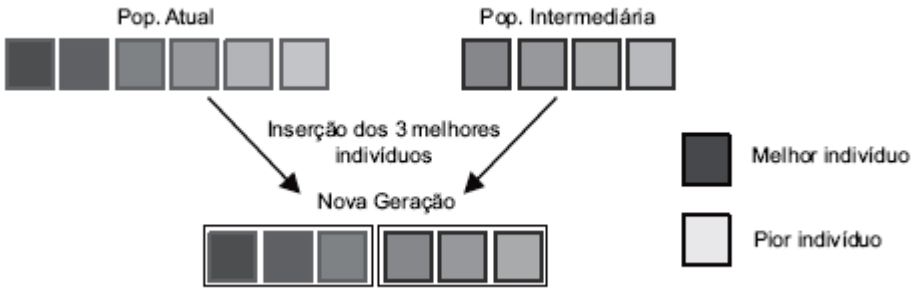

Figura 3.4: Esquema de re-inserção elitista combinada com a re-inserção baseada na aptidão.

Para a seleção de indivíduos a serem substituídos da população atual, assim como, da população intermediária, há vários esquemas possíveis. Por exemplo, pode-se re-inserir todos os indivíduos da população atual, substituindo-se os indivíduos da população intermediária dentro da vizinhança local de forma uniforme e aleatória; pode-se inserir todos os indivíduos da população intermediaria, em substituição aos indivíduos menos aptos da população intermediária pertencente à vizinhança; pode-se substituir os indivíduos da população intermediária pertencente à vizinhança; pode-se substituir os indivíduos da população intermediaria com aptidão menor que o mais fraco da vizinhança, substituindo-os por indivíduos da população atual de forma uniforme e aleatória; ou ainda, simplesmente inserir indivíduos da população intermediária melhores que os indivíduos da população atual pertencentes à vizinhança local.

Uma vez gerada a nova população atual (nova geração), cada indivíduo é avaliado. Calcula-se o valor da função objeto (função de avaliação), que é o propósito final da otimização. Se o critério de parada/otimização for atendido interrompe-se o algoritmo, caso contrário, aplicam-se novamente os operadores de seleção, mutação, recombinação e re-inserção, até que esse critério seja alcançado (figura 3.1).

3.5 Seleção de Variáveis por Algoritmos Genéticos

Intuitivamente, quanto maior o número de atributos em uma base de dados, maior será o poder discriminatório do classificador, bem como a facilidade na extração de modelos de conhecimentos da base. No mundo real encontramos evidências de que nem sempre isso é verdade absoluta, visto que vários métodos de indução sofrem da maldição da dimensionalidade, ou seja, o tempo computacional aumenta em função do número de variáveis presentes (MATHIAS e WHITLET, 1994).

Métodos automáticos de seleção de variáveis são empregados quando desejamos selecionar um subconjunto de variáveis que possa reduzir a dimensionalidade de forma que ocorra a menor queda possível no poder de distinção das classes por um classificador no espaço de características. Uma conseqüência da aplicação de um bom algoritmo de seleção de variáveis é a redução do número necessário de amostras de treinamento para obter-se bons resultados com um classificador, ou seja, a redução do problema da dimensionalidade.

Nos últimos anos, diversas aplicações de AGs na seleção de características foram relatadas na literatura. FILHO e POPPI (2002), relatam a aplicação de AG na seleção de variáveis na seleção de variáveis em dados espectroscópicos no infravermelho, em problemas onde os espectros dos analíticos estudados (glicose, maltose e frutose) possuem um alto grau de similaridade. O algoritmo genético foi iniciado com uma população inicial de 100 cromossomos, com um número máximo de gerações igual a 100, a taxa de *crossover* de 0,9 e probabilidade de mutação de 0,01, os critérios de parada utilizados foram o erro absoluto de 1% e o número de variáveis selecionadas igual a 10. De forma a verificar o desempenho do AG foi realizado uma comparação entre os resultados e os valores obtidos pela seleção de variáveis utilizando o método dos mínimos quadrados e regressão linear múltipla.

Para o método dos mínimos quadrados foi utilizado um número de componentes principais igual a quatro. Os resultados mostram que o AG é uma ferramenta poderosa no que tange à robustez dos modelos propostos, visto que os erros não apresentaram discrepância elevada em nenhum dos modelos utilizados para a comparação neste estudo. Os autores ressaltam ainda que apesar do AG ser mais complexo que o método dos mínimos quadrados, sua utilização é mais simples e pode ser realizada sem a

intervenção ou ajuda do usuário, fato este que confere uma grande vantagem para ser utilizado quando existe pouco ou nenhum conhecimento sobre o modelo em estudo.

SOFGE (2002) relata a utilização do AG na seleção de variáveis para a aplicação em redes neurais MLP em processos químicos industrial. Para tal utiliza 200 dados oriundos de 20 sensores, onde os dados apresentam ruídos e características irrelevantes para o modelo, totalizando 20 variáveis. O objetivo do modelo proposto era encontrar um conjunto de variáveis de entrada para o modelo rede neural treinada com o algoritmo de *Levenberg-Marquardt* que minimize a soma dos erros médio quadrático no conjunto de validação cruzada. A população foi inicializada de forma aleatória, com taxa de mutação de 0,1. Os resultados obtidos pelo modelo neural proposto, com o conjunto de entrada definido pelas 11 variáveis selecionadas pelo AG no conjunto de 1000-2000 simulações foi capaz de minimizar o valor do erro médio quadrático no conjunto de validação. Fato este que possibilitou o autor concluir que o AG é uma ferramenta promissora na seleção de variáveis, podendo ser aplicada em inúmeras aplicações onde a as variáveis de entrada apresentam altos valores de ruído, dificultando a seleção dessas variáveis.

NEUMANN *et al*. (2004) apresentam a utilização de AG na seleção de variáveis na identificação de fármacos anti-inflamatórios a partir de dados de infravermelho, visando determinar e otimizar o grau de similaridade entre as amostras de medicamentos manipulados contendo diclofenaco de sódio ou diclofenaco de potássio de três diferentes estabelecimento de manipulação. Primeiramente, selecionaram um conjunto de regiões no espectro (onda) que apresentava melhor correlação em função da quantificação do princípio ativo para os fármacos em estudo. A esse conjunto de dados aplicou-se o AG e os resultados obtidos na construção dos dendogramas com as variáveis selecionadas possibilitou não apenas uma melhora na similaridade entre as duplicatas, mas também uma maior discriminação entre os diferentes grupos formados, facilitando a interpretação dos resultados. Os autores afirmam também que a metodologia empregada pode ser uma excelente alternativa para a otimização no estudo quimiométrico de fármacos.

TAN *et al* (2008) desenvolveram uma estrutura baseada em algoritmo genético para a seleção de um subconjunto de variáveis que combina várias técnicas existentes para a seleção de variáveis, visando selecionar um subconjunto pequeno de variáveis que proporcione um alto índice classificação correta quando comparada a outras técnicas existentes. Para avaliar o método os autores comparam os resultados obtidos

com três modelos diferentes de seleção de variáveis baseado na entropia (DASH e LIU, 1999), estatística-T e SVM recursiva na eliminação de característica – RFE (GUYON *et al*., 2002). Aplicados em dois conjuntos de dados de microarranjos (dados de câncer de cólon e prostata). Os resultados obtidos por TAN *et al* (2008) mostram que o modelo proposto é robusto e eficaz na determinação em um pequeno subconjunto de variáveis, visto que foi capaz de selecionar características importantes que não foram selecionadas pelos modelos individuais utilizados, tornando-se uma metodologia bastante atraente no problema de seleção de característica.

LI *et al*.(2001) desenvolveram um método, denominado GA/*k*NN, onde o AG foi usado para a seleção de um subconjunto de variáveis de uma matriz de expressão gênica. As variáveis foram classificadas utilizando um classificador de vizinhos próximos (*k*NN), com $k = 3$. Ao invés de utilizar as variáveis selecionadas em uma única simulação do AG, o método proposto seleciona as variáveis mais freqüentes de um conjunto de 20000-40000 simulações e forma o classificador *k*NN final baseado nessas variáveis usando como métrica a distância Euclidiana. A função objeto, utilizada foi a razão de amostras corretamente classificadas no conjunto de treino. Como os AGs são algoritmos estocásticos, simulações diferentes irão normalmente gerar diferentes conjuntos de variáveis selecionadas. Entretanto, a coerência da abordagem reside no fato de variáveis mais capazes de gerar classificações corretas irão aparecer no subconjunto selecionado com maior freqüência e são mais adequadas à inclusão no classificador final.

O método foi utilizado ao conjunto de leucemia aguda, cujo objetivo era diferenciar entre duas classes de leucemia aguda (linfóide e mielóide). O resultado demonstrou que o classificador final, baseado nas 50 variáveis, mais freqüente selecionadas pelo AG em 40000 simulações, pôde classificar corretamente 33 das 34 amostras de teste. Resultado esse que levam os autores a concluírem que a aplicação do AG na seleção de variáveis é bastante promissor.

Neste capítulo foram apresentados os AGs como técnica de seleção de variáveis. Foram detalhados os principais métodos de seleção e operadores genéticos. Finalmente foram descritos alguns trabalhos com a aplicação dos AGs na seleção de variáveis. Todos esses conceitos serão usados ao problema de previsão de resistência aos inibidores da protease do HIV-1, com o classificador de *Kernel* Discriminante de Fisher, o qual será apresentado no próximo capítulo.

Capítulo 4

Classificador não linear de Fisher

4.1 Introdução

Provavelmente o mais conhecido exemplo de classificador linear é o classificador de Fisher (FISHER, 1938). A idéia de Fisher consiste basicamente em determinar a direção $\mathbf{w}$[1] que maximize a variância entre as classes (inter-classes), enquanto minimiza a variabilidade entre os elementos da mesma classe (intra-classes). A figura (4.1) ilustra a idéia de Fisher.

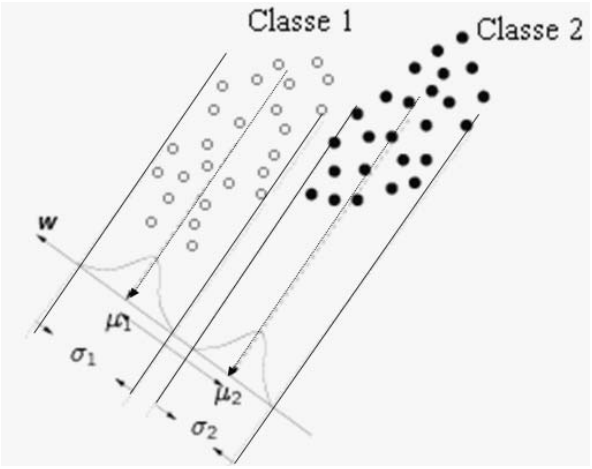

Figura 4.1: Ilustra o classificador linear de Fisher para 2 classes. Cujo objetivo é procurar uma direção $\mathbf{w}$, de modo que a diferença entre as médias projetadas na

[1] Variáveis vetoriais serão denotadas em negrito.

direção $(\mu_1 \; e \; \mu_2)$ seja a maior possível, enquanto a média da variância aproximada entre os elementos das classes $(\sigma_1 \; e \; \sigma_2)$ seja pequena.

Matematicamente, temos: Seja χ o espaço de observação, ou seja, $\chi \subseteq R^N$ e y o conjunto de possíveis rótulos, definidos por y = {+1, -1}. Além disso, seja $X = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\} \subseteq \chi \times y$ o conjunto de treinamento de tamanho n e $X_1 = \{(\mathbf{x}, y) \in X \mid y = 1\}$ e $X_2 = \{(\mathbf{x}, y) \in X \mid y = -1\}$ a partir de duas classes de tamanho $n_i = |X_i|$, definimos $\mathbf{m}_1$ e $\mathbf{m}_2$ a média empírica dos exemplos das classes.

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{x \in z_i} \mathbf{x} \tag{4.1}$$

De forma similar podemos calcular a média dos dados projetados sobre a direção do vetor $\mathbf{w}$, por:

$$\mu_i = \frac{1}{n_i} \sum_{x \in z_i} \mathbf{w}^T \mathbf{x} \tag{4.2}$$

como $\frac{1}{n_i} \sum_{x \in z_i} \mathbf{x} = \mathbf{m}_i$, a equação (4.2) pode ser reescrita, resultando em :

$$\mu_i = \mathbf{w}^T \mathbf{m}_i \tag{4.3}$$

ou seja, a média μ_i das projeções são as projeções das média $\mathbf{m}_i$. As variâncias σ_1 e σ_2 dos dados projetados podem ser expressos como:

$$\sigma_i = \sum_{x \in z_i} \left(\mathbf{w}^T \mathbf{x} \; - \; \mu_i \right)^2 \tag{4.4}$$

A maximização da variância entre as classes e a minimização da variância entre os elementos de uma mesma classe, pode ser obtida maximizando,

$$J(\mathbf{w}) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1 + \sigma_2} \tag{4.5}$$

na direção de **w**, de modo que a relação entre a razão da variância inter classes e a variância intra classes seja máxima. Substituindo os valores de μ_i e σ_i pelas equações (4.3) e (4.4), a equação (4.5) pode ser reescrita como:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \tag{4.6}$$

onde S_B e S_W são matrizes que definem o espalhamento entre as classes e dentro das classes, respectivamente representados por:

$$S_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \quad \text{e} \quad S_W = \sum_{i=1,2} \sum_{x \in Z_i} (\mathbf{x} - \mathbf{m}_i)^2 \tag{4.7}$$

A equação (4.6) é referenciada como coeficiente de *Rayleigh*. Coeficiente este que mede a separabilidade entre classes por meio de duas quantidades: a dispersão entre as diferentes classes (numerador da equação 4.6) e a dispersão dentro das classes (denominador de 4.6). A dispersão entre as classes indica quão distante estão às médias das projeções dos dados de cada classe, que para uma maior separabilidade, deve ser maximizada. A dispersão dentro das classes, por sua vez, mede a variabilidade entre as projeções em uma mesma classe e deve ser minimizada.

Uma importante propriedade do classificador linear de Fisher é que a equação (4.6) tem solução global, entretanto não necessariamente única, e que esta solução global ótima de **w**, pode ser obtida resolvendo um problema de auto-valor e auto vetor.

$$S_B \mathbf{w} = \lambda S_W \mathbf{w} \tag{4.8}$$

Dividindo ambos os lados da equação (4.8) pela matriz de espalhamento intra classe S_W, temos:

$$\frac{S_B \mathbf{w}}{S_W} = \frac{\lambda S_W \mathbf{w}}{S_W} \quad \rightarrow \quad {S_W}^{-1} S_B \mathbf{w} = \lambda \mathbf{w} \tag{4.9}$$

A equação (4.9) pode ser vista como uma equação de auto-vetores. Se S_W é uma matriz não singular, ou seja, possui determinante não nulo, então essa taxa é maximizada quando os vetores colunas da matriz de transformação são os auto-vetores de ${S_W}^{-1} S_B$ (FISHER, 1938).

Apesar do classificador linear de Fisher apresentar resultados satisfatórios quando aplicado em problemas linearmente separáveis, o mesmo não é verificado quando os dados não apresentam essa característica. A fim de contornar esse problema MIKA *et al* (1999), propuseram uma modificação denominada truque do *kernel* (*kernel trick*) a fim de possibilitar sua utilização em problemas não linearmente separáveis.

4.2 Classificador de Kernel Discriminante de Fisher

Os métodos baseados na teoria de *kernel* provocaram uma verdadeira revolução nos algoritmos da teoria de aprendizado estatístico supervisionado e não supervisionado por possibilitar a criação de versões não lineares dos algoritmos clássicos lineares. A idéia do classificador não linear de Fisher é solucionar o problema do discriminante linear de Fisher no espaço de característica baseado na matriz de *kernel*.

Considere que **x** represente um vetor retirado do espaço de entrada, que é assumido como tendo dimensão n_0. Considere também que $\{\varphi_j(\mathbf{x})\}_{j=1}^{n_i}$ represente um conjunto de transformações não-lineares do espaço de entrada para o espaço de características n_1 é a dimensão do espaço de características e $\varphi_j(\mathbf{x})$ seja definido a *priori* para todo j. Dado este conjunto de transformações não-lineares, podemos definir um hiperplano como a superfície de decisão caracterizado por:

$$\sum_{j=0}^{n_i} w_j \varphi_j(\mathbf{x}) + b = 0 \qquad (4.10)$$

onde w_j representa um conjunto de pesos lineares conectando o espaço de características com o espaço de saída, b é um fator de ajuste (*bias*) e $\varphi_j(\mathbf{x})$ representa a entrada fornecida ao peso w_j através do espaço de características.

Considere também que o vetor $\varphi_j = \left[\varphi_0(\mathbf{x}), \varphi_1(\mathbf{x}), ..., \varphi_{n_1}(\mathbf{x})\right]^T$ e onde por definição, $\varphi_0(\mathbf{x}) = 1 \ \forall \ \mathbf{x}$. Logo o vetor $\varphi(\mathbf{x})$ representa a "imagem" induzida no espaço de características pelo vetor de entrada **x** (HAYKIN, 2001). Assim, em termos desta imagem, podemos definir a superfície de decisão calculada no espaço de características como:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \varphi^T(\mathbf{x_i}) \varphi(\mathbf{x}) = 0 \qquad (4.11)$$

onde $\varphi^T(\mathbf{x_i})\varphi(\mathbf{x})$ representa o produto interno de dois vetores induzidos no espaço de características pelo vetor de entrada **x** e o padrão de entrada $\mathbf{x}_i$ relativo ao i-ésimo exemplo, α_i os multiplicadores de *Lagrange* e $\varphi(\mathbf{x}_i)$ o vetor de características. Logo podemos definir o *kernel* do produto interno ($K(\mathbf{x},\mathbf{x}_i)$), por

$$K(\mathbf{x},\mathbf{x}_i) = \varphi^T(\mathbf{x})\varphi(\mathbf{x}_i),$$

$$= \sum_{j=0}^{n_i} \varphi_j(\mathbf{x})\varphi_j(\mathbf{x}_i) \quad \text{para } i = 1,2,...,n \qquad (4.12)$$

Desta definição podemos concluir que as funções φ devem pertencer a um domínio em que seja possível o cálculo de produtos internos, ou seja, funções contínuas e simétricas. Atendendo às condições do Teorema de Mercer (COURANT e HILBERT 1953), os núcleos devem ser matrizes positivamente definidas. Logo, a matriz *K*, onde $K_j = K(x,x_j) \ \forall\, j = 1,...,n$, deve ter autovalores maiores que zero.

Apesar das restrições, muitas funções candidatas podem ser aplicadas como funções de núcleo. As mais comumente utilizadas são as funções polinomiais de grau *p*, RBF (Função de Base Radial) Gaussiana e *perceptron* de múltiplas camadas (tabela 4.1).

Tabela 4.1: Resumo das principais funções de núcleo utilizadas

Tipo de Núcleo	Função $K(x,x_j)$	Comentários
polinomial de grau *p*	$\left[\left(x^T \times x_j\right)+1\right]^p$	A potência *p* deve ser especificada pelo usuário.
RBF Gaussiana	$exp\left(-\frac{1}{2\sigma^2}\left\|x-x_j\right\|^2\right)$	A amplitude σ^2 é especificada pelo usuário.
perceptron múltiplas camadas	$tanh\left(\beta_0 x \cdot x_j + \beta_1\right)$	Utilizado somente para alguns valores de β_0 e β_1.

De acordo com a tabela 4.1, podemos destacar:

- Para os núcleos polinomiais, os mapeamentos φ também são funções polinomiais com complexidade crescente conforme o expoente *p* aumenta.
- O *kernel* Gaussiano equivale a um espaço de características de dimensão infinita. Podendo-se afirmar que a maioria das formas de mapeamento existentes podem ser implementadas por esta função em particular.

A utilização de função de *kernel* traz alguns benefícios, dentre os quais podemos citar: a não necessidade de se conhecer diretamente o mapeamento φ, o mapeamento φ é

computado implicitamente, a dimensão do espaço de característica não necessariamente afeta o desempenho computacional, as propriedades de representações de núcleo são auto contidas e podem ser usadas com diferentes teorias de aprendizagem e o método de núcleo dissocia os algoritmos e a teoria de aprendizagem das especificidades da área de aplicação, que deve ser codificada no projeto de uma função de núcleo apropriada.

Com isso, todo algoritmo linear que utiliza somente produtos escalares pode implicitamente ser executado em um espaço de característica *X* (com dimensão potencialmente alta), isto é, pode-se construir uma versão não linear de um algoritmo linear (MIKA *et al*., 2001).

O Classificador de *Kernel* Discriminante de Fisher (*Kernel Fisher Discriminant* – KDF) (MIKA *et al*., 1999), generaliza o algoritmo do classificador linear de Fisher para o caso não-linear, através do artifício do uso das funções de *Kernel*, em uma estratégica semelhante à utilizada pelas Máquinas de Vetores de Suportes (SVMs) e pelo algoritmo de *Perceptron* denominado VP (*Voted-Perceptron*). Logo, as equações 4.1, 4.6 e 4.7 podem ser então reescritas como:

$$\mathbf{m}_i^{\mathbf{\Phi}} = \frac{1}{n_i}\sum_{j=1}^{n_i}\Phi\left(\mathbf{x}_j^i\right) \tag{4.13}$$

$$J\left(\mathbf{w}\right) = \frac{\mathbf{w}^T S_B^{\mathbf{\Phi}}\mathbf{w}}{\mathbf{w}^T S_w^{\mathbf{\Phi}}\mathbf{w}} \tag{4.14}$$

$$S_B = (\mathbf{m}_1^{\mathbf{\Phi}} - \mathbf{m}_2^{\mathbf{\Phi}})(\mathbf{m}_1^{\mathbf{\Phi}} - \mathbf{m}_2^{\mathbf{\Phi}})^T \tag{4.15}$$

$$S_W = \sum_{i=1,2}\sum_{x\in Z_i}(\mathbf{\Phi}\left(\mathbf{x}\right) - \mathbf{m}_\mathbf{i}^{\mathbf{\Phi}})\ (\mathbf{\Phi}\left(\mathbf{x}\right) - \mathbf{m}_\mathbf{i}^{\mathbf{\Phi}})^T \tag{4.16}$$

A utilização de funções *kernel* sobre este problema requer que as equações 4.13 a 4.16 sejam ainda reformuladas em função de produtos internos entre os dados. Seja *F* o espaço de características dado pelo mapeamento **Φ**. A teoria de *Kernel* afirma que a solução $\mathbf{w} \in F$ deve ser definida em função dos exemplos de treinamento mapeados neste espaço, como exemplificado na equação 4.17.

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i \mathbf{\Phi}(\mathbf{x}_i) \tag{4.17}$$

Utilizando manipulações algébricas, o problema de otimização do KDF no espaço de características pode ser então escrito como:

$$J(\alpha) = \frac{\alpha^T \mathbf{M} \alpha}{\alpha^T \mathbf{N} \alpha} \tag{4.18}$$

onde, $\mathbf{M} = (\mathbf{M}_1 - \mathbf{M}_2)(\mathbf{M}_1 - \mathbf{M}_2)^T$, com $(\mathbf{M}_i)_j = \frac{1}{n_i} \sum_{k=j}^{n_i} K(x_j, x_k^i)$ e $\mathbf{N} = \sum_{j=1,2} K_j (\mathbf{I} - \mathbf{1}_{n_j}) K_j^T$, onde $(K_j)_{mn} = K(x_n, x_m^j)$ é a matriz de *kernel* para a classe *j*, **I** a matriz identidade e $\mathbf{1}_{n_j}$ uma matriz com entradas $1/n_j$.

Na definição acima existe um problema decorrente do fato da dimensão de *F* ser igual ou superior ao número de exemplos *n*, tornando-o mal condicionado. De modo a solucionar esse problema, uma regularização é realizada somando-se um múltiplo da matriz identidade a **N**.

$$\mathbf{N}_\mu = \mathbf{N} + \mu \mathbf{I} \tag{4.19}$$

Resolvido o problema de otimização do KDF, a projeção de um novo padrão **x** em **w** é obtida através da equação 4.20.

$$(\mathbf{w} \cdot \Phi(x)) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i - \mathbf{x}) \tag{4.20}$$

Para utilizar esta projeção na classificação do novo padrão, devemos estabelecer um limiar que será aplicado na separação entre as classes. Este limiar pode ser obtido pela média das projeções médias das duas classes ou pelo treinamento.

O KDF pode ser solucionado encontrando-se o autovetor α com maior autovalor λ no sistema $\mathbf{M}\alpha = \lambda \mathbf{N}\alpha$ ou calculando-se $\alpha = \mathbf{N}^{-1}(\mathbf{M}_2 - \mathbf{M}_1)$ (KODIPAKA *et al*., 2007). Os métodos porém são computacionalmente caros quando temos um grande conjunto de dados, sendo factíveis apenas para um conjunto de *n* pequeno.

Para solucionar, MIKA *et al*. (2001) propuseram uma reformulação do KDF como um problema de otimização quadrática que realiza a minimização da variância dos dados projetados enquanto maximizam a distância entre a saída média de cada classe, objetivos primordiais do classificador linear de Fisher.

$$\min_{\alpha,b,\xi} \|\xi\|^2 + CP(\alpha) \qquad (4.21)$$

com as restrições:

$$K\alpha + \mathbf{1}b = y + \xi$$

$$\mathbf{1}_{\xi}^{T}\xi = 0 \text{, para } K = 1, 2.$$

onde α, $\xi \in \Re^n$ e b, $C \in \Re$. O parâmetro P representa o termo de regularização e $(\mathbf{1}_k)_i$, é igual a 1 caso y_i pertença à classe k e 0 caso contrário. A primeira restrição pode ser lida como $(\mathbf{w} \cdot \mathbf{x}_i) + b = y_i + \xi_i$, i = 1,...,n e "impõe" que a saída de cada exemplo seja uma classe desejada. O fator $\|\xi\|^2$ representa a minimização da variância dos erros cometidos, enquanto as restrições $\mathbf{1}_{\xi}^{T}\xi = 0$ asseguram que a média das saídas de cada classe seja o rótulo desejado.

Essa formulação possibilitou a derivação de algoritmos de treinamento mais eficientes para o KDF. Um desses métodos denominado *sparse-greedy* (MIKA *et al.*, 2001) permitiu a construção da solução do problema iterativamente. Começando por uma solução vazia, adiciona-se a cada iteração um novo padrão na equação 4.17. A escolha desse padrão se dá por meio de heurísticas, onde o padrão escolhido é aquele que proporciona o maior decréscimo na função objetivo. O algoritmo é interrompido quando o valor da função objetivo se torna menor que um limiar pré-estabelecido.

Pode se considerar que o KDF maximiza a margem média de separação entre as classes, definida como a distância média entre exemplos de classes diferentes (JAIN e ZONGRKER, 1997). Apesar de não poder definir o conceito de margem explicitamente para essa técnica, o KDF apresenta desempenho equivalente aos das SVMs, apresentando em geral, boa capacidade de generalização, ou seja, produzir respostas para padrões de entrada que são similares, mas não idênticos aos padrões que o classificador já conhece.

4.3 Classificação por Kernel Discriminante de Fisher (KDF)

Na literatura internacional, vários autores concentram esforços na utilização de algoritmos de aprendizagem baseados no KDF. Os resultados apontados mostram que essa tecnologia é promissora.

ZANGH e MA (2004) apresentam a utilização do *kernel* discriminante de Fisher (KDF) na classificação de padrões de textura em diversas imagens teste com diferentes complexidades. Os resultados experimentais mostram claramente que o método proposto apresenta excelente desempenho na classificação e resultados superiores a outros métodos clássicos de reconhecimento de textura. Os autores também demonstram em seu estudo que é possível extrair boas características descritivas da textura sem a necessidade de se aplicar o processo de filtragem na imagem como um processo de pré-processamento.

KODIPAKA *et al*. (2007) apresentam a aplicação do *kernel* discriminante de Fisher na análise estatística das deformações que indicam a posição hemisférica de um foco epilético. Para tal utilizam duas classes de varredura em pacientes portadores de epilepsia, uma direta e a outra com foco temporal medial anterior esquerdo do lóbulo (RATL e LATL), confirmado através do consenso clínico e pela cirurgia subseqüente. Os resultados obtidos mostraram um aumento significativo na melhora da classificação dos pacientes pertencentes a uma das classes LATL ou RATL, respectivamente.

No estudo realizado por DeCOSTE (2001), encontramos a aplicação e avaliação do desempenho das máquinas de vetores de suporte (SVM) e o *kernel* discriminante de Fisher (KDF) em uma série de dados gerado de um polímero. A habilidade dos classificadores de SVM e do KDF em identificar a classe funcional correspondente (categoria) de um produto químico baseado em sua assinatura eletrônica é comparada e avaliada com outros métodos tradicionais, incluindo os vizinhos mais próximos e o discriminante linear de Fisher (LDF). Os resultados mostram que o método baseado no *kernel* discriminante de Fisher obteve um desempenho de 85% para a classificação do grupo de teste, levando os autores a afirmarem que o mesmo possui potencial para as aplicações no mundo real.

YANG *et al*. (2005) propuseram um novo algoritmo baseado no método de análise *kernel* discriminante de Fisher, denominado KDF completo (CKFD). Os autores ressaltam que o modelo proposto possui duas vantagens sobre os algoritmos tradicionais

existentes de KDF. Primeiramente, sua execução é dividida em duas etapas, sendo a primeira composta de uma análise da componente principal do *kernel* (KPCA) seguida da análise discriminante linear de Fisher (LFD), fato este que torna o modelo proposto mais transparente e simples. Na segunda etapa, o CKFD pode utilizar as duas categorias de informação obtida pelo discriminante, tornando o modelo mais robusto. O algoritmo proposto foi aplicado em um conjunto de dados composto de 200 imagens. Os resultados experimentais mostraram que o modelo proposto CKFD foi significativamente melhor quando comparado com os resultados obtidos pelos algoritmos KFD existentes.

BO *et al.* (2006) propuseram uma variação no método proposto por MIKA *et al*. (2001), para a definição dos parâmetros de *kernel* e regularização (FS-KDF), baseado na otimização do erro de validação cruzada *leave-one-out* via método do gradiente descendente. Os resultados obtidos na classificação pelo método FS-KDF foram superiores quando, comparados com os demais métodos, entretanto, não apresentaram diferença estatística ao nível de significância de 5%.

Apesar do crescente interesse dos pesquisadores, na utilização de classificadores baseado na tecnologia de *kernel*, sua utilização no reconhecimento de padrões gênicos associados à resistência aos fármacos utilizados na terapia antiretroviral é insipiente.

SING e BEERENWINKEL (2006) utilizaram um modelo de mistura baseado no *kernel* de Fisher e árvore (MTreeMix Fisher *kernel*) para a previsão de resistência do HIV aos inibidores NNRTIs, NRTIs e IPs. Para tal, os autores utilizaram um conjunto de dados variando de 305 a 858 amostras, onde as mutações de resistência selecionadas foram definidas pela Sociedade Internacional de AIDS (JOHNSON *et al*., 2008). Os resultados obtidos pelo método proposto apresentaram coeficientes de determinação (coeficiente de correlação ao quadrado r^2) superiores para todas as classes de inibidores (RTs e PIs), quando comparado com o método da regressão linear, indicando que as mutações selecionadas podem possibilitar previsões mais confiáveis no que tange a resistência a terapia antiretroviral, bem como os métodos de *kernel* podem desempenhar papel importante no contexto da classificação de resistência.

SAIGO *et al*., (2007) propuseram um método de modo a predizer a resistência aos anti retrovirais a partir de dados de genótipo. Consistindo basicamente em um método de regressão não-linear no espaço de características no conjunto de mutações. Para verificar sua potencialidade, são comparados os resultados obtidos pelo método de SVM com *kernel* gaussiano e polinomial, utilizando amostras de genotipagem obtidas

junto bancos de dados de Stanford. SAIGO *et al* afirmam que apesar do custo computacional do método ser elevado, esse possibilita a descoberta de associação marcante de mutação. Os resultados obtidos na previsão para os inibidores NRTIs, apresentaram valores de precisão superiores aos SVMs, fato que não foi verificado para os inibidores NNRTIs e IPs.

Neste capítulo inicialmente foi apresentado o classificador linear de Fisher, seguido do classificador de KDF. Foram detalhadas as principais funções de *kernel* utilizadas.Foi descrito o modelo proposto por MIKA *et al*, em 2001 que reformula o KDF como um problema de otimização quadrática. Finalmente foram descritos trabalhos encontrado na literatura científica com a aplicação do KDF em problemas de classificação. Todos os conceitos, vistos nos capítulos anteriores, serão usados para formular e aplicar o KDF ao problema de classificação de resistência em pacientes em falha terapêutica aos inibidores de protease do HIV-1 no Brasil, o qual será apresentado no próximo capítulo.

Capítulo 5

Materiais e Métodos

5.1 Conjunto de Dados

O conjunto de dados utilizado nesse estudo consistiu de 1092 seqüências do gene da protease provenientes de isolados séricos de pacientes portadores do HIV-1, resistentes à terapia antiretroviral, obtidos junto ao Laboratório de Virologia Molecular da Universidade Federal do Rio de Janeiro (UFRJ/Brasil), integrante da rede de laboratórios de genotipagem do Ministério da Saúde (RENAGENO). O conjunto de dados também possui informações clínicas referentes à contagem de linfócitos T-CD4^{+}, carga viral e descrição dos regimes terapêuticos utilizados pelo paciente. Um resumo das características do conjunto de dados pode ser vista nas tabelas 5.1 e 5.2.

Tabela 5.1: Classificação da população de acordo com os subtipos baseado na seqüência da protease.

Subtipo[1]	Total por subtipo	Percentual	Percentual sexo Feminino
B	754	69,04	28,10
C	190	17,40	37,90
F	123	11,26	40,50
Outros[2]	25	2,30	35,00
Total	1092	100,00	31,40

[1] - O subtipo foi baseado na análise filogenética (*Kimura* 2-parâmetros), avaliado pela seqüência obtida de Los Alamos.

[2] – Outros incluem os não subtipos B/C/F , bem como os mosaicos entre esses subtipos.

Tabela 5.2: Dados clínicos da infecção do HIV-1 para os diferentes subtipos[1].

Subtipos e Variáveis	Valor
Total	
Idade (anos)[média ± desvio padrão]	37,10 ± 13,20
CD4 (células/mm^3)[média ± desvio padrão]	283,90 ± 198,00
Carga Viral (log_{10} cópias/ml)[média ± desvio padrão]	4,61 ± 0,60
Subtipo B	
Idade (anos)[média ± desvio padrão]	38,20 ± 12,10
CD4 (células/mm^3)[média ± desvio padrão]	280,60 ± 196,80
Carga Viral (log_{10} cópias/ml)[média ± desvio padrão]	4,60 ± 0,61
Subtipo C	
Idade (anos)[média ± desvio padrão]	30,49 ± 1,38
CD4 (células/mm^3)[média ± desvio padrão]	320,30 ± 225,87
Carga Viral (log_{10} cópias/ml)[média ± desvio padrão]	4,59 ± 0,58
Subtipo F	
Idade (anos)[média ± desvio padrão]	37,00 ± 15,10
CD4 (células/mm^3)[média ± desvio padrão]	260,31 ± 173,78
Carga Viral (log_{10} cópias/ml)[média ± desvio padrão]	4,58 ± 0,59
Subtipo Outros	
Idade (anos)[média ± desvio padrão]	36,20 ± 13,70
CD4 (células/mm^3)[média ± desvio padrão]	262,31 ± 132,50
Carga Viral (log_{10} cópias/ml)[média ± desvio padrão]	4,40 ± 0,40

O percentual de pacientes em falha terapêutica que foram experimentados no último regime terapêutico aos inibidores de protease no conjunto de dados é exemplificado na tabela 5.3.

Tabela 5.3: Distribuição do número de pacientes tratados no último regime terapêutico com inibidores de protease.

Inibidor Protease	Número e Percentual (%) para cada grupo de subtipo do HIV-1				
	Total (*n* =554)	Subtipo B (*n* = 400)	Subtipo C (*n* =80)	Subtipo F (*n* =64)	Outros[2] (*n* =10)
Saquinavir	44(7,94)	34(8,50)	3(3,75)	5(7,81)	2(20,00)
Indinavir	79(14,26)	60(15,0)	7(8,75)	9(14,06)	3(30,00)
Nelfinavir	196(35,38)	132(33,00)	34(42,50)	23(35,94)	7(70,00)
Lopinavir	125(22,56)	92(23,00)	17(21,25)	16(25,00)	0(0,0)
Amprenavir	39(7,04)	27(6,8)	8(10,00)	4(6,25)	0(0,0)
Atazanavir	50(9,03)	38(9,5)	9(11,25)	3(4,69)	0(0,0)

A distribuição de freqüência de mutação de resistência no gene da protease, em relação ao último esquema terapêutico é mostrada no ANEXO I.

5.2 Metodologia

O método proposto nesse estudo, consiste em um modelo computacional híbrido (AG/KDF) na seleção de posições de mutações do gene da protease, de pacientes em falha terapêutica, no Brasil via algoritmo genético (AG) combinado ao classificador não-linear de Fisher (KDF). O fluxograma da metodologia pode ser vista na Figura 5.1, e suas etapas são descritas a seguir.

Categorização das amostras em função do subtipo viral (B, C, F e Outros), tendo como base a análise filogenética de Kimura de 2-parâmetros; eliminação de atributos (resíduos de aminoácidos da seqüência da protease do HIV-1) estáveis, ou seja, que não "mudaram", tendo como base para o subtipo B a seqüência de referência do vírus selvagem (HXB-2) e para o subtipo C a seqüência de referência definida por GAO (2002).

Os resíduos de aminoácidos (tabela 5.4) foram codificados utilizando o fator de hidrofobicidade (hidrofóbicos ou hidrofílicos) dos aminoácidos, visto que a codificação possibilita à observação apropriada das variações estequiométricas (espaciais) na molécula funcional da protease do HIV-1. Para tal, foi utilizada a escala de hidrofobicidade de KYTE e DOOLITTLE (1982), ponderada pelos respectivos pesos moleculares, através da equação (5.1), e os valores obtidos estão apresentados na tabela (5.5).

$$\text{hidro\_matrix}[i][j] = \frac{abs(hidroAA[i] * PM - hidroAA[j] * PM)}{100} \qquad (5.1)$$

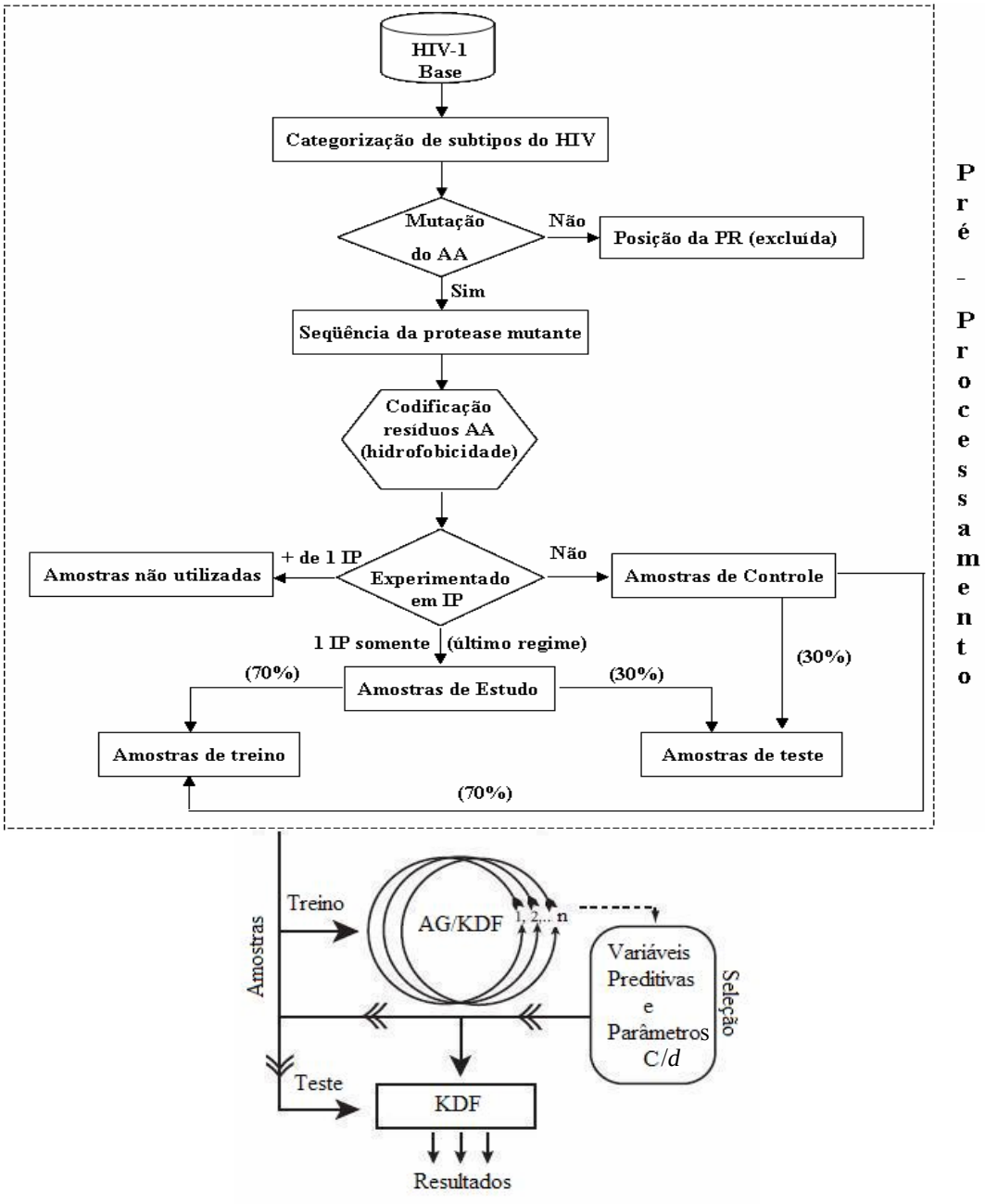

Figura 5.1: Fluxograma da metodologia proposta no modelo computacional AG/KDF.

Tabela 5.4: Valores atribuídos a cada aminoácido e sua respectiva classificação quanto ao valor de sua hidrofobicidade, pela escala de Kyte e Doolittle.

Aminoácido	Símbolo	Valor hidrofobicidade	Categoria
Isoleucina	**I**	+ 4,5	Hidrofóbico
Valina	**V**	+ 4,2	Hidrofóbico
Leucina	**L**	+ 3,8	Hidrofóbico
Fenilalanina	**F**	+ 2,8	Hidrofóbico
Cisteina	**C**	+ 2,5	Hidrofóbico
Metionina	**M**	+ 1,9	Hidrofóbico
Alanina	**A**	+ 1,8	Hidrofóbico
Glicina	**G**	- 0,4	Neutro
Treonina	**T**	- 0,7	Neutro
Serina	**S**	- 0,8	Neutro
Triptofano	**W**	- 0,9	Neutro
Tirosina	**Y**	- 1,3	Neutro
Prolina	**P**	- 1,6	Neutro
Histidina	**H**	- 3,2	Hidrofílico
Glutamina	**Q**	- 3,5	Hidrofílico
Asparagina	**N**	- 3,5	Hidrofílico
Ácido glutâmico	**E**	- 3,5	Hidrofílico
Ácido aspártico	**D**	- 3,5	Hidrofílico
Lisina	**K**	- 3,9	Hidrofílico
Arginina	**R**	- 4,5	Hidrofílico

Tabela 5.5 – Matriz de valores de mutação dos aminoácidos em função hidrofobicidade após normalização da escala de Kyte e Doolittle.

		I	**V**	**L**	**F**	**C**	**M**	**A**	**G**	**T**	**S**	**W**	**Y**	**P**	**H**	**E**	**Q**	**D**	**N**	**K**	**R**
		5,90	4,91	4,98	4,62	3,03	2,83	1,60	-0,30	-0,83	-0,84	-1,84	-2,35	-1,84	-4,96	-5,15	-5,11	-4,66	-4,62	-5,69	-7,83
I	5,90	0,00	0,00	0,06	4,56	-1,53	4,36	-2,76	2,46	-3,29	2,45	-4,29	1,94	-3,78	-1,18	-3,96	-1,15	-3,51	-1,11	-4,58	-3,25
V	4,91	0,98	0,00	0,06	-0,29	-1,89	-2,08	-3,31	-5,21	-5,75	-5,75	-6,75	-7,27	-6,75	-9,87	-10,06	-10,02	-9,57	-9,53	-10,61	-12,74
L	4,98	0,92	-0,06	0,00	-0,36	-1,95	-2,15	-3,38	-5,28	-5,81	-5,82	-6,81	-7,33	-6,82	-9,94	-10,12	-10,09	-9,63	-9,60	-10,67	-12,81
F	4,62	1,28	0,29	0,36	0,00	-1,60	-1,79	-3,02	-4,92	-5,45	-5,46	-6,46	-6,97	-6,46	-9,58	-9,77	-9,73	-9,28	-9,24	-10,31	-12,45
C	3,03	2,87	1,89	1,95	1,60	0,00	-0,19	-1,42	-3,33	-3,86	-3,87	-4,86	-5,38	-4,87	-7,99	-8,17	-8,14	-7,68	-7,65	-8,72	-10,86
M	2,83	3,06	2,08	2,15	1,79	0,19	0,00	-1,23	-3,13	-3,66	-3,67	-4,67	-5,18	-4,67	-7,79	-7,98	-7,94	-7,49	-7,45	-8,53	-10,66
A	1,60	4,29	3,31	3,38	3,02	1,42	1,23	0,00	-1,90	-2,44	-2,44	-3,44	-3,96	-3,44	-6,56	-6,75	-6,71	-6,26	-6,22	-7,30	-9,43
G	-0,30	6,20	5,21	5,28	4,92	3,33	3,13	1,90	0,00	-0,53	-0,54	-1,54	-2,05	-1,54	-4,66	-4,85	-4,81	-4,36	-4,32	-5,39	-7,53
T	-0,83	6,73	5,75	5,81	5,45	3,86	3,66	2,44	0,53	0,00	-0,01	-1,00	-1,52	-1,01	-4,13	-4,31	-4,28	-3,82	-3,79	-4,86	-7,00
S	-0,84	6,74	5,75	5,82	5,46	3,87	3,67	2,44	0,54	0,01	0,00	-1,00	-1,51	-1,00	-4,12	-4,31	-4,27	-3,82	-3,78	-4,85	-6,99
W	-1,84	7,73	6,75	6,81	6,46	4,86	4,67	3,44	1,54	1,00	1,00	0,00	-0,52	0,00	-3,12	-3,31	-3,27	-2,82	-2,78	-3,86	-5,99
Y	-2,35	8,25	7,27	7,33	6,97	5,38	5,18	3,96	2,05	1,52	1,51	0,52	0,00	0,51	-2,61	-2,79	-2,76	-2,30	-2,27	-3,34	-5,48
P	-1,84	7,74	6,75	6,82	6,46	4,87	4,67	3,44	1,54	1,01	1,00	0,00	-0,51	0,00	-3,12	-3,31	-3,27	-2,82	-2,78	-3,85	-5,99
H	-4,96	10,86	9,87	9,94	9,58	7,99	7,79	6,56	4,66	4,13	4,12	3,12	2,61	3,12	0,00	-0,19	-0,15	0,31	0,34	-0,73	-2,87
E	-5,15	11,04	10,06	10,12	9,77	8,17	7,98	6,75	4,85	4,31	4,31	3,31	2,79	3,31	0,19	0,00	0,03	0,49	0,52	-0,55	-2,69
Q	-5,11	11,01	10,02	10,09	9,73	8,14	7,94	6,71	4,81	4,28	4,27	3,27	2,76	3,27	0,15	-0,03	0,00	0,46	0,49	-0,58	-2,72
D	-4,66	10,55	9,57	9,63	9,28	7,68	7,49	6,26	4,36	3,82	3,82	2,82	2,30	2,82	-0,31	-0,49	-0,46	0,00	0,04	-1,04	-3,18
N	-4,62	10,52	9,53	9,60	9,24	7,65	7,45	6,22	4,32	3,79	3,78	2,78	2,27	2,78	-0,34	-0,52	-0,49	-0,04	0,00	-1,07	-3,21
K	-5,69	11,59	10,61	10,67	10,31	8,72	8,53	7,30	5,39	4,86	4,85	3,86	3,34	3,85	0,73	0,55	0,58	1,04	1,07	0,00	-2,14
R	-7,83	13,73	12,74	12,81	12,45	10,86	10,66	9,43	7,53	7,00	6,99	5,99	5,48	5,99	2,87	2,69	2,72	3,18	3,21	2,14	0,00

Após a codificação, foi realizada a separação da amostra de pacientes experimentados somente a um inibidor de PR no último regime e não experimentados para essa classe de inibidor.

Os pacientes experimentados no último regime terapêutico a mais de um inibidor de PR foram excluídos do banco de dados. Sendo dividida a amostra em dois grupos independentes, o conjunto de treinamento e o conjunto de teste.

O conjunto de treinamento composto por 70% dos pacientes experimentados ou não para o inibidor em estudo e o conjunto teste composto com os 30% restante. A esse conjunto de procedimentos denominamos pré-processamento do modelo AG/KDF.

O AG/KDF é um algoritmo evolutivo, sendo empregado basicamente para a seleção de atributos (resíduos de aminoácidos da PR viral) mutáveis (informativos), sendo utilizada como representação genotípica dos indivíduos da população cromossomos de dois alelos binários (figura 5.2). Sendo o primeiro de comprimento igual à dimensão de entrada (número de resíduos de aminoácidos mutáveis na presença / ausência de falha terapêutica), onde o valor 1 indicará que o resíduo será incorporado no treinamento do classificador KDF e 0 indicará o contrário. E o segundo com quatro alelos binários, sendo dois reservados para a seleção dos parâmetros de regularização C, onde os genótipos [0,0], [0,1], [1,0] e [1,1] representam respectivamente, os fenótipos {0,1, 1,0, 10, 100} e os dois restantes utilizados na codificação da variância da função do núcleo Gaussiana *d,* do classificador KDF, onde os genótipos [0,0], [0,1], [1,0] e [1,1], representam os fenótipos {0,5, 0,25, 0,125 e 0,0625}.

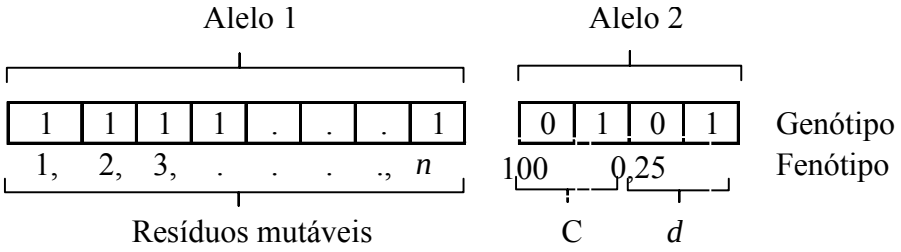

Figura 5.2: Representação genotípica utilizada para o AG/KDF. Cada solução (cromossomo) contém dois alelos, o primeiro responsável pela representação do número de resíduos de aminoácidos mutantes da seqüência da PR e o segundo responsável pela codificação do parâmetro de regularização C e de distribuição gaussiana *d* utilizados pelo classificador KDF.

O AG/KDF visa minimizar a soma ponderada do erro de treinamento do classificador KDF e a razão dos resíduos de aminoácidos, sendo descrito como:

$$F(x,C,d) = w_1 \cdot Err + w_2 \cdot \frac{N_{res}}{N_{Tres}} \qquad (5.2)$$

onde w_1 é o peso do erro de treinamento do classificador KDF *(Err)* e w_2 o peso da razão entre o número de resíduos de aminoácidos selecionados pelo AG (N_{res}) e o número total de resíduos mutável (NT_{res}).

Para cada inibidor de protease são realizadas 20 simulações, o AG/KDF evolui enquanto não é satisfeito um dos seguintes critérios de parada: (1) Valor mínimo da função objetivo; ou, (2) Número máximo de gerações. Não sendo satisfeito qualquer critério de parada definido, os indivíduos são modificados pelo operador de recombinação (*crossover*) de ponto único e mutação, com razão de recombinação e re-inserção selecionadas no intervalo [0,6, 1,0]. Gerada a população intermediária pela modificação dos indivíduos da população atual, o operador de re-inserção elitista selecionará 10% dos indivíduos mais aptos em função de suas aptidões, selecionadas da população atual, assim como os 90% mais aptos da população intermediária.

Ao final do AG/KDF, ou seja, quando um dos critérios de parada é/são obtido(s), o melhor indivíduo, aquele de menor valor para a função objeto, será escolhido como a melhor solução para o problema e será empregado no cálculo do erro de teste na amostra independente, bem como, o erro de generalização obtido pela

validação cruzada *leave-one-out* (LOOCV). As posições originais dos resíduos de aminoácidos da protease viral serão obtidas a partir da codificação genotípica do melhor indivíduo, juntamente com o parâmetro de regularização C e de distribuição gaussiana *d* selecionados nesse subconjunto.

Esses parâmetros são então empregados no treinamento do classificador KDF e a performance dos subconjuntos de variáveis selecionadas na previsão de resistência são avaliados pela acurácia de classificação, ou precisão total (*ACC*), sensibilidade (*S*) e especificidade (*E*), definidos respectivamente por:

$$Acc = \left(V_P + V_N\right) / n \tag{5.3}$$

$$S = V_P / \left(V_P + F_N\right) \tag{5.4}$$

$$E = V_N / \left(V_N + F_P\right) \tag{5.5}$$

Ao final das 20 simulações, são selecionadas as que apresentarem os melhores resultados em termos da validação cruzada (LOOCV) no conjunto de treino. Para tal, os valores de acurácia da LOOCV são ordenados em ordem crescente e as simulações que apresentarem resultado superior ao limite de distribuição do primeiro quartil serão selecionadas.

Após essa seleção, o mesmo processo se realizará nesse conjunto de simulações, onde serão utilizadas somente no classificador KDF, na previsão de resistência das amostras independente de teste, as posições de mutação obtidas pelo modelo computacional que apresentarem freqüência superior a distribuição do primeiro quartil para o subtipo B e superior ou igual para o subtipo C. Evita-se, assim, que posições selecionadas raramente (não polimórficas), sejam incorporadas no modelo computacional (SHAFER *et al*., 2007).

Os parâmetros utilizados no modelo AG/KDF na seleção das variáveis (mutações) foram obtidos de forma experimentalmente. A relação dos parâmetros está sumarizada na Tabela 5.6.

Tabela 5.6: Parâmetros utilizados pelo modelo AG/KDF na seleção/previsão de mutações de resistência na PR em pacientes em falha terapêutica antiretroviral.

Parâmetros	Valor
Simulações	20
População (n)	100
Gerações	100
Valor mínimo função objetivo	0,01
Taxa de recombinação	0,7
Taxa de mutação	0,009
Taxa re-inserção	0,9
Formato da variável	Binário (0,1)
Função de *kernel*	Guassiana
Parâmetro de regularização (C)	[0,1; 1,0; 10; 100]
Variância da função de *kernel* (d)	[0,5; 0,25; 0,125; 0,0625]

O modelo computacional AG/KDF proposto nesse estudo utilizou-se do pacote *GEATbx* (POHLHEIM, 2007), disponível em *http://www.geatbx.com*, na implementação do algoritmo genético e o *software MATLAB*® (The MathWorks, Inc; *http://www.mathworks.com*) na otimização do classificador *Kernel* Discriminante de Fisher (KDF).

Neste capítulo, foi apresentado um modelo baseado em Algoritmos Genéticos e no classificador de *Kernel* Discriminante de Fisher, na identificação de possíveis novas mutações de resistência no gene da aspartil-protease do HIV-1 de subtipos B e C para os inibidores SQV, NFV e LPV, bem como a previsão de resistência. A seguir, o modelo proposto (AG/KDF) foi detalhado em termos dos parâmetros utilizados. No próximo capítulo serão apresentados os resultados obtidos com a aplicação do modelo proposto.

Capítulo 6

Resultados

A utilização do modelo computacional AG/KDF proposto nesse trabalho foi avaliado para três inibidores de protease Saquinavir (SQV), Nelfinavir (NFV) e Lopinavir (LPV) utilizados no tratamento de pacientes em falha terapêutica para os subtipos B e C circulante no Brasil.

6.1 Saquinavir

O conjunto de dados de pacientes portadores do HIV-1 subtipo B e experimentado ao SQV no último regime terapêutico, contém 34 amostras. Sendo 24 amostras utilizadas no conjunto de treinamento e 10 no conjunto de teste. O conjunto controle (*naive* na protease ou *naive* puro) para o subtipo B é composto de 236 amostras. Para o subtipo C o número de pacientes experimentados no último regime terapêutico ao SQV contém somente 3 amostras, fato que inviabiliza a aplicação do modelo proposto para esse subtipo.

O desempenho obtido no conjunto das 20 simulações realizadas pelo modelo AG/KDF na categorização de resistência, ordenado em função da validação cruzada *leave-one-out* (LOOCV) no conjunto de treino, bem como os residuos de aminoácidos (posições de mutações) selecionados e o valor do parâmetro C, são apresentados na tabela 6.1 .

Tabela 6.1: Resumo dos resultados obtidos pelo modelo AG/KDF para a acurácia de treino, validação cruzada *leave-one-out* (LOOCV), acurácia de teste, parâmetro de regularização C e posições de mutações selecionadas para o inibidor SQV.

Simulação	Acurácia de Treino (%)	LOOCV (%)	Acurácia de Teste (%)	C	Posições de mutações selecionadas
6	87,83	62,96	86,42	1	12 15 20 35 36 37 43 46 60 69 70 75 77 84 90 93
11	87,83	63,49	75,31	1	35 36 37 54 62 63 71 73 77 84 93
19	87,83	64,02	71,60	10	15 35 37 54 57 62 63 64 69 70 72 90 93
4	88,36	65,61	69,14	1	10 13 15 20 35 41 57 62 63 69 70 77 89 93
7	87,83	67,20	86,42	0,1	15 17 35 57 62 63 70 71 77 82 90 93
10	87,83	67,72	80,25	100	20 35 37 41 46 57 63 69 71 73 93
13	89,42	68,78	71,60	1	20 33 35 57 60 62 63 69 71 73 90
1	91,01	72,49	69,14	0,1	14 15 19 35 36 54 55 60 63 69 72 77 82 93
20	89,42	73,02	67,90	10	13 41 45 46 54 61 63 71 73 90 93
14	92,06	74,60	56,79	10	10 13 20 35 36 41 62 64 69 71 77 93
8	92,06	79,89	87,65	0,1	13 17 19 20 33 35 48 54 57 61 62 63 69 72 73 74 77 93
9	89,95	79,89	79,01	0,1	10 24 36 57 61 63 69 71 72 77 84 93
16	89,95	79,89	87,45	1	10 15 33 36 37 63 70 72 82 84 93
3	91,53	82,01	83,95	0,1	10 12 20 57 69 70 72 77 90
2	89,95	83,07	71,60	0,1	10 18 36 37 41 55 71 72 90
5	89,95	84,13	87,65	0,1	15 20 35 36 43 62 71
18	87,30	85,19	85,19	0,1	10 24 48
12	88,36	85,71	80,25	0,1	14 54 69 73 90
15	88,89	86,77	69,14	1	20 66 71 82 89 90
17	92,00	88,00	89,00	0,1	14 20 36 48 54 61 62 63 69 72 73 77 90 93

A região hachurada representa as simulações que apresentaram valores superiores ao limite de distribuição do primeiro quartil no conjunto do erro LOOCV.

A simulação que apresentou melhor desempenho no conjunto da validação cruzada *leave-one-out*, obteve valor de 88,00 % na caracterização de resistência para o SQV, selecionando as posições de mutações K14, K20, M36, G48, I54, Q61, I62, L63, H69, I72, G73, V77, L90 e I93. Das posições de mutações selecionadas pelo modelo AG/KDF para o SQV, somente a posição I84, descrita na literatura como uma posição associada a resistência primária para o inibidor em estudo, não foi caracterizada nessa simulação.

Visando eliminar os piores resultados obtidos no conjunto da validação cruzada *leave-one-out* (LOOCV), utilizou-se o valor da distribuição do primeiro quartil como ponto de corte, sendo o conjunto das simulações que apresentaram valores superiores ao limite estabelecido, assinalados na tabela 6.1.

A distribuição de freqüência das posições de mutações da seqüência da protease selecionadas pelo modelo computacional no conjunto das melhores simulações, está representada na tabela (6.2). As posições assinaladas representam as posições de mutações selecionadas com freqüência de mutação superior ao primeiro quartil (> 1,31), considerando as freqüências em ordem crescente.

Tabela 6.2: Distribuição das freqüências de posições de mutações selecionadas pelo modelo AG/KDF para o SQV.

Posição	**Freq.(%)**	**Posição**	**Freq.(%)**	**Posição**	**Freq.(%)**
L10	3,92	S37	1,96	H69	5,88
T12	0,65	R41	2,61	K70	1,31
I13	1,96	K43	1,31	A71	5,23
K14	1,96	M46	1,31	I72	4,58
I15	1,96	G48	1,96	G73	3,92
G17	1,31	I54	3,27	T74	0,65
Q18	0,65	K55	1,31	V77	3,92
L19	1,31	R57	3,27	V82	1,96
K20	5,23	D60	1,31	I84	1,31
L24	1,31	Q61	2,61	L89	0,65
L33	1,96	I62	3,27	L90	4,58
E35	3,92	L63	5,23	I93	5,23
M36	4,58	I64	0,65		

Visando comparar o desempenho do classificador KDF com a utilização das 24 posições de mutações mais freqüentes selecionadas na categorização de resistência para o SQV, foram utilizadas as posições clássicas de resistência para o SQV: L10, L24, G48, I54, I62, A71, G73, V77, V82, I84 e L90 descrita por JONHSON *et al*.,(2008). Os resultados obtidos na categorização de resistência pelo classificador KDF no conjunto de teste nas 20 simulações, estão representados na tabela (6.3).

Tabela 6.3: Resultados obtidos na categorização de resistência para o SQV, utilizando posições de mutações mais freqüentes selecionadas pelo modelo computacional AG/KDF e posições clássicas descritas na literatura.

	Modelo com 24 Posições			**Modelo com Posições Clássicas**		
Simulações	**Acurácia (%)**	**S (%)**	**E (%)**	**Acurácia (%)**	**S (%)**	**E (%)**
1	93,83	100,00	92,96	70,37	100,00	66,20
2	97,53	100,00	97,18	77,78	100,00	74,65
3	97,53	100,00	97,18	77,78	100,00	74,65
4	87,65	100,00	85,92	81,48	100,00	78,87
5	87,65	100,00	85,92	77,78	100,00	74,65
6	87,65	100,00	85,92	81,48	100,00	78,87
7	87,65	100,00	85,92	80,25	100,00	77,46
8	95,06	94,89	94,37	88,89	90,00	88,73
9	87,65	100,00	85,92	65,43	100,00	60,56
10	96,30	100,00	95,77	76,54	100,00	73,24
11	87,65	100,00	85,92	61,73	100,00	56,34
12	87,42	100,00	85,92	58,02	100,00	52,11
13	87,65	100,00	85,92	60,49	100,00	54,93
14	97,53	100,00	97,18	96,30	100,00	95,77
15	87,65	100,00	85,92	64,20	100,00	59,15
16	96,30	97,36	95,77	87,65	80,00	88,73
17	97,53	100,00	97,18	86,58	100,00	85,62
18	88,89	100,00	87,32	80,25	100,00	77,46
19	87,65	100,00	85,92	79,01	90,00	77,46
20	87,65	100,00	85,92	51,85	100,00	45,07

6.2 Nelfinavir

O conjunto de pacientes portadores do HIV-1, subtipo B, e experimentado no último regime terapêutico ao inibidor de protease NFV, possui 132 amostras. Sendo 93 amostras utilizadas no conjunto de treinamento e 39 no conjunto de teste. Para o subtipo C, o número de pacientes experimentados no último regime terapêutico ao NFV foi de 34 amostras, sendo 24 utilizadas no conjunto de treino e 10 no conjunto de teste. O grupo controle (*naive* na protease ou *naive* puro) para o subtipo B é composto de 236 amostras, sendo 165 utilizadas no conjunto de treino e 71 no conjunto de teste. Para o subtipo C esse grupo é composto de 88 amostras, sendo 62 empregadas no conjunto de treino e 26 no conjunto de teste.

O resumo do desempenho obtido na categorização de resistência ao NFV para cada uma das 20 simulações, ordenada em função da validação cruzada *leave-one-out*, bem como os resíduos de aminoácidos (posição de mutação) selecionados e o valor do parâmetro C, para os subtipos B e C, são apresentados respectivamente nas tabelas (6.4) e (6.5).

Para o subtipo B a simulação que apresentou melhor desempenho na validação cruzada *leave-one-out*, obteve valor de 87,00 % na caracterização de resistência para o NFV, selecionando as posições de mutações: L10, I15, D30, M36, S37, R41, M46, I62, L63, I64, A71, I72, V77, I84 e N88. Das posições de mutações que conferem resistência a esse inibidor descritas na literatura, somente a posição L90 não foi caracterizada nessa simulação.

Já para o subtipo C, a simulação que obteve o melhor resultado na categorização de resistência para o NFV selecionou as posições de mutação: L10, D30 e N88; tendo conseguido 81,25 % no desempenho na validação cruzada *leave-one-out*.

Tabela 6.4: Resumo dos resultados obtidos pelo modelo AG/KDF para a acurácia de treino, validação cruzada *leave-one-out* (LOOCV), acurácia de teste, parâmetro de regularização C e posições de mutações selecionadas para o NFV, tendo como base o subtipo B.

Simulação	Acurácia de Treino (%)	LOOCV (%)	Acurácia de Teste (%)	C	Posições de mutações selecionadas
6	84,00	60,00	79,00	1	10 12 20 35 37 41 43 50 57 63 88
12	85,00	65,40	77,00	100	10 19 35 36 37 46 51 54 57 63 64 65 69 71 72 77 93
3	86,00	66,50	80,00	100	10 14 15 20 30 35 37 41 62 63 72 77 90
8	85,00	68,40	79,00	0,1	10 12 13 20 35 36 37 45 57 63 64 69 72 74 77 88
10	88,00	69,10	77,00	0,1	10 14 15 19 36 37 41 49 63 64 69 71 72 73 77
9	86,00	69,22	78,00	100	10 13 35 36 37 41 57 63 69 72 82 88 89 93
14	84,00	69,22	71,00	1	12 13 15 19 20 36 37 57 64 82 88
13	87,00	70,00	76,00	0,1	10 12 15 36 37 46 53 54 55 63 64 69 71 72 77 90
20	87,00	70,00	56,00	1	10 14 15 30 36 37 39 43 46 47 62 63 64 77 84 89 93
7	86,00	71,00	79,00	100	10 15 20 35 36 37 41 45 48 63 64 69 70 72 77 88 93
16	86,00	72,00	68,00	0,1	15 28 30 33 36 37 41 45 46 47 54 62 63 64 71 77 84 93
18	85,00	72,00	64,00	0,1	13 14 20 36 41 54 62 64 69 70 77 93
19	85,00	74,00	58,00	0,1	12 13 15 19 35 36 41 54 57 71 82 93
17	81,00	75,00	68,00	0,1	13 15 20 35 54 62 63 74
15	85,00	78,00	68,00	0,1	10 15 35 36 41 69 72
2	87,00	81,00	80,00	100	14 19 20 30 35 36 37 48 88 89 90
5	87,00	83,00	79,00	100	10 15 30 36 37 41 46 62 63 64 72 73 77 84 90 93
11	86,00	83,00	77,00	0,1	10 15 19 20 23 28 33 39 41 63 64 69 71 72 77 93
4	88,00	86,90	79,00	1	10 19 35 36 37 46 51 54 63 64 65 69 77 88 93
1	90,00	87,00	81,00	10	10 15 30 36 37 41 46 62 63 64 71 72 77 84 88

A região hachurada representa as simulações que apresentaram valores superiores ao limite do primeiro quartil no conjunto do erro LOOCV.

Tabela 6.5: Resumo dos resultados obtidos pelo modelo AG/KDF para a acurácia de treino, validação cruzada *leave-one-out* (LOOCV), acurácia de teste, parâmetro de regularização C e posições de mutações selecionadas para o NFV, tendo como base o subtipo C.

Simulação	Acurácia de Treino (%)	LOOCV (%)	Acurácia de Teste (%)	C	Posições de mutações selecionadas
12	85,00	57,50	52,94	10	10 13 19 20 30 63 71
7	85,00	60,00	50,00	1	14 20 46 60 61 64 88 89
20	76,74	62,79	57,22	1	14 16 20 36 82
17	85,00	63,75	47,06	0,1	10 12 13 14 16 63 90
4	82,50	67,75	64,71	0,1	13 14 20 23 30
1	82,50	68,75	61,76	1	20 36 43 57 62 71 82
16	81,25	70,00	61,76	10	12 14 30 43 63
11	78,75	71,25	73,53	100	11 30 45 75 88 89 90
2	80,00	72,50	70,59	1	10 72 88
15	83,75	72,50	52,94	100	14 20 35 61 74
8	82,50	73,75	61,76	0,1	30 36 46 75 82
13	80,00	73,75	70,59	100	20 35 88
14	81,25	73,75	67,65	100	16 71 72 88
6	86,25	75,00	58,82	1	10 45 62 63 70 71
3	80,00	76,25	73,53	100	12 14 45 46 90
9	80,00	76,25	70,59	0,1	30 36 63 74
10	80,00	76,25	73,53	0,1	16 30 70
18	82,50	76,25	61,76	1	30 72 89
5	80,23	78,44	77,46	10	10 30 36 41 54 82 89 90
19	83,75	81,25	61,76	1	10 30 74 88

A região hachurada representa as simulações que apresentaram valores superiores ao limite do primeiro quartil no conjunto do erro LOOCV.

As simulações que apresentaram resultados no conjunto de validação cruzada *leave-one-out* superiores ao valor limite estabelecido pelo ponto de corte (primeiro quartil), para os subtipos B e C estão assinaladas nas tabelas (6.6) e (6.7), respectivamente.

A tabela (6.6) mostra as distribuições de freqüência das posições de mutações da seqüência da protease, selecionadas pelo modelo computacional no conjunto das melhores simulações (LOOCV), para o subtipo B, tendo como ponto de corte o valor da distribuição superior do primeiro quartil. As posições assinaladas representam as posições de mutações selecionadas com freqüência de mutação superior ao primeiro quartil (> 0,97), considerando as freqüências em ordem crescente. Em função do número limitado de exemplos para o subtipo C, o valor de ponto de corte utilizado, inclui o valor igual ou maior que a distribuição do primeiro quartil (≥ 2,98). A tabela (6.7), mostra as posições de mutações selecionadas para o subtipo C

Tabela 6.6: Distribuição das freqüências de posições de mutações selecionadas pelo modelo AG/KDF para o NFV no subtipo B.

Posição	Freq.(%)	Posição	Freq.(%)	Posição	Freq.(%)
L10	4,35	R41	4,35	E65	0,48
T12	1,45	K43	0,48	H69	3,38
I13	2,42	K45	0,97	K70	0,97
K14	1,45	M46	2,90	A71	2,42
I15	5,32	I47	0,97	I72	3,38
L19	2,42	G48	0,97	G73	0,48
K20	2,90	G49	0,48	V74	0,48
L23	0,48	G51	0,48	V77	3,35
A28	1,93	F53	0,48	V82	1,45
D30	2,42	I54	2,90	I84	1,93
L33	0,97	K55	0,48	N88	2,90
E35	3,38	R57	1,45	L89	1,45
M36	6,28	I62	2,90	L90	1,45
S37	4,83	L63	4,83	I93	3,35
P39	0,97	I64	4,83		

Tabela 6.7: Distribuição das freqüências de posições de mutações selecionadas pelo modelo AG/KDF para o NFV no subtipo C.

Posição	Freq.(%)	Posição	Freq.(%)	Posição	Freq.(%)
L10	5,97	R41	1,49	K70	2,98
V11	1,49	K43	2,98	A71	4,48
T12	2,98	K45	4,48	T74	4,48
K14	4,48	M46	2,98	V82	4,48
G16	2,98	I54	1,49	N88	7,49
K20	4,48	R57	1,49	L89	4,48
D30	11,94	Q61	1,49	L90	4,48
E35	2,98	I62	2,98		
M36	5,97	L63	4,48		

A análise comparativa dos desempenhos obtidos pelo classificador KDF na previsão de resistência para o NFV nos subtipos B e C no conjunto de teste nas 20 simulações, com a utilização das posições de mutações mais freqüentes selecionadas pelo modelo proposto e os obtidos com a utilização das posições clássicas de resistência para o NFV: L10, D30, M36, M46, A71, V77, V82, I84, N88 e L90 descrita por JONHSON *et al*.(2008), são apresentados respectivamente, nas tabelas (6.8) e (6.9).

Tabela 6.8: Resultados obtidos na categorização de resistência para o NFV em pacientes do subtipo B, utilizando posições de mutações mais freqüentes selecionadas pelo modelo computacional AG/KDF e posições clássicas descritas na literatura.

	Modelo com 29 Posições			**Modelo com Posições Clássicas**		
Simulações	**Acurácia (%)**	**S (%)**	**E (%)**	**Acurácia (%)**	**S (%)**	**E (%)**
1	74,74	100,00	66,20	60,00	100,00	46,48
2	97,30	100,00	97,18	95,95	100,00	95,77
3	92,63	100,00	90,14	68,42	100,00	57,75
4	76,84	100,00	69,02	64,21	95,85	53,52
5	97,30	100,00	97,18	72,97	100,00	71,80
6	97,30	100,00	97,18	97,68	100,00	97,20
7	97,30	100,00	97,18	56,84	100,00	42,25
8	74,70	100,00	66,20	54,74	100,00	39,40
9	95,95	100,00	95,77	95,95	100,00	95,80
10	97,30	100,00	97,18	78,38	100,00	77,50
11	97,30	100,00	97,18	69,47	91,67	61,97
12	86,31	100,00	81,69	67,37	100,00	56,34
13	71,57	97,50	64,97	82,23	65,00	86,63
14	74,62	100,00	68,15	67,51	92,50	61,15
15	72,08	97,50	65,61	69,54	87,50	64,97
16	72,08	97,50	65,61	69,54	87,50	64,97
17	74,62	100,00	68,15	67,51	92,50	61,15
18	78,17	97,50	73,25	72,59	100,00	65,60
19	74,62	100,00	68,15	67,51	92,50	61,15
20	67,01	100,00	58,60	51,27	100,00	38,85

Tabela 6.9: Resultados obtidos na categorização de resistência para o NFV em pacientes do subtipo C, utilizando posições de mutações mais freqüentes selecionadas pelo modelo computacional AG/KDF e posições clássicas descritas na literatura.

	Modelo com 20 Posições			**Modelo com Posições Clássicas**		
Simulações	**Acurácia (%)**	**S (%)**	**E (%)**	**Acurácia (%)**	**S (%)**	**E (%)**
1	94,12	100,00	92,31	85,29	50,00	96,15
2	73,53	100,00	65,38	73,53	75,00	73,08
3	76,47	100,00	69,23	73,53	87,50	69,23
4	82,35	100,00	76,92	82,35	75,00	84,62
5	88,89	70,00	96,15	86,11	70,00	92,31
6	91,18	100,00	88,46	76,47	75,00	76,92
7	91,18	100,00	88,46	85,29	75,00	88,46
8	76,47	90,00	69,23	73,53	75,00	73,08
9	70,59	100,00	61,54	76,47	75,00	76,92
10	64,71	100,00	53,85	67,65	75,00	63,38
11	76,47	100,00	69,23	76,47	75,00	76,92
12	88,24	100,00	84,62	55,88	100,00	42,31
13	76,47	100,00	69,23	76,47	75,00	76,92
14	76,47	100,00	69,23	76,47	75,00	76,92
15	85,30	87,50	84,62	73,53	75,00	73,08
16	73,53	100,00	65,38	67,65	87,50	61,54
17	94,12	100,00	92,31	94,12	100,00	92,31
18	76,47	100,00	69,23	73,53	100,00	65,38
19	91,18	87,50	92,31	82,35	75,00	84,62
20	63,89	100,00	50,00	72,22	70,00	73,08

6.3 Lopinavir

O conjunto de dados dos pacientes resistentes a terapia antiretroviral pertencente ao subtipo B e experimentado ao inibidor LPV no último regime terapêutico, contêm 92 amostras, sendo 65 utilizadas no conjunto de treinamento e 27 no conjunto de teste. Para o subtipo C, o número de pacientes experimentados no último regime terapêutico ao LPV foi de 17 amostras, sendo 12 utilizadas no conjunto de treino e 5 no conjunto de teste. O grupo controle (*naive* na protease ou *naive* puro) para o subtipo B é composto de 236 amostras, sendo 165 utilizadas no conjunto de treino e 71 no conjunto de teste. Para o subtipo C esse grupo é composto de 88 amostras, sendo 62 empregadas no conjunto de treino e 26 no conjunto de teste.

O resumo do desempenho obtido na categorização de resistência ao LPV para cada uma das 20 simulações, ordenada em função da validação cruzada *leave-one-out*, bem como os resíduos de aminoácidos (posição de mutação) selecionados e o valor do parâmetro C, para os subtipos B e C, são apresentados respectivamente, nas tabelas (6.9) e (6.10).

Dentre as simulações realizadas no grupo de pacientes portadores do HIV-1 pertencentes ao subtipo B, a que apresentou melhor desempenho em termo da validação cruzada *leave-one-out*, obteve valor de 80,22 % na caracterização de resistência ao LPV, selecionando as posições de mutações: L10, T12, I15, L19, L24, E35, M36, S37, I50, L63, H69, A71, V77 e L90.

Já para o subtipo C, a simulação que obteve o melhor resultado na categorização de resistência para o LPV, selecionou as posições de mutação: L33, E35, M36, K45, M46, I54 A71, V82 e L90. Tendo conseguido, 84,93 % no desempenho na validação cruzada *leave-one-out*.

Tabela 6.10: Resumo dos resultados obtidos pelo modelo AG/KDF para a acurácia de treino, validação cruzada *leave-one-out* (LOOCV), acuraria de teste, parâmetro de regularização C e posições de mutações selecionadas para o LPV, tendo como base o subtipo B.

Simulação	Acurácia de Treino (%)	LOOCV (%)	Acurácia de Teste (%)	C	Posições de mutações selecionadas
18	79,91	50,66	84,85	100	10 15 20 35 36 41 43 54 61 63 77 90
14	74,67	54,93	67,68	10	10 12 14 15 17 19 30 35 36 37 57 71 77 82 90 93
13	73,36	58,86	68,69	100	10 12 14 15 17 32 35 37 41 46 57 63 72 77 82 90
10	72,05	61,00	67,68	10	10 12 15 18 35 41 46 55 57 62 64 69 71 72 73 74 77 82
20	74,24	62,31	62,53	10	12 13 20 35 41 45 46 57 63 64 69 71 72 74 77 89
3	73,76	62,75	72,31	0,1	15 20 33 35 41 62 63 69 70 90 93
1	76,81	65,08	72,31	0,1	15 41 54 62 63 64 70 77 90 93
2	77,95	65,46	72,31	0,1	15 20 35 46 53 62 69 73 74 77 89 90 93
7	72,49	65,81	69,70	100	15 19 35 36 37 41 46 53 57 61 62 63 71 72 77 93
17	76,86	66,64	67,68	10	14 15 35 36 37 41 62 63 70 71 73 82 90 93
4	77,19	66,65	72,31	0,1	13 17 54 57 63 64 70 71 73 77 93
8	73,36	67,12	67,68	0,1	15 17 19 20 36 37 41 46 62 63 77 90 93
11	79,48	67,99	65,86	100	12 14 15 17 35 37 41 46 62 63 64 69 71 77 93
12	74,67	69,69	67,68	100	15 17 20 34 35 36 41 62 63 69 70 72 77 88
15	73,36	70,13	71,72	10	10 12 13 15 30 36 37 62 63 70 71 72 77 82 90 93 95
6	79,91	71,05	86,87	10	15 36 37 41 46 50 60 62 63 64 69 77 82 88 91 93
19	80,35	71,48	87,38	100	10 19 20 35 37 41 46 50 60 63 64 70 72 77 90 93
16	78,60	71,92	87,88	100	10 20 33 35 37 41 43 46 54 57 60 63 64 70 71 77 89 90
5	74,24	72,24	68,69	10	13 14 15 19 20 35 36 41 60 62 63 70 77 90 93
9	81,22	80,22	84,85	0,1	10 12 15 19 24 35 36 37 50 63 69 71 77 90

A região hachurada representa as simulações que apresentaram valores superiores ao limite do primeiro `quartil no conjunto do erro LOOCV`.

Tabela 6.11: Resumo dos resultados obtidos pelo modelo AG/KDF para a acurácia de treino, validação cruzada *leave-one-out* (LOOCV), parâmetro de regularização C e posições de mutações selecionadas para o LPV, tendo como base o subtipo C.

Simulação	Acurácia de Treino (%)	LOOCV (%)	Acurácia de Teste (%)	C	Posições de mutações selecionadas
2	86,30	69,86	87,10	100	19 20 43
5	89,04	75,34	83,87	100	36 54 61 62 77 82
7	90,41	75,34	45,16	100	20 57 58 63
18	89,04	76,70	74,19	10	13 89
4	87,67	78,07	74,19	1	13 70
6	90,41	78,10	61,29	0,1	12 13 16 20 54
8	86,30	78,13	70,97	0,1	54 61 89
1	91,78	79,45	74,19	0,1	20 43 61 62 70
3	89,04	80,82	74,19	10	10 74 89
15	86,30	80,82	67,74	1	20 36 63 89
16	89,04	80,82	70,97	10	58 62 71
20	90,41	80,82	77,42	10	19 20 63 82
10	86,30	81,75	80,65	0,1	70 89
9	87,67	82,19	77,42	100	37 77 82
14	89,04	82,19	67,74	0,1	12 19 23 60
17	87,67	82,19	80,65	0,1	16 20 36
11	87,67	82,30	77,42	10	10 46 82
12	86,30	83,56	74,19	1	12 16 20 33 54 74
19	89,04	83,93	67,74	1	16 37 90
13	86,30	84,93	83,87	0,1	33 35 36 45 46 54 63 71 82 90

A região hachurada representa as simulações que apresentaram valores superiores ao limite do primeiro quartil no conjunto do erro LOOCV.

As simulações que apresentaram resultados no conjunto de validação cruzada *leave-one-out*, superiores ao valor limite estabelecido pelo ponto de corte (primeiro quartil), para os subtipos B e C estão assinaladas nas tabelas (6.10) e (6.11), respectivamente.

A tabela (6.12) mostra as distribuições de freqüência das posições de mutações da seqüência da protease, selecionadas pelo modelo computacional no conjunto das melhores simulações (LOOCV), para o inibidor de protease LPV em pacientes portadores do HIV-1 de subtipo B, tendo como ponto de corte o valor da distribuição superior do primeiro quartil. As posições assinaladas representam as posições de mutações selecionadas com freqüência de mutação superior ao primeiro quartil (> 0,93), considerando as freqüências em ordem crescente. Em função do número limitado de exemplos para o subtipo C, o valor de ponto de corte utilizado inclui o valor igual ou maior que a distribuição do primeiro quartil (≥ 3,07). A tabela (6.13) mostra as posições de mutações selecionadas para o subtipo C.

Tabela 6.12: Distribuição das freqüências de posições de mutações selecionadas pelo modelo AG/KDF para o LPV no subtipo B.

Posição	**Freq.(%)**	**Posição**	**Freq.(%)**	**Posição**	**Freq.(%)**
L10	1,86	S37	4,23	K70	4,23
T12	1,41	R41	5,16	A71	3,29
I13	1,41	K43	0,47	I72	1,86
K14	1,41	M46	3,29	G73	1,40
I15	5,63	I50	1,41	T74	0,47
G17	1,86	F53	0,94	V77	6,10
L19	2,35	I54	1,41	V82	1,41
K20	3,29	R57	1,88	N88	0,94
L24	0,47	D60	1,86	L89	0,94
D30	0,47	Q61	0,47	L90	4,23
L33	0,94	I62	5,16	T91	0,47
E34	0,47	L63	6,57	I93	6,10
E35	4,70	I64	2,82	C95	0,47
M36	3,76	H69	2,82		

Tabela 6.13: Distribuição das freqüências de posições de mutações selecionadas pelo modelo AG/KDF para o LPV no subtipo C.

Posição	Freq.(%)	Posição	Freq.(%)	Posição	Freq.(%)
L10	3,51	M36	3,51	I62	3,51
T12	5,26	S37	5,26	L63	5,26
I13	1,75	K43	1,75	A71	3,51
G16	7,02	K45	1,75	T74	3,51
L19	3,51	M46	3,51	V77	1,75
K20	10,53	I54	7,02	V82	7,02
L33	3,51	D60	1,75	L89	7,02
E35	1,75	Q61	3,51	L90	3,51

A análise comparativa do desempenho obtido pelo classificador KDF, na previsão de resistência para o LPV nos subtipos B e C, no conjunto de teste, nas 20 simulações, com a utilização das posições de mutações mais freqüentes, selecionadas pelo modelo proposto e o obtido com a utilização das posições clássicas de resistência para o inibidor LPV: L10, K20, L24, V32, L33, M46, I47, I50, F53, I54, L63, A71, G73, L76, V82, I84, e L90 descrita por JONHSON *et al*.(2008), são apresentados respectivamente, nas tabelas (6.14) e (6.15).

Tabela 6.14: Resultados obtidos na categorização de resistência para o LPV em pacientes do subtipo B, utilizando posições de mutações mais freqüentes selecionadas pelo modelo computacional AG/KDF e posições clássicas descritas na literatura.

	Modelo com 29 Posições			**Modelo com Posições Clássicas**		
Simulações	**Acurácia (%)**	**S (%)**	**E (%)**	**Acurácia (%)**	**S (%)**	**E (%)**
1	96,92	100,00	95,74	84,62	94,44	80,85
2	93,85	100,00	91,49	84,62	83,33	85,11
3	72,31	98,00	61,70	49,23	100,00	29,79
4	95,38	100,00	93,62	86,15	66,67	93,62
5	71,72	100,00	60,56	53,54	100,00	35,21
6	88,89	100,00	84,51	63,64	100,00	49,30
7	70,71	100,00	59,15	56,57	100,00	39,44
8	71,72	100,00	60,56	56,57	100,00	39,44
9	88,89	100,00	84,51	83,84	67,86	90,14
10	90,91	100,00	87,32	65,66	100,00	52,11
11	90,91	100,00	87,32	67,68	100,00	54,93
12	71,72	100,00	60,56	54,55	100,00	36,62
13	71,72	100,00	60,56	56,57	100,00	39,44
14	71,72	100,00	60,65	55,56	100,00	38,03
15	71,72	100,00	60,56	53,54	100,00	35,21
16	71,72	100,00	60,56	75,76	100,00	66,20
17	71,72	100,00	60,56	59,60	92,86	46,48
18	90,91	100,00	87,32	71,72	100,00	60,56
19	89,99	100,00	85,92	77,78	82,14	76,06
20	70,71	100,00	59,15	55,56	100,00	38,03

Tabela 6.15: Resultados obtidos na categorização de resistência para o LPV em pacientes do subtipo C, utilizando posições de mutações mais freqüentes selecionadas pelo modelo computacional AG/KDF e posições clássicas descritas na literatura.

	Modelo com 18 Posições			**Modelo com Posições Clássicas**		
Simulações	**Acurácia (%)**	**S (%)**	**E (%)**	**Acurácia (%)**	**S (%)**	**E (%)**
1	70,97	100,00	65,38	80,65	80,00	80,77
2	87,10	100,00	84,62	87,10	60,00	92,31
3	70,97	100,00	65,38	70,97	80,00	69,23
4	83,87	100,00	80,77	54,84	100,00	46,15
5	96,77	100,00	96,15	77,92	100,00	73,08
6	67,74	100,00	61,54	83,87	60,00	88,46
7	58,06	100,00	50,00	38,71	100,00	26,92
8	61,29	100,00	53,85	80,65	60,00	84,92
9	64,52	100,00	57,69	58,06	80,00	53,85
10	83,87	100,00	80,77	80,65	100,00	76,92
11	77,42	100,00	73,08	70,97	80,00	69,23
12	83,87	100,00	80,77	83,87	80,00	84,62
13	74,19	100,00	69,23	83,87	80,00	84,62
14	80,65	80,00	80,77	87,10	60,00	92,31
15	64,52	100,00	57,69	67,74	80,00	65,38
16	96,77	100,00	96,15	93,55	80,00	96,15
17	77,42	100,00	73,08	83,65	80,00	84,62
18	74,19	100,00	69,23	80,65	80,00	80,77
19	70,97	100,00	65,38	51,61	100,00	42,31
20	87,10	100,00	84,62	90,32	60,00	96,15

Nas tabelas (7.1 e 7.2) apresentamos a comparação entre os resultados obtidos pelo modelo computacional AG/KDF, em função da média (± desvio-padrão) para a acurácia (*ACC*), sensibilidade (*S*) e especificidade (*E*), com a utilização das posições de mutações mais freqüentes selecionadas pelo AG (modelo) e as mutações descritas por JONHSON (2007) para os inibidores de protease SQV, LPV e NFV para os subtipos B e C, no conjunto de teste nas 20 simulações.

Tabela 6.16: Comparação entre os resultados obtidos (média ± desvio-padrão) para o subtipo B, com a aplicação do classificador KDF, na previsão de resistência aos inibidores de protease, utilizando as posições de mutações mais freqüentes selecionadas pelo AG (Modelo) e as posições clássicas descritas pela literatura como promotoras de resistência para esses inibidores (Padrão - IAS).

	Subtipo B					
	Modelo AG/KDF			**Padrão - IAS**		
	SQV	**NFV**	**LPV**	**SQV**	**NFV**	**LPV**
ACC	$91,22 \pm 4,47^{(*)}$	$83,49 \pm 11,53^{(*)}$	$80,71 \pm 10,51^{(*)}$	$\mathbf{75,19 \pm 11,61^{(*)}}$	$71,48 \pm 13,03^{(*)}$	$65,64 \pm 12,44^{(*)}$
S	$99,61 \pm 1,26$	$99,50 \pm 1,03^{(*)}$	$99,90 \pm 0,45^{(*)}$	$98,00 \pm 5,23$	$95,25 \pm 8,40^{(*)}$	$94,37 \pm 10,73^{(*)}$
E	$90,00 \pm 5,08^{(*)}$	$79,23 \pm 14,83^{(*)}$	$73,12 \pm 14,65^{(*)}$	$72,03 \pm 13,58^{(*)}$	$65,01 \pm 17,94^{(*)}$	$54,33 \pm 20,60^{(*)}$

(*) Resultados significativamente diferente p-valor < 0,05.

Tabela 6.17: Comparação entre os resultados obtidos (média ± desvio-padrão) para o subtipo C, com a aplicação do classificador KDF, na previsão de resistência aos inibidores de protease, utilizando as posições de mutações mais freqüentes selecionadas pelo AG (Modelo) e as posições clássicas descritas pela literatura como promotoras de resistência para esses inibidores (Padrão - IAS).

	Subtipo C			
	Modelo AG/KDF		Padrão - IAS	
	NFV	LPV	NFV	LPV
ACC	80,58 ± 9,45	76,61 ± 10,92	**76,45 ± 8,18**	75,34 ± 14,44
S	96,75 ± 7,61(*)	99,00 ± 4,47(*)	78,25 ± 11,81(*)	80,00 ± 14,51(*)
E	75,38 ± 13,64	72,31 ± 13,12	75,86 ± 12,44	74,46 ± 19,11

Neste capítulo, foram apresentados os resultados obtidos com a aplicação do AG/KDF na seleção de posições mutações de resistência e previsão de resistência para os inibidores utilizados na terapia antiretroviral SQV, NFV e LPV. Para efeito comparativos, foram formulados dois modelos de classificação com a aplicação do KDF: o primeiro baseado nas posições mais freqüentes selecionadas no conjunto das 20 simulações realizadas e o segundo com as posições definidas pela IAS/2008. Posteriormente os resultados obtidos em função das métricas de acurácia (ACC), sensibilidade (S) e especificidade (E) foram avaliados em termos da média e desvio padrão, visando verificar a existência de diferença estatística significante (p-valor < 0,05). Os resultados finais mostram a existência de diferença estatística para os modelos propostos nos subtipos B e C.

(*) Resultados significativamente diferente p-valor < 0,05.

Capítulo 7

Discussão

O acúmulo de resistências às drogas antiretrovirais e as conseqüentes falhas terapêuticas são um problema mundial para o sucesso da terapia da AIDS. A replicação residual sob pressão seletiva resulta no aparecimento de mutações no genoma do HIV-1 que diminui a suscetibilidade aos fármacos, reduzindo progressivamente a potência dos componentes do esquema terapêutico. A seleção de novas posições de mutações na seqüência do gene *pol* da polimerase do HIV-1, que promovem ou colaboram na resistência a terapia antiretroviral possibilita o aumento da acurácia de classificação de resistência, possibilitando assim, a diminuição no custo terapêutico e o desenvolvimento de modelos mais eficientes na pesquisa de resistência antiretroviral.

Nesse trabalho apresenta-se um novo método de seleção de resíduos de aminoácidos (posições de mutação), na seqüência da protease do HIV-1 por algoritmo genético (AG), associado ao classificador de *Kernel* Discriminante de Fisher (KDF) na categorização de resistência. A codificação dos resíduos de aminoácidos pela escala de hidrofobicidade possibilitou o modelo computacional proposto levar em consideração as propriedades físico-químicas dos aminoácidos, propriedades estas que são independentes ao subtipo do HIV-1. A ponderação da escala de Kyte e Doolitte se fez necessário para evitar que o método não incorporasse aminoácidos com valores de hidrofobicidade equivalentes apesar da presença de mutação.

A utilização dos inibidores SQV, NFV e LPV neste estudo se deu em função primeiramente do número de pacientes experimentados no último regime terapêutico, há existência de posições de mutações de resistência excluídas as posições de assinatura com diferença significativa (p-valor < 0,05) entre as proporções para os subtipos B e C. No caso específico do inibidor SQV, apesar de não apresentar um número significativo de exemplos para o subtipo C, possui posições específicas de resistência como é o caso da G48 entre outras.

A utilização da seqüência de consenso HXB-2 para o subtipo B definida por TANURI (1999) e a seqüência definida por GAO (2001) para o subtipo C, pode ter contribuído para a melhora do nível de sensibilidade do modelo computacional no conjunto de teste para o subtipo C, quando comparado aos resultados obtidos com a utilização da seqüência HXB-2 como referência para o subtipo C.

A existência de diferença significativa entre as médias obtidas pelo modelo proposto utilizando as posições mais freqüentes de mutação selecionada pelo AG e as mutações de resistência descrita pelo IAS, foi verificada através da aplicação do teste de comparação das médias com p-valor < 0,05.

Os resultados médios obtidos na previsão com o modelo computacional proposto, utilizando as posições de mutações mais freqüentes selecionadas, apresentaram resultados significativamente diferentes $^{(*)}$ em termos da acurácia (*ACC*), sensibilidade (S) e especificidade (*E*) quando comparados aos obtidos com as posições de resistências, já descritas na literatura para os inibidores em estudo. Indicando assim, que entre as posições de mutações selecionadas pelo modelo computacional proposto e as posições já categorizadas possam existir algumas que estejam contribuindo para a melhora no desempenho do modelo.

O desempenho obtido na categorização de resistência ao LPV no conjunto de treinamento em função da validação cruzada *leave-one-out* (LOOCV) para o subtipo B apresentou resultado inferior ao obtido para o subtipo C. Fato que pode ser justificado em função da barreira genética do HIV-1 subtipo B para o LPV.

Apesar da utilização do ponto de corte maior que a freqüência de mutação superior ao primeiro quartil para o subtipo B, ter conseguido selecionar as principais mutações de resistências para o NFV e LPV. O mesmo não foi verificado para o inibidor SQV, visto que esse corte (maior que) não possibilitou ao modelo AG/KDF selecionar importantes posições de mutação de resistência, como é o caso das posições M46 e I84.

O número médio de posições de mutações selecionadas pelo AG, nas seqüências da protease de pacientes em falha terapêutica do subtipo B foi superior ao obtido no grupo de pacientes do subtipo C, indicando que o HIV-1 de subtipo C possui barreira genética menor, ou seja, essas cepas necessitam de um número menor de mutações para se tornarem resistentes. No caso específico do LPV que possui uma alta barreira genética, o modelo para o subtipo B selecionou no conjunto de simulações realizadas, uma média de 15 mutações, entretanto, não verificamos essa característica no subtipo C, onde o modelo conseguiu obter uma média de 4 mutações, para o conjunto total das simulações realizadas.

Quando comparamos os resultados obtidos na previsão de resistência para o inibidor de protease NFV nos subtipos B e C, tendo com entrada no classificador KDF as posições de mutação, definida pela literatura verificamos que a acurácia média obtida no subtipo B na categorização de resistência é superior ao valor obtido no subtipo C, levando a concluir que o perfil mutacional do subtipo C é diferente quando comparado ao subtipo B.

No caso da utilização das posições mais freqüentes selecionadas, apesar dos resultados obtidos para o NFV no subtipo C serem também inferiores aos obtidos no subtipo B, tal diferença pode ser explicada em função do número de amostras de subtipo C, em falha terapêutica, no último regime terapêutico ser bem menor quando comparada ao do subtipo B, fato que pode dificultar o processo de seleção e classificação pelo modelo.

Como a categorização de resistência ao LPV não está associada a determinadas posições de mutação, e sim a uma combinação de mutações, o desempenho obtido pelo classificador em pacientes portadores do HIV-1 de subtipo B apresentou desempenho inferior ao obtido nesse mesmo subtipo para os inibidores SQV e NFV.

Apesar da melhor simulação ter conseguido identificar a principal mutação para o inibidor NFV, a substituição da asparagina por ácido aspártico na posição 30 do gene da protease (D30N), a mesma não conseguiu identificar a mutação L90M (substituição da leucina por metionina), mutação que ocorre em alguns casos de falha terapêutica para o NFV, provocando uma resistência cruzada aos inibidores de protease (TUPINANBÁS *et al*., 2005). Entretanto, a mutação 90M foi selecionada em 20,00% das simulações realizadas para o NFV no subtipo B.

O baixo nível de seleção de mutação das posições V82 (15,00%) e I84 (20%) em pacientes do subtipo B, experimentados ao NFV no último regime terapêutico selecionado pelo modelo AG/KDF é corroborado pelo estudo de KANTOR *et al*. (2005). Em SHAFER (2002) argumenta-se que a mtação na posição V82 não tem qualquer efeito sobre a resistência fenotípica para o NFV, entretanto, contribui para a resistência associada a outras posições de mutação. Tal fato não se verifica no subtipo C, onde a seleção pelo modelo proposto para posição de mutação V82 apresenta alto nível de ocorrências, apesar de existir na base de dados utilizada nesse estudo, somente 20,00% dos casos com presença de mutação nessa posição.

O valor de ponto de corte maior ou igual ao valor do primeiro quartil para os inibidores de protease NFV e LPV no subtipo C foram aproximados (primeiro quartil = 3,0), mostrando uma homogeneidade do modelo para o subtipo C, fato este que corrobora os achados de SHAFER *et al* (2007).

A análise das posições de mutações da seqüência da protease, selecionadas pelo modelo computacional, para o SQV em pacientes do subtipo B, tendo como ponto de corte valor de freqüência de mutação superior a 1,31 (valor do 1º quartil), identificou 24 posições de mutações, dentre as quais selecionou importantes posições de resistência primária G48 e L90 e polimorfismos I13, K20, E35, M36, I62 e V82 descritas na literatura para esse inibidor.

Comparando a melhor simulação obtida pelo modelo AG/KDF para os subtipos B e C em pacientes tratados com NFV no último regime terapêutico, verificamos que o modelo conseguiu selecionar as principais posições de mutações de resistência a D30 e N88, indicando que estas posições estão correlacionadas entre si, independente do subtipo B ou C.

Analisando os resultados das posições de mutações selecionadas pelo modelo computacional proposto, verificou-se que o mesmo foi capaz de identificar posição de mutação específica para o subtipo C (T74), bem como, selecionar posições ainda não descritas na literatura, como promotora de resistência para os inibidores de protease em pacientes do subtipo B.

Dentre as possíveis posições selecionadas pelo modelo para o inibidor NFV, ainda não caracterizada como posição promotora de resistência (primária/compensatória) ou de polimorfismo, encontramos as posições I72 e L19 que aparecem com uma freqüência respectivamente a (3,38% e 2,44%) no conjunto das simulações, sinalizando que as mesmas possam ser uma nova posição de mutação associada a resistência antiretroviral para o NFV.

A posição de mutação I72 também foi selecionada pelo modelo para os inibidores SQV e LPV, com freqüência de (4,58% e 1,86 %) respectivamente. Em trabalho recente KING *et al*. (2007) associaram à referida posição a resistência ao LPV. Apesar da posição I72 estar associada conjuntamente aos diversos inibidores, a associação do inibidor ritonavir (RTV) aos inibidores SQV (SQV/r) e LPV (LPV/r), pode estar contribuindo para a seleção seletiva dessa posição de mutação.

A aplicação da escala de hidrofobicidade de aminoácidos de Kyte e Doolittle ponderada pelos respectivos pesos moleculares, na codificação dos aminoácidos da seqüência do gene da protease, mostrou-se eficiente, visto que o valor de hidrofobicidade é independente ao subtipo do HIV-1, além de possibilitar que aminoácidos com valores de hidrofobicidade iguais quando sofressem mutação fossem incorporados pelo modelo como, por exemplo, a mutação do aminoácido Asparagina (N) e Ácido aspártico (D).

Os resultados obtidos pelo modelo, em função da previsão de resistência utilizando a codificação pela escala de hidrofobicidade, são equivalentes aos encontrados na literatura, que utilizam outras técnicas de codificação (DRAGHICI e POTTER, 2003).

Apesar de nenhuma simulação ter utilizado o critério de parada referente ao número máximo de iterações, os resultados nos levam a acreditar que há a necessidade de se avaliar outros critérios de parada disponível no pacote *GEATbx*, visando permitir avaliar o modelo proposto com outro modelos de seleção.

Embora a incorporação das posições de mutação com freqüências superiores ao valor do primeiro quartil no modelo computacional tenha proporcionado valores de acurácia superiores aos valores obtidos pelo modelo que levou em consideração as posições de resistência já descritas no gene da protease aos inibidores utilizados nesse estudo, não podemos afirmar a priori que as referidas posições, sejam mutações que promovam resistência.

Capítulo 8

Conclusões

Nesse estudo, tratamos o problema da seleção de novas posições de mutação em seqüências da protease, na previsão de resistência à terapia antiretroviral nos subtipos B e C, um problema de difícil solução devido a grande complexidade mutacional do HIV-1.

A contribuição desse trabalho consistiu no desenvolvimento de um modelo computacional híbrido (AG/KDF) que combina a busca paralela dos AGs na seleção de atributos e o classificador de *kernel* de Fisher (KDF), na seleção de variáveis (posições de mutações) e previsão de resistência, em seqüências da protease de pacientes portadores do HIV-1 dos subtipos B e C em falha terapêutica no Brasil para os inibidores Saquinavir, Nelfinavir e Lopinavir.

Os resultados obtidos pelo modelo computacional AG/KDF mostraram-se promissores, quanto à utilização de algoritmos genéticos na seleção de variáveis (posições de mutação) associadas à resistência aos inibidores de protease saquinavir, nelfinavir e lopinavir, a partir de informações genotípicas do gene da protease do HIV-1, visto que a aplicação dessa técnica no modelo computacional proposto foi capaz de selecionar as principais posições de mutações de resistência para os inibidores em estudo.

A acurácia obtida pelo modelo na classificação da resistência a terapia antiretroviral do HIV-1 para o subtipo B, encontra-se próxima aos valores obtidos na literatura com a utilizações de classificadores não lineares (CAO *et al*.,2005).

A incorporação das posições de mutações mais freqüentes selecionadas pelo modelo AG/KDF no classificador KDF possibilitou um desempenho superior em todos os índices de avaliação de desempenho utilizado. Principalmente ao nível de sensibilidade, onde o valor médio obtido no conjunto das 20 simulações foi maior que 99,00% para os inibidores SQV, NFV e LPV em pacientes portadores do HIV-1 de subtipo B e maior que 95,00% para o subtipo C.

O modelo proposto foi capaz de selecionar posição de mutação específica do subtipo C, como as posições G16 e T74 que foram selecionadas simultaneamente para os inibidores NFV e LPV.

Não houve diferença estatística significante ao nível de p-valor <0,05 no desempenho do modelo AG/KDF na previsão de resistência, tendo como valores de entrada as posições de mutações mais freqüentes selecionadas pelo AG, para os subtipos B e C.

Apesar dos bons resultados obtidos com a aplicação do modelo na previsão de resistência em seqüências de subtipo C, há a necessidade de aumentar o número de amostras de pacientes resistentes pertencentes a esse subtipo.

Visando confirmar se as posições de mutações selecionadas pelo modelo computacional proposto são de resistência ou de polimorfismos, estão sendo desenvolvidos experimentos *in vitro* no Laboratório de Virologia Molecular (Dep. Genética/IB/UFRJ).

Como trabalhos futuros, pretende-se verificar: O do desempenho do modelo AG/KDF na caracterização e previsão para outros inibidores de protease, Darunavir e Tipranavir utilizados na terapia antiretroviral; Aplicação de novas técnicas de seleção de variáveis visando comparar a técnica de AGs em dados de resistência do HIV-1; Quanto ao AG proposto, pretendemos investigar outros operadores genéticos e outros critérios de parada, visando verificar se conseguimos melhorar o modelo final em termos de precisão; Desenvolvimento de aplicativo computacional baseado no modelo proposto para a plataforma .NET.

Referências Bibliográficas

ADESOKAN, A.A., ROBERTS, V.A., LEE, K.W., *et al*. ,2004. "Prediction of HIV-1 Integrase/Viral DNA Interactions in the Catalytic Domain by Fast Molecular Docking". *J. Med. Chem.*, v.47, pp. 821-828.

BANZHAF, W., NORDIN, P., KELLER, R., *et al*., 1998, *Genetics Programming: An Introduction*, San Mateo, Morgan Kaufmann.

BAXTER, D., MAYERS, D., WENT WORTH, D., *et al*., 2000. "A radomized study of antiretroviral management based on plasma genotypic antiretroviral failing therapy", *AIDS*, v.14(9), pp.83-92.

BENTLEY, P.J., 2002, *Digital Biology: How Natureis Transforming our Technology and your Lives*. NY, Simont Schuster. Inc.

BLICKLE, T., THIELE, L., A , 1995, "Comparison of Selection Schemes used in Genetics Algorithms", Technical Report 11, Computer Engineering and Communication Networks Lab (TIK), Swiss Federal Institute of Thechnology (ETA), Zürich, Switzerland, 2 nd edition.

BO, L., WANG, L., JIAO, L., 2006. "Feature Scaling for Kernel Fisher Discriminant Analysis Using Leave-One Cross Validation". *Neural Computation*, v.18, pp. 961-978.

BRINDEIRO, R., DIAZ, R., SABINO, E., *et al*., 2003. "Brazilian Network for HIV Drug Resistance Surveillance (HIV-BResNet): a survey of chronically infected individuals", *AIDS*, v.17, pp.1063-1069.

CAO, Z., W., HAN, L., Y., ZHENG, C., J., *et al*., 2005, "Computer Prediction of drug resistance mutations in proteins", *Drug Discovery Today: BIOSILICO*, v.7, pp. 521-529.

CARIDE, E., BRINDEIRO, R., HERTOGS, K., *et al*., 2000. "Drug-resistant reverse transcriptase genotyping and phenotyping of B and non-B subtypes (F and A) of human immunodeficiency virus type I found in Brazilian patients failing HAART", *Virology*, v.275, pp.107-115.

CARMONA, R., PÉREZ-ALVAREZ, L., MUÑOZ, M., *et al*., 2005. "Natural resistance-associated mutations to enfuvirtide(T20) and polymorphisms in the g41 region of different HIV-1 genetic formas from T20 naive patients", *Journal of Clinical Virology*, v.32, pp.248-253.

CDC, 2006. "HIV/AIDS Guidelines and Recommendations ", *Printable Vesion*.

CHAN, D., KIM, P., 1998. "HIV entry and its inhibition", *cell*, v.93, pp.681-684.

COFFIN, N.J., 1996, "Human Immunodeficiency Viruses and their replication. *Fundamental Virology", Fields* BN, Knipe DM, Howley PM, Eds. Lippincott-Raven, Philadelphia-NY, pp.845-916.

COURANT, R., HILBERT, D., 1953, *Methods of Mathematical Physics*, v. 1, Interscience Publishers Inc., New York.

CRAIGIE, R., 2001. "HIV Integrase, a Brief Overview from Chemistry to Therapeutics". *J. Biol. Chem*. V.276, pp. 23213-23216.

CRISTIANINI, N., SHAWE, T., 2000, *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

DALGLEISH, A.,BEVERLEY, P., CLAPHAM, P., *et al*., 1984, "The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus", *Nature*, v. 312, pp. 763-767.

DASH, M., LIU, H., 1999, "Handling large unsupervised data via dimensionality reduction". ACM SIGMOD workshop on research issues in data mining and knowledge discovery.

DeCOSTE, D., BURL, M., HOPKINS, A., *et al*., 2001, "Support Vector Machines and kernel Fisher Discriminants: A Case Study using Eletronic Nose Data.". *In: Fourth Workshop on Mining Scientific Dataset*.

DEFORCHE, K. SILANDER, T., CAMACHO, R., *et al*., 2006. "Analysis of HIV-1 pol sequences using Bayesian networks: implications for drug resistance". Bioinformatics, v22, p. 2975-2979.

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

DIRIENZO, G., DeGRUTTOLA, V., LARDER, B., *et al*., 2003. "Non-parametric methods to predict HIV drug susceptibility phenotype from genotype". Statistics in Medice, v. 22, p. 2785-2798.

DOMS, R., 2004. "Unwelcome Guests with Master Keys:How HIV Enters Cells and How it cam be Stopped", *International AIDS Society – Topics in HIV Medicine*, v.12, pp. 100-103.

DRAGHICI, S., POTTER, R., 2003, "Predicting HIV drug resistance with neural networks", *Bioinformatics*, v.19, pp.98-107.

ESHELMAN, L.J., 2000. *Genetics Algorithms*. Editora Bäck *et al*., v. 1, pp: 64-80, Chapter 8.

FILHO, P., POPPI, R., 2002, "Aplicação de AGs na seleção de variáveis em espectroscopia no infravermelho médio. Determinação simultânea de glicose, maltose e frutose", *Química Nova*, v.25(1), pp. 46-52.

FISHER, R., 1938, "The statistical utilization of multiple measurements". *In Annals of Eugenics*, v.8, pp. 376-386.

GALLO, R., 2002, "Historical essay. The early years of HIV/AIDS", *Science*. v. 298, pp. 1728-1730.

GAO, F., VIDAL, N., LI, Y., *et al*., 2001. "Evidence of Two Distinct Subsubtypes withis the HIV-1 Subtype A Radiation", *AIDS Research and Human Retroviruses*, v. 17(8), pp. 675-688.

GAO, F., YUE, L., WHITE, A., *et al*., 1992. "Human infection by genetically diverse SIVSM-related HIV-2 in west Africa", *Nature*, v. 358, pp.495-499.

GOLDBERG, D., 1989, *Genetics Algorithms in Serach, Optimization, and Machine Learning Reading*, MA, USA, Addison-Wesley.

GONDA, M. A., WONG-STALL, F., GALLO, R.C., *et al*., 1986. "Human T-Cell Lymphotropic Virus Type III Shares sequence Homology with a Family of Pathogenic Lentiviruses", *Proceedings of the National Academy of Sciences of the United States of America*, v.83, pp.4007-4011.

GUYON, I., WESTON, J., BARNHILL, S., *et al*., 2002, " Gene Selection for Cancer Classification using Support Vector Machines". *Machine Learning*, v.46, n. 1-3, pp. 389-422.

HAHN, B., SHAW, G., COCK, K., *et al*., 2000, "AIDS as a zoonosis: scientific and public health implications", *Science*. Review, v.287, pp.607-614.

HAYKIN, S., 2001, *Redes Neurais: Princípios e Prática*. 2nd ,Bookman.

HIRSCH, V., OLMSTED, R., MURPHEY-CORB, M., *et al*., 1989, "An African primate lentivirus (SIVsm) closely related to HIV-2", *Nature*, v. 339, pp.389-392.

HOLLAND, J., 1975, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI, USA, University of Michigan Press.

JAIN, A., ZONGRKER, D., 1997, "Feature-Selection: Evaluation, application, and small sample performance". IEEE *Transactions on Pattern Analysis and Machine Intelligence*, v.19(2), pp. 152-157.

JONHSON, A., FRANÇOISE, B., BONAVENTURA, C., *et al*., 2008, "Update of the Drug Resistance Mutations in HIV-1 :Spring 2008", Topics in HIV Medicine, v.16. p. 62-68.

KANTOR, R., KATZENSTEIN, D., EFRON, E., *et al*., 2005, "Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration". *PLoS Medicine*, v. 2, pp. 325-337.

KODIPAKA, S., VEMURI, B., RANGARAJAN, A., *et al*., 2007, "Kernel Fisher Discriminant for Shape Based Classification in Epilepsy.", *Med. Image Anal*, v.1191, pp. 79-90.

KYTE, R., DOOLITTLE, F.,1982, "A Simple Method for Displaying the Hidropathic Character of a Protein", *Journal Molecular Biology*, pp. 157-165.

LI, L., WEINBERG, C. R., DARDEN, T. A., *et al*.,2001, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/kNN method", *Bioinformatics*, v.17, pp. 1131-1142.

LUCIW, P.A., 1996, "Retroviridae: The virus and their replication", *In Fundamental Virology*, Fields BN, Knip DM, Howley PM, *et al*, Lippincott-Raven: Philadelphia, pp. 763-843.

MADDON, P., DALGLEISH, A., MCDOUGAL, J., S., *et al*., 1986, "The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain", *Cell*, v. 47, pp.333-48.

MATHIAS, K., WHITLEY, L.,D.,1994, "Changing representations during serach a comparative study of delta coding", In: Bäck, T., Fogel, D.B., Michalewicz Z., (eds.), *Evolutionary Computation 2: Advanced Algorithms an Operators*.

MAYR, E., 1987, *Toward a New Philosophy of Biology: Observations of on Evolutionist* . Cambridge, MA, USA, Belknap.

MERLUZZI, V., HARGRAVE, K., LABADIA, M., *et al*., 1990. **"**Inhibition of HIV-1 replication by a nonnucleoside reverse transcriptase inhibitor", *Science*, v.250, pp.1411-1413.

MIKA, S., RÄTSCH, J., WESTON, B., *et al.*, 1999, "Fisher discriminant analysis with kernels.", *In Neural Networks for Signal Processing*, v.9, pp.

MIKA, S., RÄTSCH, G., MÜLLER, K.,R.,2001, "A mathematical programming approach for kernel fisher discriminants". In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, v. 13, pp. 591-597.

MINISTÉRIO DA SAÚDE DO BRASIL, 2003. Programa Nacional de DST/AIDS. Recomendações para terapia anti-retroviral em adultos e adolescentes infectados pelo HIV – 2002/2003.

MINISTÉRIO DA SAÚDE DST/AIDS (http://www.aids.gov.br), acessado em 03/2007.

MORGADO, M., GUIMARAES, M., GRIPP, C., *et al*., 1998. "Molecular epidemiology of HIV-1 in Brazil: high prevalence of HIV-1 subtype B and identification of an HIV-1 subtype D infection in the city of Rio de Janeiro, Brazil. Evandro Chagas Hospital AIDS Clinical Research Group". *J Acquir Immune Defic Syndr Hum Retrovirol*, v.18, pp.488-494.

MORGADO, M., 2000. "A Diversidade do HIV na América do Sul", *Boletim Vacinas*, v. 5, pp.28-30.

NELSON, D.L., COX, M.M., 2000. *Lehninger Principles of Biochemistry*. fourth edition. W.H. Freeman.& Co.

NEUMANN, L.G., ROSA, T.F., JUNIOR, V.S., *et al*., 2004. "Otimização combinatorial empregando algoritmo genético aplicada na análise multivariada de medicamentos manipulados". *XXIV ENRGEP*, pp. 3175-3182.

PEÇANHA, E., ANTUNES, O., TANURI, A., 2002. "Estratégias Farmacológicas para a terapia anti-AIDS", *Química Nova*, v.25, pp. 1108-1116.

PEETERS, M., 2000. "Recombinant HIV sequences: Their Role in the Global Epidemic", *Laboratoire Retrovirus – Reviews*, pp. 39-54.

PERELSON, A., ESSUNGER, P., CAO, Y., *et al*., 1997. "Decay characteristics of HIV-1-infected compartments during combination therapy" *Nature*, v. 387, pp.188-191.

PETROPOULOS, C., PARKIN, T., LIMONI, K., *et al*. 2000. "A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1", *Antimicrob ts Chemoter*, v.44, pp. 920-928.

PINTO, M., STRUCHINER, C., 2006.“ A diversidade do HIV-1: Uma ferramenta para o estudo da pandemia”, C*aderno de Saúde Pública*, Rio de Janeiro, v.22, pp.473-484.

POHLHEIM, H., 2007, “GEATbx : Genetic and Evolutionary Algorithm Toolbox for Use with Matlab”. Disponível em http://www.geatbx.com/ . Acesso em 18 nov. 2007.

REEVES, C.R., ROWE, J.E., 2002. *Genetics Algorithms – Principles and Perspectives. A Guide to GA Theory*. Vol.20, Springer.

RICHMAN, D., WRIN, T., PETROPOULOS, C., *et al*., 2003. “Rapid evolution of the neutralizing antibody response to HIV type 1 infection”, *PNAS*, v.100(7), pp. 4144-4149.

SAIGO, H., UNO, T., TSUDA, K., 2007. “Mining Complex Genotype Features for Predicting HIV-1 Drug Resistance”. *Bioinformatics*, v.23, pp: 2455-2462.

SANTOS, N.S., ROMANOS, M.T., WIGG, M.D., 2002. *Introdução a Virologia Humana*. Rio de Janeiro, Editora Guanabara Koogan, v.1.

SEVIN, A.D., DeGRUTTOLA, V., NIJHUIS, M.., *et al*., 2000. “Methods for investigation of the relationship between drug-susceptibility phenotype and human immunodeficiency virus type 1 genotype with applications in AIDS clinical trials group 333”. *J. Infect. Dis*., v.182, pp. 59-67.

SHAFER, R.W., RHEE, S.Y., PILLAY, D., *et al*., 2007. “HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance”, *AIDS*, v.21, pp. 215-223.

SHAFER, R. W., 2002, “Genotypic testing for human immunodeficiency virus type 1 drug resistence”, *Clinical Microbiology Reviews*, v.15, pp.247- 277.

SING, T., BEERENWINKEL, N., 2007, *Mutagenetic tree Fisher kernel improves prediction of HIV drug resistance from viral genotype*, Advances in Neural Information Processing Systems 19, MA, MIT, USA, pp. 1-9.

SOARES, M., DE OLIVEIRA. T., BRINDEIRO, R., *et al*., 2003. “A specific subtype C of human immunodeficiency virus type 1 circulates in Brazil”, *AIDS* v.17, pp.11-21.

SOFGE, D.A., 2002, "Using Genetic Algorithm Based Variable Selection to Improve Neural Network Models for Real-Word Systems", *Proceedings of the International Conference on Machine Learning & Applications*, pp.1-4.

SOUZA, M., ALMEILDA, M., 2003. "Drogas Anti-VIH: Passado, Presente e Perspectivas Futuras", *Química Nova*, v.26, pp.366-372.

SYSWERD, G., 1989, "Uniform crossover in genetic algorithms", *In ICGA3*, pp. 2-9.

TAN, F., FU, X., ZANG, Y., *et al*., 2008, "A genetic algorithm-based method for feature subset selection". *Soft. Comput*., v.12, pp.111-120.

TANURI, A., SWANSON, P., DEVARE, S., *et al*., 1999. "HIV-1 subtypes among blood donors from Rio de Janeiro, Brazil", *J Acquir Immune Defic Syndr Hum Retrovirol* v. 20, pp. 60-66.

TUPINANBÁS, U., ALEIXO, A.,GRECO, D., 2005, "HIV-1 Genotype Related to Failure of Nelfinavir as the First Protease Inhibitor Treatment", *The Journal Infectious Diseases*, v.9, pp.62-68.

TURNER, B., SUMMERS, M., 1999. "Structural biology of HIV", *J Mol Biol.*, v.285, pp. 1-32.

UNAIDS, 2008. "AIDS epidemic update", pp. 1-9. (http://www.unaids.org).

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.

WANG, D., LARDER, B., 2003. "Enhanced predictions of lopinavir resistance from genotype by use of artificial neural networks". *J. Infect. Dis.*, v.188, pp.653-660.

WLODAWER, A., GUSTCHINA, A., 2000. "Structural and biochemical studies of retroviral proteases", *Biochimica et Biophysica Acta*, v. 1477, pp. 16-34.

YANG, J., FRANGI, J., ZHANG, D., *et al*., 2005, "KPCA plus LDA: A complete Kernel Fisher Discriminant Framework for Feature Extraction and recognition", IEEE *Transactions on Pattern Analysis and Machine Intelligence*, v. 27(2), pp. 230-243.

ZHANG, J., MA, K., 2004, "Kernel Fisher Discriminant for Texture Classification", *School of Electrical and Electronic Engineering*.

ANEXO I

Distribuição de freqüência de mutação de resistência no gene da protease em relação ao último esquema terapêutico. Os valores assinalados com $^{(*)}$ e $^{(**)}$ indicam a ocorrência de diferença significativa (p-valor < 0,05) entre as proporções para os subtipos (BxC) e (BxF), excluindo posições associadas a assinatura molecular dos subtipos C (I15V, M36I, R41K, H69K e L89M) (GONZALES, *et al*., 2003) e F (I15V, E35D, M36I, R41K, R57K, Q61N, L63P e L89M) (TANURI, 1999).

Distribuição da freqüência de posições de mutações na protease de pacientes tratados com SQV no último regime terapêutico.

Posição	Subtipo B (%)	Subtipo C (%)	Subtipo F (%)
L10	58,80	66,70	60,00
I13	20,60	33,30	0,00
G16	11,80	33,30	0,00
K20	47,10	33,30	60,00
L23	0,00	33,30	0,00
L24	11,80	33,30	0,00
D30	0,00	0,00	0,00
V32	0,00	0,00	0,00
L33	8,80	33,30	0,00
E35	38,20	66,70	100,00
M36	50,00	100,00	100,00
K43	5,90	0,00	0,00
M46	38,20	33,30	40,00
I47	0,00	0,00	0,00
G48	5,90	0,00	0,00
I50	0,00	0,00	0,00
F53	20,60	33,30	20,00
I54	29,40	66,70	40,00
Q58	0,00	0,00	0,00
D60	26,50	0,00	40,00
I62	26,50	33,30	0,00
L63	88,20	100,00	60,00
H69	23,50	100,00	0,00
A71	61,80	66,70	40,00
G73	23,50	0,00	0,00
T74	14,70	33,30	40,00
L76	0,00	0,00	0,00
V77	23,50	33,30	40,00
V82	32,40	0,00	20,00
N83	0,00	33,30	0,00
I84	41,20	33,30	20,00
I85	2,90	0,00	20,00
N88	8,80	0,00	20,00
L89	2,90	100,00	60,00
L90	67,60	66,70	40,00
I93	32,40	100,00	40,00

Distribuição da freqüência de posições de mutações na protease de pacientes tratados com IDV no último regime terapêutico.

Posição	Subtipo B (%)	Subtipo C (%)	Subtipo F (%)
L10	47,50	33,30	66,70
I13	18,60	16,70	11,10
G16	8,50	33,30	11,10
K20	32,20	66,70	77,80
L23	0,00	0,00	0,00
L24	6,80	0,00	11,10
D30	1,70	0,00	0,00
V32	5,10	0,00	22,20
L33	5,10	0,00	11,10
E35	28,80	16,70	100,00
M36	32,20	83,30	88,90
K43	10,20	0,00	11,10
M46	40,70	0,00	22,20
I47	0,00	0,00	11,10
G48	0,00	0,00	11,10
I50	0,00	0,00	0,00
F53	0,00	0,00	22,20
I54	22,00	16,70	55,60
Q58	8,50	0,00	22,20
D60	18,60	16,70	22,20
I62	39,00	16,70	44,40
L63	89,80	83,30	77,80
H69	15,30	100,00	22,20
A71	42,40	16,70	66,70
G73	16,90	0,00	0,00
T74	6,80	0,00	33,30
L76	1,70	0,00	11,10
V77	22,00	0,00	0,00
V82	33,90	33,30	77,80
N83	0,00	0,00	0,00
I84	15,30	16,70	0,00
I85	10,20	0,00	11,10
N88	5,10	0,00	0,00
L89	6,80	100,00	77,80
L90	32,20	16,70	44,40
I93	42,40	100,00	77,80

Distribuição da freqüência de posições de mutações na protease de pacientes tratados com NFV no último regime terapêutico.

Posição	Subtipo B (%)	Subtipo C (%)	Subtipo F (%)
L10	44,70(*)	17,60(*)	65,20
I13	36,40	47,10	39,10
G16	1,50(*)	14,70(*)	21,70
K20	40,20(*),(**)	64,70(*)	87,00(**)
L23	2,30	8,80	0,00
L24	2,30	0,00	4,30
D30	43,20	29,40	21,70
V32	2,30	0,00	0,00
L33	9,10	5,90	4,30
E35	50,80	61,80	100,00
M36	56,80	88,20	100,00
K43	5,30	5,90	8,70
M46	24,20	11,80	26,10
I47	0,80	0,00	0,00
G48	0,00	0,00	0,00
I50	0,80	0,00	0,00
F53	1,50	0,00	0,00
I54	10,60	8,80	21,70
Q58	7,60	5,90	0,00
D60	13,60	5,90	30,40
I62	37,10	32,40	26,10
L63	87,90(*)	64,70(*)	52,20
H69	14,40	100,00	8,70
A71	37,10	29,40	26,10
G73	5,30	0,00	4,30
T74	15,9(*)	50,00(*)	30,40
L76	0,00	0,00	0,00
V77	40,20(*),(**)	0(*)	8,70(**)
V82	9,10	17,60	8,70
N83	0,00	0,00	8,70
I84	3,80	0,00	0,00
I85	7,60	0,00	4,30
N88	37,90	20,60	21,70
L89	5,30	76,50	52,20
L90	34,10	44,10	52,20
I93	34,80(*)	100,00(*)	13,00

Distribuição da freqüência de posições de mutações na protease de pacientes tratados com LPV no último regime terapêutico.

Posição	Subtipo B (%)	Subtipo C (%)	Subtipo F (%)
L10	64,10(*)	25,00(*)	87,50
I13	35,90(*)	0,00(*)	12,50
G16	13,00	18,70	25,00
K20	38,00(**)	56,20	81,20(**)
L23	2,20	6,20	0,00
L24	7,60	0,00	18,70
D30	5,40	0,00	6,20
V32	5,40	0,00	0,00
L33	21,70	18,70	12,50
E35	35,90	18,70	93,70
M36	55,40	93,70	100,00
K43	9,80	6,20	12,50
M46	41,30	25,00	43,70
I47	13,00	0,00	0,00
G48	6,50	0,00	6,20
I50	6,50	0,00	6,20
F53	6,50	0,00	12,50
I54	48,90	25,00	56,20
Q58	13,00	12,50	6,20
D60	12,00	6,20	6,20
I62	55,40(*)	18,70(*)	25,00
L63	84,80	68,70	68,70
H69	12,00	100,00	0,00
A71	42,40(*)	12,50(*)	25,00
G73	17,40	6,20	6,20
T74	10,90	12,50	18,70
L76	5,40	0,00	6,20
V77	26,10(*),(**)	0,00(*)	0,00(**)
V82	44,60	31,20	68,70
N83	3,30	0,00	0,00
I84	14,10	0,00	0,00
I85	3,30	0,00	6,20
N88	6,50	0,00	6,20
L89	5,40	87,50	62,50
L90	28,30	18,70	37,50
I93	39,10(*)	100,00(*)	31,20

Distribuição da freqüência de posições de mutações na protease de pacientes tratados com APV no último regime terapêutico.

Posição	Subtipo B (%)	Subtipo C (%)	Subtipo F (%)
L10	74,10	71,40	75,00
I13	44,40	14,30	75,00
G16	11,10	28,60	25,00
K20	33,30(*)	100,00(*)	75,00
L23	3,70	0,00	0,00
L24	7,40	28,60	0,00
D30	0,00	0,00	0,00
V32	18,50	0,00	0,00
L33	37,00	0,00	25,00
E35	29,60	71,40	100,00
M36	40,70	100,00	100,00
K43	11,10	0,00	0,00
M46	63,00	42,90	25,00
I47	11,10	0,00	0,00
G48	0,00(*)	28,60(*)	0,00
I50	11,10	14,30	0,00
F53	7,40	0,00	25,00
I54	59,30	57,10	75,00
Q58	14,80	0,00	0,00
D60	11,10	0,00	50,00
I62	55,60(*)	0,00(*)	75,00
L63	92,60	85,70	100,00
H69	11,10	100,00	0,00
A71	48,10	42,90	50,00
G73	22,20	0,00	25,00
T74	14,80	57,10	25,00
L76	11,10	0,00	25,00
V77	25,90	0,00	0,00
V82	55,60	71,40	75,00
N83	0,00	0,00	0,00
I84	22,20	0,00	25,00
I85	3,70(**)	14,30	50,00(**)
N88	0,00	0,00	0,00
L89	14,80	75,70	75,00
L90	37,00	14,30	25,00
I93	37,00(*)	100,00(*)	50,00

Distribuição da freqüência de posições de mutações na protease de pacientes tratados com ATZ no último regime terapêutico.

Posição	Subtipo B (%)	Subtipo C (%)	Subtipo F (%)
L10	36,80	44,40	33,30
I13	23,70	22,20	33,30
G16	13,20	44,40	33,30
K20	36,80	33,30	66,70
L23	2,60	0,00	0,00
L24	5,30	0,00	0,00
D30	7,90	22,20	33,30
V32	7,90	0,00	0,00
L33	15,80	11,10	0,00
E35	44,70	66,70	100,00
M36	50,00	77,80	100,00
K43	5,30	11,10	33,30
M46	36,80	0,00	33,30
I47	2,60	0,00	0,00
G48	0,00	11,10	33,30
I50	21,10	33,30	0,00
F53	10,50	0,00	33,30
I54	13,20	22,20	66,70
Q58	0,00	11,10	0,00
D60	7,90	22,20	0,00
I62	34,20	11,10	0,00
L63	81,60	66,70	33,30
H69	23,70	100,00	0,00
A71	57,90	33,30	33,30
G73	23,70	11,10	0,00
T74	7,90	33,30	33,30
L76	0,00	0,00	0,00
V77	34,20	22,20	0,00
V82	10,50	22,20	33,30
N83	2,60	11,10	0,00
I84	7,90	0,00	0,00
I85	10,50	0,00	0,00
N88	15,80	33,30	33,30
L89	7,90	77,80	33,30
L90	34,20	11,10	0,00
I93	31,60(*)	100,00(*)	33,30