COPPE/UFRJ

ANÁLISE DE AGLOMERADOS DAS SEQUÊNCIAS DA PROTEASE DO HIV-1 DE INDIVÍDUOS INFECTADOS UTILIZANDO MAPA AUTO ORGANIZÁVEL E K-MÉDIAS

Roberta Cândido de Souza Fernandes

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

Orientador: Flávio Fonseca Nobre

Rio de Janeiro
Maio de 2010

ANÁLISE DE AGLOMERADOS DAS SEQUÊNCIAS DA PROTEASE DO HIV-1 DE INDIVÍDUOS INFECTADOS UTILIZANDO MAPA AUTO ORGANIZÁVEL e K-MÉDIAS

Roberta Cândido de Souza Fernandes

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

Examinada por:

_____

Prof. Flávio Fonseca Nobre, Ph.D.

_____

Prof. Rodrigo de Moraes Brindeiro, D.Sc.

_____

Prof. Antônio Maurício Ferreira Leite Miranda de Sá, D.Sc

RIO DE JANEIRO, RJ - BRASIL
MAIO DE 2010

Fernandes, Roberta Cândido de Souza

Análise de aglomerados das sequências da protease do HIV-1 de indivíduos infectados utilizando mapa auto organizável e k-médias/ Roberta Cândido de Souza Fernandes – Rio de Janeiro: UFRJ/COPPE, 2010.

XI, 69 p.: il.; 29,7 cm.

Orientador: Flávio Fonseca Nobre

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia Biomédica, 2010.

Referências Bibliográficas: p. 64 - 69.

1. Resistência. 2. Análise de cluster. 3. SOM. I. Nobre, Flávio Fonseca. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Biomédica. III. Titulo.

Agradecimentos

Primeiramente, agradeço a DEUS pelo término de mais esta etapa em minha vida.

Agradeço aos meus pais e irmãos que sempre estiveram presentes no decorrer deste trabalho, em especial à minha mãe, que sempre me incentiva nas minhas decisões.

Ao meu orientador Prof. Flávio Fonseca Nobre pelos ensinamentos, paciência, dedicação, carinho e amizade.

Ao professor Rodrigo Brindeiro e demais colegas do Laboratório de Virologia Molecular do CCS/UFRJ que gentilmente cederam o material biológico considerado neste estudo e ao CNPq pelo apoio financeiro.

À amiga de graduação Bárbara Dias, que me incentivou na escolha do Programa de mestrado e aos amigos da Fiocruz Caio e Roberta que participaram da minha decisão em vir para a COPPE.

Aos demais amigos do PEB que tornaram a rotina de estudos agradável, em um ótimo ambiente de crescimento intelectual e pessoal e pela amizade, destacando-se a Carol, Fernanda, Aninha, Ângelo.

À minha companheira da disciplina de redes neurais e hoje, grande amiga, Adriane, pelo apoio, conselhos e ensinamentos que tanto me auxiliaram no desenvolvimento deste trabalho.

Aos demais amigos presentes na minha vida e professores do programa que sempre contribuíram e me apoiaram neste período de grande dedicação.

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

ANÁLISE DE AGLOMERADOS DAS SEQUÊNCIAS DA PROTEASE DO HIV-1 DE INDIVÍDUOS INFECTADOS UTILIZANDO MAPA AUTO ORGANIZÁVEL E K-MÉDIAS

Roberta Cândido de Souza Fernandes

Maio/2010

Orientador: Flávio Fonseca Nobre

Programa: Engenharia Biomédica

Este trabalho propõe a separação de seqüências da protease do HIV-1 na identificação de grupos naturais de acordo com o perfil de aminoácidos semelhantes, que podem ser úteis para a caracterização de resistência às drogas. Foi utilizada a técnica Mapas auto-organizáveis (*Self Organizing Map*, SOM) para redução da dimensionalidade, seguido da análise de aglomerados utilizando k-médias. A seleção de variáveis (posições da protease) mantidas no modelo foi feita observando a variabilidade e utilizando a concordância Kappa na escolha destes atributos. O número de grupos e a avaliação dos grupos foram obtidos comparando os métodos Davies-Bouldin e Silhueta. Ambos os índices mostraram que as sequências poderiam ser divididas em três grupos. A análise destes grupos apresentou uma boa separação dos dados, com um grupo, contendo apenas indivíduos resistentes e mais do subtipo B e um segundo grupo, com 93,29% de seus indivíduos caracterizados pelo subtipo C. Analisando o perfil de mutações e polimorfismos de cada grupo, observamos características concordantes com descrições na literatura.

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

ANALYSIS OF CLUSTERS OF PROTEASE SEQUENCES OF HIV-1 INFECTED INDIVIDUALS USING SELF ORGANIZING MAP AND K-MEANS

Roberta Cândido de Souza Fernandes
May/2010

Advisor: Flávio Fonseca Nobre

Department: Biomedical Engineering

This paper proposes the separation of the protease sequences of HIV-1 in the identification of natural groups according to the amino acid profile similar, which may be useful for the characterization of drug resistance. Technique was used Self Organizing Map (SOM) to reduce the dimensionality, followed by cluster analysis with k-means. The selection of variables in the model was made by observing the variable and using the Kappa in the choice of these attributes. The number of groups and assessment of groups were obtained by comparing the methods Davies-Bouldin and Silhouette, both showed that the sequences could be divided into three groups. The analysis of these groups showed a group composed only of individuals resistant and majority subtype B and a second group, with 93.29% of individuals characterized by subtype C. Analyzing the profile of mutations and polymorphisms in each group, we found features consistent with descriptions in the literature.

SUMÁRIO

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AIDS – Síndrome da Imunodeficiência Adquirida (*Acquired Immune Deficiency Syndrome)*

ANN – Rede Neural Artificial (A*rtificial Neural Networks)*

AZT - Zidovudina

BPTT - *Backpropagation through time* (BPTT).

BRNN – Redes Neurais Recorrentes Bidirecionais *(Bidirectional Recurrent Neural Networks)*

CBA – Classificador baseado em Associação (*Classification Based Association*)

CDC - Centro de Controle de doenças (*Center of Desease Control*)

CRF – Formas Circulantes Recombinantes (C*irculating Recombinant Forms)*

CTL – Células T com atividade citotóxica

DB - Davies-Bouldin

DNA - Ácido desoxirribonucéico (D*eoxiribonucleic Acid*)

DT – Árvore de Decisão (*Decision Tree)*

HAART - Terapia com alta atividade antiretroviral (*Highly Active Antiretroviral Therapy)*

HIV – Vírus da Imunodeficiência Humana (*Human Immunodeficiency Virus*)

IAS – Sociedade Internacional de AIDS *(International AIDS Society)*

KNN – Técnica de K-vizinhos mais próximos *(K-Nearest Neighbor Techinique)*

ML – Aprendizado de Máquina *(Machine Learning)*

MLP - *Multilayers Perceptrons*

NNRTI – Inibidores de Transcriptase Reversa não-nucleosídicos *(Non Nucleoside Reverse Transcriptase)*

NRTI – Inibidor Nucleosídico de Transcriptase Reversa *(Nucleoside Reverse Transcriptase Inhibitors)*

PI – Inibidor de Protease *(Protease Inhibitor)*

RT – Transcriptase Reversa *(Reverse Transcriptase)*

SIV – Vírus da Imunodeficiência Símia *(Simian Immunodeficiency Virus)*

SOM – Mapa Auto-organizável (S*elf Organizing Map)*

SVM – Máquina de Vetor de Suporte (S*uport Vector Machine)*

Capítulo 1

INTRODUÇÃO

A infecção pelo vírus da imunodeficiência humana (H*uman Immunodeficiency Virus,* HIV) atinge cerca de 33 milhões de pessoas ao redor do mundo, incluindo aquelas cuja infecção progrediu para a síndrome da imunodeficiência adquirida (A*cquired Immune Deficiency Syndrome,* AIDS), segundo o Boletim epidemiológico anual de AIDS de 2007. A África concentra 23 milhões destes indivíduos e mais de 1,5 milhões residem na América Latina (UNAIDS, 2009).

Somente no ano de 2007, 2,7 milhões de novos casos da infecção foram notificados, enquanto cerca de 2,0 milhões de indivíduos, entre adultos e crianças, foram a óbito (UNAIDS, 2009). Na América Latina, a epidemia HIV/AIDS acomete hoje em torno de 1,6 milhões de indivíduos, com 100.000 novas infecções notificadas em 2007. No Brasil, assim como no resto do mundo, os números desta epidemia são bastante significativos, representando um terço dos indivíduos infectados na América Latina (UNAIDS, 2009). A figura 1.1 mostra a distribuição mundial de pessoas vivendo com o HIV em 2007.

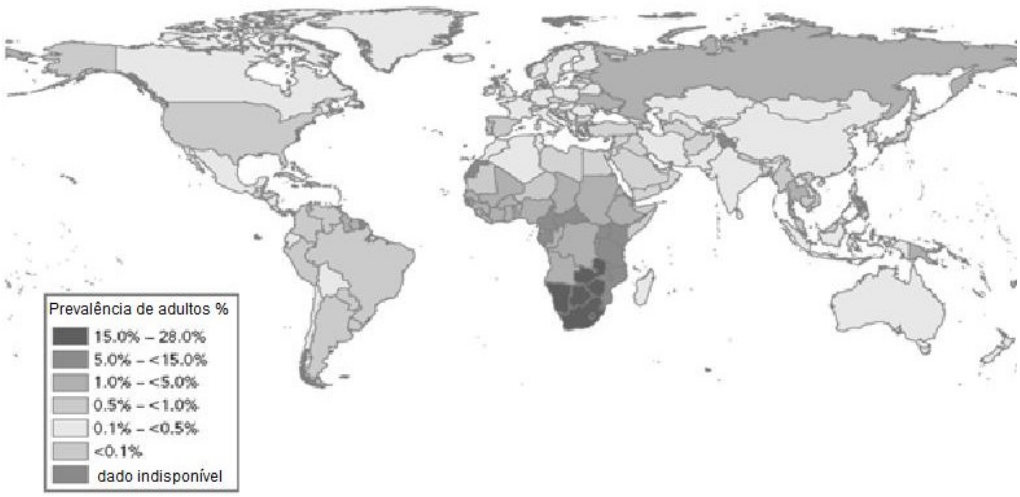

Figura 1.1: Estimativa de pessoas - 33 milhões [30-36 milhões] - vivendo com HIV em 2007 – Adaptada de UNAIDS (2009) – *A global view of HIV infection*

A análise filogenética do HIV em humanos classifica o vírus em dois grandes tipos HIV-1 e HIV-2. O HIV-1 é o de maior ocorrência em todo o mundo e pode ser classificado em diferentes subtipos de acordo com características associadas a variações genéticas (KANTOR e KATZENSTEIN, 2004).

Dentre os subtipos do HIV-1, o mais prevalente no mundo é o subtipo C, que concentra a sua área de ocorrência em regiões muito populosas e com altas taxas de infecção pelo HIV-1, como o Sul da África e Índia. Nos países centrais, Europa e Estados Unidos, o subtipo circulante predominante é o B. No Brasil, os subtipos virais mais freqüentes são o B e F (WALERIA-ALEIXO *et al,* 2008, GONZALEZ *et al,* 2007), seguidos do subtipo C, que se concentra principalmente na região Sul do país (BRINDEIRO *et al*, 2003).

O advento da terapia com alta atividade antirretroviral (*Highly Active Antiretroviral Therapy,* HAART) ocasionou a diminuição da mortalidade e morbidade pela infecção do HIV-1, no entanto, a evolução do vírus revela mecanismos de escape e mutações que acarretam resistência viral e consequente falha terapêutica (RAMBAUT *et al,* 2004).

Esta grande diversidade e complexidade das mutações relacionadas à resistência no HIV-1 às drogas utilizadas na HAART dificultam o tratamento e o controle da infecção pelo HIV-1 e da evolução à AIDS. A resistência aos fármacos é considerada o maior obstáculo ao sucesso da terapia contra a infecção pelo HIV-1, representando a grande responsável pelas falhas terapêuticas que impedem a supressão completa do vírus. É importante considerar as possíveis posições de mutações nos diferentes subtipos virais para auxiliar a escolha das drogas utilizadas, principalmente após a ocorrência de falha terapêutica.

As diferenças genômicas entre os distintos subtipos virais podem ocasionar a codificação de aminoácidos diferentes nas proteínas, inclusive nas enzimas alvo dos medicamentos, como a protease (SANCHES *et al*, 2007). Estes variantes precisam ser considerados na escolha das drogas durante a terapia antirretroviral, uma vez que este tratamento tem sido desenvolvido baseando-se na sequência do subtipo B e os efeitos decorrentes da combinação das mutações de resistência a droga e os polimorfismos naturais de subtipos não-B ainda vêm sendo elucidados (SANCHES *et al*, 2007).

Os testes para determinação de resistência podem ser de dois tipos:

a) fenotípicos, que mensuram diretamente a replicação viral, porém exigem muito tempo e os custos são elevados;

b) genotípicos, que são baseados no sequenciamento das enzimas alvo de ação dos fármacos (protease e transcriptase reversa), sendo mais rápidos e menos custosos, entretanto, apresentam difícil interpretação dos dados (BEERENWINKEL *et al*, 2002).

Um sistema padronizado para interpretação genotípica vem sendo utilizado, no qual as sequências são comparadas com a sequência consenso do subtipo B. Geralmente as mutações são indicadas por uma letra (referente ao aminoácido) seguida por um número (posição do resíduo do aminoácido) e outra letra referente à substituição (por exemplo, L10F). Estas posições são comparadas à sequência

original do vírus (selvagem) para determinar as modificações. São muitos os *sites* disponíveis na *Web* para tal fim, destacando-se a Sociedade Internacional de AIDS (*International AIDS Societ*y, IAS), Los Alamos *database*, Stanford HIV *Drug resistence* e HIV *Resistence Web* (KANTOR e KATZENSTEIN, 2004). Uma grande limitação destes sistemas que atinge diretamente os países periféricos, incluindo o Brasil, é a restrição ao subtipo B, visto que outros subtipos, como o C, também são circulantes nestas regiões.

Mesmo na ausência de tratamento, as sequências da Transcriptase Reversa (*Reverse Transcriptase*, RT) e da protease dos subtipos B e não–B são polimórficas em aproximadamente 40% dos primeiros 240 aminoácidos da RT e em 30% dos 99 aminoácidos da protease (KANTOR e KATZENSTEIN, 2004). Algumas destas substituições são associadas à resistência no subtipo B, como as posições 10, 20, 36, 63, 71 e 77 da protease e posições 69, 75, 98, 106, 118 e 179 da RT (KANTOR e KATZENSTEIN, 2004). Sendo assim, tornam-se relevantes estudos de resistência considerando subtipos não-B para melhor compreensão dos mecanismos e substituições nas sequências envolvidas nestes casos.

A determinação de um modelo de predição do fenótipo de resistência do HIV-1 às drogas, a partir do genótipo, vem sendo pesquisado por diferentes técnicas, destacando-se a aplicação de métodos estatísticos e algoritmos de aprendizagem de máquina (*Machine Learning,* ML) (BONET *et al*, 2007), programas computacionais que melhoram o desempenho automaticamente por meio da experiência.

Dentre as técnicas de ML citadas na literatura, destaca-se a utilização de redes neurais artificiais (A*rtificial Neural Networks,* ANNs) (WANG e LARDER, 2003; ROGNVALDSSON, 2004; BONET, 2007), a análise de correspondências múltiplas (BUNGNON *et al*, 2000 *apud* BEERENWINKEL *et al*, 2002), a utilização máquinas de vetor de suporte (S*uport Vector Maquine*, SVMs) (BONET *et al*, 2007), aprendizagem não-supervisionada pela análise de agrupamento *(cluster)* e de discriminante linear

(SEVIN *et al*, 2000), aplicação de algoritmos com regras: HIV*database*, VIG, ANRS, Rega e análises lineares (WANG e LARDER, 2003).

O estudo de modelos preditores de resistência geralmente considera a influência de algumas drogas isoladamente. BEERENWINKEL *et al* (2002) realizou a construção de árvore de decisão para a determinação de resistência ou susceptibilidade do vírus a 14 drogas antirretrovirais. WANG e LARDER (2003) utilizaram dois modelos de redes neurais para a previsão de resistência ao inibidor de protease (*Protease Inhibitor,* PI) Lopinavir. Um modelo considerou 11 posições de resistência descritas na literatura e o outro utilizou 28 posições de um banco de dados, comparando os resultados, o modelo de 28 posições mostrou-se mais explicativo na previsão de resistência ao Lopinavir. JAMES (2004) utilizou árvore de decisão e um modelo baseado em distâncias utilizando a técnica de K-vizinhos mais próximos (*K-nearest neighbor technique*, KNN) para predizer a resistência a alguns PIs.

Muitos dos algoritmos que vêm sendo utilizados são baseados em treinamentos supervisionados que classificam a sequência do HIV-1 quanto à susceptibilidade à droga em dois grupos: susceptível ou resistente (SRISAWAT e KIJSIRIKUL, 2004). A utilização de um método de aprendizagem não supervisionado pode ser interessante para determinar outras possíveis características como, por exemplo, o subtipo e as drogas utilizadas que aproximam grupos naturais baseando-se no genoma do vírus. Segundo DRAGHICI e POTTER (2003), a utilização de um mapa auto-organizável (S*elf Organizing Map*, SOM) permitiu a determinação de três grupos: alta, média e baixa resistência.

1.1. OBJETIVOS

Objetivo Geral

Separar indivíduos infectados pelo HIV-1 em grupos naturais baseando-se na sequência da enzima protease do vírus, verificando as posições determinantes na discriminação dos aglomerados a fim de relacioná-las a mutações de resistência ou variações nos subtipos não-B.

Objetivos específicos

Avaliar o emprego da técnica do mapa auto-organizável (SOM) na prévia organização dos dados em vetores protótipos e posterior utilização da técnica de k - médias na separação de sequências da protease do HIV-1 de indivíduos infectados em grupos naturais.

Comparar dois métodos de avaliação de *cluster*: o índice de Davies-Boudin (DB) e o índice da Silhueta na determinação do número de grupos mais apropriado para a separação dos dados.

Determinar as posições da protease determinantes na discriminação dos grupos, considerando os diferentes padrões de resistência dos subtipos virais, além de possíveis polimorfismos dos subtipos não-B.

Capítulo 2

FUNDAMENTOS TEÓRICOS

2.1. BIOLOGIA DO HIV

2.1.1. CLASSIFICAÇÃO DO HIV

A descoberta clínica do HIV ocorreu no início da década de 80 no Centro Norte-Americano de controle e prevenção de doenças (*Center of Desease Control*, CDC). A observação do aumento de infecções oportunistas, principalmente em homens homossexuais, como a pneumonia por *Pneumocystis Carinii* associada a um conjunto padrão de sinais clínicos levou à classificação deste quadro como uma imunodeficiência característica da AIDS (RAMBAUT *et al,* 2004).

O primeiro isolamento do vírus ocorreu em 1983 (GALLO *et al*, 1983, BARRÉ-SINOUSSI *et al,* 1983). Posteriormente, outros pesquisadores estudaram o HIV e o vírus foi classificado como membro do gênero *Lentivirinae*, da família *Retroviridae,* sendo esta a classificação atual (LEVY *et al,* 1984).

A análise filogenética do HIV em humanos divide o grupo em dois tipos: HIV-1 e HIV-2, com diferentes organizações do genoma. As análises filogenéticas revelam que os dois tipos se aproximam de diferentes lentivírus que infectam primatas, denominados como Vírus da Imunodeficiência Símia (*Simian Immunodeficiency Virus*, SIV). Analisando a distribuição geográfica juntamente aos fatores filogenéticos, estudos revelaram que o HIV-1 se originou a partir do SIVcpz, que infecta chimpanzés

(*Pan troglodytes*) (GAO *et al*., 1999), enquanto o HIV-2 está mais relacionado com o SIVsm, que infecta *Cercocebus atys* (HIRSCH *et al*., 1989).

A infecção pelo HIV-1 é predominante no mundo. A análise de diferentes linhagens do HIV-1 revela a grande diversidade genética do vírus, sendo divididos em grupos, subtipos e formas circulantes recombinantes (C*irculating Recombinant Forms*, CRFs), baseando-se nas diferenças filogenéticas (KANTOR e KATZENSTEIN, 2004; SANCHES *et al*, 2007). Algumas classificações remetem para a existência de três grandes grupos de HIV-1: M (M*ajor*), O (O*utlier*), N (*New*, N*on*-M e N*on*-O). A maioria das linhagens identificadas no mundo pertence ao grupo M. Já foram identificados nove subtipos virais do grupo M (A – D, F – H, J e K) e 14 formas CRFs (SANCHES *et al*, 2007).

O subtipo C do HIV-1 é o subtipo de maior ocorrência no mundo todo, e é o principal subtipo encontrado em alguns países do Oriente Médio, da Ásia e do Sul e Leste Africano. Estas são regiões muito populosas e com altas taxas de infecção pelo HIV-1. Nos países centrais, Europa e Estados Unidos, o subtipo circulante predominante é o subtipo B.

No Brasil, os subtipos virais mais freqüentes são o B e F, seguido do subtipo C e CRFs, mas já foram relatados casos isolados do subtipo D (MORGADO *et al.*, 1998, SANCHES *et al*, 2007). Existem diferenças regionais na prevalência dos subtipos virais no território nacional, sendo o subtipo B relatado em diversos estudos como o predominante no país, porém com a co-circulação de pelo menos mais um subtipo não-B por região geográfica (POTTS *et al.*, 1993; MORGADO *et al.*, 1994). Diversos estudos revelam um aumento da prevalência do subtipo C principalmente na região Sul do país (BRINDEIRO *et al*., 2003).

2.1.2. PATOGENIA

Durante a fase aguda da infecção HIV-1, ocorre um controle inicial alcançado pelo sistema imunológico do hospedeiro, que pode ser observado pela queda da viremia e pelo desenvolvimento de linfócitos $TCD8^+$ com atividade citotóxica (CTLs), antes mesmo do aparecimento de anticorpos neutralizantes, sugerindo a participação de uma imunidade celular no controle inicial do vírus (KOUP *et al*., 1994). A esse quadro associa-se a participação de citocinas pró-inflamatórias para as quais um aumento é observado nesta fase inicial da infecção (COHEN *et al*., 1997, HOGAN *et al*., 2001).

Entretanto, a deterioração do sistema imune em conseqüência da infecção viral, tanto quantitativa quanto funcional parece inevitável, culminando no desenvolvimento da AIDS. O advento da HAART trouxe uma boa perspectiva de controle da replicação viral e a consequente recuperação imunológica com a reposição de células $CD4^+$ nos indivíduos infectados, embora em longo prazo a evolução do HIV no hospedeiro leve ao mecanismo de resistência e conseqüente replicação viral.

A capacidade de evolução do vírus, com mecanismos de escape como reservatório da infecção com vírus em latência em *pool* de células $CD4^+$ (RAMBAUT *et al*,.2004; PANTALEO e FAUCI, 1996) e o mecanismo de resistência do vírus às drogas dificultam o controle da infecção. No entanto, as pessoas que tem acesso a HAART prolongam o tempo além de obterem melhorias na qualidade de vida.

2.1.3. ESTRUTURA E CICLO DE REPLICAÇÃO DO HIV-1

O HIV é um lentivírus da família *Retroviridae*. Os retrovírus são assim denominados devido à presença da enzima transcriptase reversa (RT) em seu ciclo de

replicação, que catalisa a síntese de ácido desoxirribonucéico (D*eoxiribonucleic Acid*, DNA) a partir de ácido ribonucléico (*Ribonucleic Acid*, RNA).

O vírion maduro do HIV é constituído de um envelope externo, matriz, capsídeo e nucleocapsídeo. O genoma do HIV é composto de duas fitas simples de RNA. Durante a replicação, o RNA é transcrito a DNA catalisada pela RT.

Os principais genes encontrados no genoma do HIV são: *gag, pol e env*, que codificam as principais proteínas virais. O *gag* codifica proteínas estruturais (no capsídeo, matriz e nucleocapsídeo), como a proteína p24; o *pol* é responsável pela codificação das enzimas como a RT, protease e integrase e o *env* codifica proteínas do envelope do vírus, como as proteínas gp120 e gp41. Algumas proteínas acessórias importantes estão localizadas nos genes *tat, ver, vip, vpr, vpu e nef* (SIMON e HO, 2003).

O ciclo de vida do HIV inicia com a ligação da partícula viral aos receptores na membrana da célula hospedeira. A fusão do vírion com a célula-alvo é mediada por proteínas expostas na superfície do vírion e receptores CD4 ou co-receptores (CXCR4 ou CCR5) encontrados na superfície de linfócitos $TCD4^{+}$, células dendríticas e macrófagos. Após ligar-se à célula, o vírion libera o capsídeo, contendo o material genético e as enzimas virais no interior da célula hospedeira para iniciar o processo de retrotranscrição, sintetizando DNA complementar (DNAc) a partir do RNA viral, com o auxílio da RT. A fita dupla de DNAc é importada para o núcleo da célula e a integrase catalisa a incorporação do DNAc viral ao genoma celular. Finalmente, a maquinaria da célula pode ser utilizada para a síntese de proteínas virais. A formação de novos vírions ocorre pelo empacotamento das proteínas virais e do RNA viral, que são liberados por brotamento levando parte da membrana da célula hospedeira na composição de seu envelope (Figura 2.1).

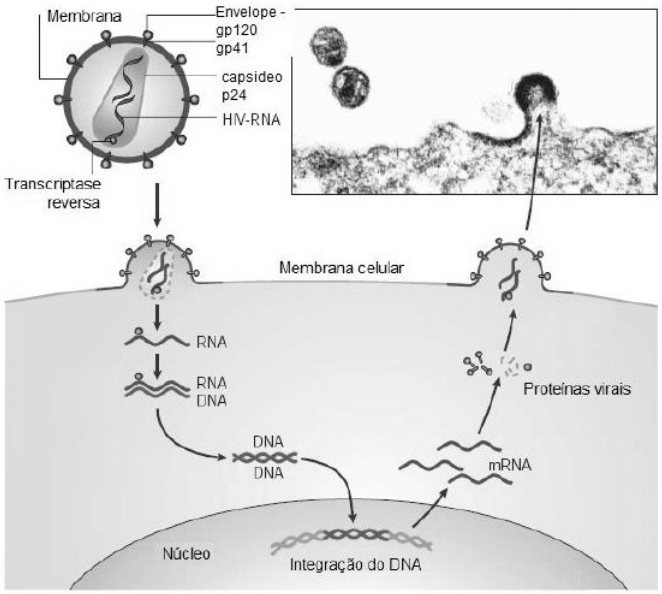

Figura 2.1: Representação esquemática do ciclo do HIV na célula hospedeira. Adaptada de (SIMON e HO, 2003).

2.1.4. TRATAMENTO E RESISTÊNCIA ÀS DROGAS

O primeiro medicamento antirretroviral utilizado para o tratamento da infecção pelo HIV-1 em adultos foi a Zidovudina (AZT), um análogo de NRTI, o qual foi capaz de promover uma diminuição da mortalidade e da frequência de infecções oportunistas em pacientes com AIDS, prolongando a sobrevida dos indivíduos, embora por apenas mais alguns meses. No entanto, a eficácia da monoterapia prolongada com AZT foi questionada devido à rápida emergência de isolados virais que apresentavam resistência ao antirretroviral, dada à presença de uma ou mais mutações não silenciosas no gene da transcriptase reversa. Estas mutações não silenciosas são alterações no DNA que resultam em mudanças na sequência de aminoácidos da proteína formada.

Os resultados obtidos com a monoterapia baseada em outros NRTIs disponíveis demonstraram-se inferiores aos alcançados com a terapia dupla, em

relação a um aumento nas contagens absolutas e percentuais de células T $CD4^+$, na redução da replicação viral, das infecções oportunistas e da progressão da doença, além de uma eficácia reduzida quanto à manutenção destas alterações e a atenuação da emergência de resistência viral (RAMBAUT *et al*., 2004). Além dos NRTIs, outro grupo de inibidores da RT tem sido extensivamente estudado: os inibidores análogos não nucleosídeos da transcriptase reversa (*Non Nucleoside Reverse Transcriptase,* NNRTIs).

A inclusão dos inibidores da enzima protease à terapia antirretroviral, na metade da década de 90, deu início a uma nova fase no tratamento de indivíduos infectados pelo HIV-1. Com o uso dos PI em combinação às outras drogas antirretrovirais, uma supressão viral muito mais profunda e sustentável e aumentos extremamente significativos nas contagens de células T CD4+ foram obtidos, em relação aos esquemas até então utilizados. A primeira combinação de antirretrovirais amplamente utilizada no tratamento de indivíduos infectados pelo HIV-1 incluía um inibidor de protease e dois NRTIs. Aos pacientes que já apresentaram falhas terapêuticas em esquemas de terapia dupla, ou potente, é recomendado um esquema mega potente – MEGAHAART (dois NRTIs, um análogo NNRTs e um PI ou dois NRTIs e dois PIs afora o Ritonavir).

Uma ampla variedade de combinações de drogas antirretrovirais encontra-se em uso atualmente. Estas várias combinações de medicamentos potentes tornou-se o padrão usual de tratamento para os indivíduos infectados pelo HIV-1. Atualmente os compostos disponíveis anti-HIV atuam na inibição no sítio de ligação das seguintes enzimas: Transcriptase Reversa (RT), Protease, proteínas de fusão e integrase.

Os inibidores de RT nucleosídicos (NRTIs) compreendem sete drogas: Zidovudina, Stavudina, Emtricitabine, Lamivudina, Didanosina, Abacavir e Tenofovir; três inibidores de RT não-nucleosídicos (NNTRIs): Nevirapina, Efavirenz e Etravirine; oito combinações de PIs: Saquinavir/ritonavir, Indinavir/ritonavir, Darunavir/ritonavir, Nelfinavir; um inibidor de fusão e outro de acoplamento: Enfuvirtide e Maraviroc.

O resultado da terapia deve ser avaliado primariamente através da carga viral do paciente. Os pacientes são denominados em falha terapêutica quando não apresentam a diminuição esperada da carga viral e as causas podem estar associadas a alguns fatores: pouca aderência; concentrações sub-ótimas da droga (por má absorção ou extrusão celular); potência inadequada da droga; ou resistência viral.

A protease é uma enzima aspártica constituída de duas (2) cadeias de 99 aminoácidos e apresenta um papel fundamental na replicação viral, sendo necessária para o processamento das poliproteínas Gag-Pol5. Os PIs se ligam à enzima com mais afinidade do que o substrato natural, promovendo a inibição por competição, inibindo a via de atuação da enzima. Os PIs utilizados na HAART retardam a evolução da doença e o possível avanço para AIDS, no entanto, a seleção de linhagens resistentes à droga impede a total supressão da replicação viral, dificultando o controle. Mais de 87 mutações já foram descritas em 49 das 99 posições de aminoácidos da protease do HIV-1 (SANCHES, 2007).

Estas mutações que conferem resistência aos PIs podem ser classificadas como primárias ou secundárias. As mutações primárias são as primeiras a serem selecionadas na presença da droga e conferem alterações que prejudicam a ação do antirretroviral, geralmente no sítio de ação da enzima. Já as secundárias, emergem tardiamente, em relação às primárias e não apresentam, por si só, um efeito significativo no fenótipo e podem, quando em conjunto com mutações primárias, melhorar a replicação viral, diminuindo o efeito da droga.

Esta seleção de mutações em linhagens resistentes impede o sucesso pleno da terapia, uma vez que linhagens do vírus resistentes a drogas distintas podem gerar vírus recombinantes resistentes a várias drogas, dificultando o sucesso terapêutico da associação de drogas (RAMBAUT *et al*, 2004). Além disso, a compreensão dos mecanismos de resistência é restrita às áreas onde o subtipo B é predominante e a terapia HAART é acessível (KANTOR e KATZENSTEIN, 2004), limitando a

compreensão dos efeitos decorrentes da combinação das mutações de resistência a droga e os polimorfismos naturais de subtipos não-B. (SANCHES *et al,* 2007).

O Brasil é um dos países que apresenta um perfil variado de distribuição dos subtipos do HIV-1. Apesar do subtipo B representar 70% - 80% das infecções no país (WALERIA - ALEIXO *et al*, 2008), a epidemia brasileira pode ser dividida em função das regiões geográficas de ocorrência. Na populosa região Sudeste, o subtipo B e CRFs do tipo BF são predominantes (GONZALEZ *et al*, 2006), a região Sul é caracterizada pelos subtipos B e C e seus recombinantes (BRINDEIRO *et al*, 2003) e nas demais áreas do país o subtipo B predomina.

Existem diversos trabalhos que reforçam a prevalência do subtipo C na região Sul do Brasil (BRINDEIRO *et al*, 2003). Em outros países da América do Sul, os relatos de ocorrência deste subtipo têm aumentado (CASTRO *et al*, 2005), apesar do subtipo F e os CRFs BF representarem os principais subtipos não-B circulantes no Brasil (BRINDEIRO *et al* 2003, COUTO-FERNANDEZ, J. C *et al*, 2005) e na América do Sul (CARRION *et al*, 2004).

Esta heterogeneidade na distribuição mundial dos subtipos do HIV-1 torna de grande valia a elucidação das diferenças existentes entre os subtipos a nível molecular e o comportamento destes em resposta ao tratamento atualmente utilizado. O entendimento da disseminação destes subtipos pode facilitar a atuação de políticas de saúde pública, auxílio clínico durante o tratamento e na síntese de medicamentos e pesquisas de vacinas.

O uso de métodos de aprendizado de máquina e estatísticos têm se mostrado uma ferramenta útil na identificação de possíveis mutações de resistência, predição e classificação das sequências a partir destas mutações no genoma do HIV-1 em pacientes em falha terapêutica.

2.2. ANÁLISE COMPUTACIONAL

2.2.1. SELEÇÃO DE ATRIBUTOS

A análise exploratória dos dados é a primeira etapa na identificação das variáveis (atributos) representativas de um modelo. Geralmente, são utilizadas técnicas que permitam reduzir a dimensionalidade dos atributos, facilitando o entendimento dos padrões de entrada e evitando o excesso de informações. A diminuição do número de variáveis para reduzir a dimensionalidade dos dados pode ser realizada por duas técnicas: extração de atributos e seleção de atributos.

Quanto à importância dos atributos para determinado problema, geralmente a classificação considera os atributos fortemente relevantes, os redundantes e os irrelevantes. Estes últimos podem ser descartados, uma vez que tem pouca influência na determinação do modelo; os atributos relevantes são fundamentais para a solução ótima e as variáveis redundantes carregam informações consideradas por outras variáveis já consideradas no modelo.

A extração consiste na obtenção de um novo conjunto de atributos a partir da combinação dos anteriores. Geralmente, a redução da dimensionalidade é alcançada, no entanto, os atributos gerados não mantêm uma correlação explícita com o problema inicial, sendo necessária a interpretação dos novos atributos no contexto do problema. A técnica mais utilizada com este propósito é a análise de componentes principais (*Principal Component Analysis*, PCA) (SMITH, 2002), na qual são realizadas combinações lineares do conjunto de variáveis tentando reter o máximo de informações possível. (SMITH, 2002).

A seleção de atributos é o processo de determinação de um subconjunto de variáveis a partir do conjunto inicial, selecionando os atributos que mais influenciam na obtenção do modelo. A escolha segue alguns critérios, removendo aqueles

redundantes ou irrelevantes. Como as variáveis pertencem ao conjunto de atributos original, existe uma relação direta com o problema, facilitando a interpretação.

Estas técnicas visam a reduzir o espaço de busca pela solução do problema, sem afetar, entretanto, a qualidade da solução final. Os métodos de busca mais comuns são divididos em dois tipos: *forward selection*, que iniciam a busca com um subconjunto vazio de atributos, considerando progressivamente os atributos segundo algum critério de avaliação a cada iteração e o *backward elimination,* que inicia o processo com um modelo completo considerando todos os atributos e os menos interessantes serão eliminados passo a passo.

A escolha dos métodos de seleção de variáveis depende do tipo de problema considerado. Em problemas com rótulos conhecidos, são utilizados métodos supervisionados; caso contrário, os métodos não supervisionados são apropriados. MITRA *et al* (2002) divide as técnicas não supervisionados: aquelas que buscam maximizar o desempenho em problemas de agrupamento e as que selecionam os atributos considerando a relevância e a interdependência dos mesmos.

A redução do número de variáveis é considerada uma etapa importante em ML, onde a aquisição automática do conhecimento é realizada baseando-se nas experiências passadas. A seleção de atributos antes do aprendizado pode ser benéfica uma vez que a redução da dimensionalidade dos dados diminui o espaço de hipóteses, permitindo a operação mais rápida e eficiente do algoritmo.

Existem vários ML que vêm sendo amplamente utilizados como técnicas de classificação e agrupamento para facilitar o entendimento e a análise de dados de diversas áreas. Com a crescente quantidade de informação genômica disponível decorrente de sequenciamentos automatizados, tais técnicas vêm sendo aplicadas à biologia.

No presente estudo, a seleção de atributos foi realizada observando-se a variação destes dentre os exemplos, excluindo do modelo as variáveis consideradas irrelevantes. Além da eliminação destes atributos, pode-se observar, ainda, o

desempenho dos agrupamentos obtidos na seleção de variáveis, pela avaliação dos grupos formados (MITRA *et al*, 2002). Uma vez que os rótulos não foram considerados na obtenção dos grupos, os métodos utilizados são não supervisionados, tais como a técnica do mapa auto-organizável de Kohonen (SOM), um tipo de rede neural artificial (ANN), que foi realizada posteriormente.

2.2.2. REDES NEURAIS ARTIFICIAIS

A fundamentação original desta metodologia é baseada na tentativa de modelar a rede de neurônios humanos para a compreensão do funcionamento do cérebro. Portanto, sua motivação inicial foi a de realizar tarefas complexas que o cérebro é capaz de executar através da simulação de seu funcionamento. McCulloch e Pitts (1943) propuseram um modelo inicial de neurônio (unidade básica da ANN) como uma unidade de processamento binária e mostraram que estas unidades eram capazes de executar várias operações lógicas (ZANINI, 2004).

Esta analogia com a neurofisiologia foi apenas uma motivação inicial do método. O conceito de redes neurais empregado atualmente é definido como um processador capaz de extrair conhecimento experimental e disponibilizá-lo para utilização na prática em modelos de previsão e métodos de classificação.

As ANNs extraem o conhecimento através do aprendizado e treinamento, a informação é armazenada em pesos e cada neurônio possui uma função de ativação (geralmente não-linear) que utiliza os padrões de entrada ponderados para definir a saída. Logo, o processo de aprendizado de uma rede ocorre utilizando um algoritmo que deve ser capaz de ajustar iterativamente os pesos para alcançar o objetivo proposto.

A etapa de aprendizado da rede depende do algoritmo de aprendizagem do processo iterativo de alteração dos pesos de conexão, na qual os neurônios de uma

ANN estão estruturados (interconectados). Existem diferentes algoritmos de aprendizagem para ANNs, destacando-se dois grandes grupos: (1) Métodos de aprendizagem supervisionados e (2) Métodos de aprendizagem não-supervisionados. No primeiro, a rede neural recebe um conjunto de entradas e seus correspondentes padrões de saída, onde ocorrem ajustes nos pesos sinápticos até que o erro entre os padrões de saída gerados pela rede tenha um valor desejado, no outro, a rede neural trabalha os dados de forma a determinar algumas propriedades similares entre os padrões de entrada. A partir destas propriedades é que o aprendizado é constituído.

Este trabalho utiliza um método não supervisionado de rede neural conhecido como mapa auto-organizável de Kohonen que será descrito a seguir.

MAPAS AUTO-ORGANIZÁVEIS DE KOHONEN (SOM)

O SOM é um modelo de rede neural artificial baseado em um aprendizado competitivo ou auto-organizado e a sua arquitetura é composta somente por duas camadas. As redes auto-organizáveis de Kohonen são caracterizadas pela formação de um mapa topográfico dos padrões de entrada (KOHONEN *et al*, 1996) que revelam suas características.

A arquitetura desta rede compreende somente uma camada de entrada e uma de saída que são conectadas por pesos. Possuem um aprendizado competitivo, no qual os neurônios de saída competem entre si para serem atualizados (ZANINI, 2004). A localização espacial dos neurônios auto-organizados indica características dos padrões de entrada, agrupando dados semelhantes em neurônios. Seu principal objetivo é a redução da dimensão dos dados, onde um conjunto de dados de entrada é organizado em um mapa unidimensional (1D) ou bidimensional (2D) de saída (KOHONEN *et al*, 1996), ou seja, os exemplos semelhantes são agrupados em neurônios reduzindo a dimensão do espaço.

O aprendizado competitivo é baseado na distância entre as unidades do neurônio (saída) e a entrada. A unidade (neurônio) que apresentar o vetor peso mais próximo dos padrões de entrada sofre atualização, que também ocorre em seus vizinhos de forma proporcional à distância calculada (ZANINI, 2004). A determinação da vizinhança é baseada em uma função de vizinhança, que limita o raio de atualização.

Desta forma, o SOM agrupa *i* elementos de um banco de dados de entrada em *J* neurônios, ou seja, projeta-se o espaço de entrada em um espaço de menor dimensão (Figura 2.2). A conexão entre as camadas de entrada e de saída do algoritmo é feita através dos pesos (W_{ij}), ou seja, a cada neurônio no espaço de saída corresponde um vetor-peso com as características do espaço original.

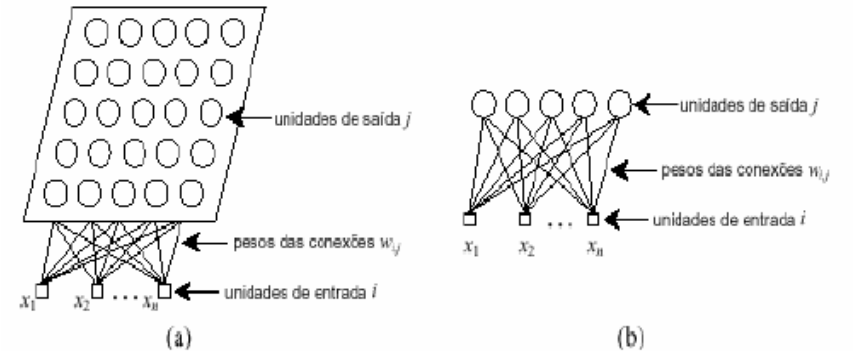

Figura 2.2: Arquiteturas de um SOM. (a) bidimensional. (b) unidimensional
Extraída de (ZANINI, 2004).

O SOM preserva a topologia dos dados de entrada manifestando a organização dos dados em agrupamentos. Desta forma, o SOM pode ser utilizado associado a técnicas de *clusters* a fim de reduzir o espaço dimensional pela organização dos dados em protótipos para posterior agrupamento destes (VESANTO e ALHONIEMI, 2000).

2.2.3. TÉCNICAS DE AGLOMERADOS

A análise de agrupamentos visa à organização dos dados em um pequeno número de grupos (*clusters*), de forma que os elementos similares estejam alocados no mesmo grupo e os padrões muito distintos estejam em grupos diferentes (PETRÓVIC, 2006; JAIN *et al*,1999).

Segundo KASK (1997), apesar de existirem diversos algoritmos para realizar esta tarefa, há dois aspectos relacionados à natureza dos métodos de agrupamento que devem ser considerados:

a) o método é não supervisionado, ou seja, os exemplos são organizados em agrupamentos considerando que os grupos não são conhecidos no início; não existe pré-classificação dos dados nem distinção entre variáveis dependentes ou independentes. O método lida diretamente com os dados, visando determinar a estrutura dos padrões.

b) a escolha da distância utilizada para determinar a similaridade entre os exemplos. A similaridade entre os padrões é essencialmente realizada por suas distâncias, onde a função de distância é a quantificação da similaridade, ou seja, quanto menor a distância, maior a similaridade.

Existem duas classes básicas de algoritmos: a) agrupamento hierárquico ou b) particional.

O agrupamento hierárquico busca reunir sucessivamente grupos menores, formando grupos maiores, ou dividir grupos grandes em outros de maior similaridade interna. Os métodos diferem pela regra adotada para decidir quais grupos devem ser reunidos ou divididos. O resultado do algoritmo é um gráfico tipo árvore chamado de "dendograma" que mostra como os grupos são interrelacionados. A figura 2.3 mostra um exemplo deste diagrama, no qual cada ramo representa um elemento e a raiz mostra o agrupamento de todos os elementos.

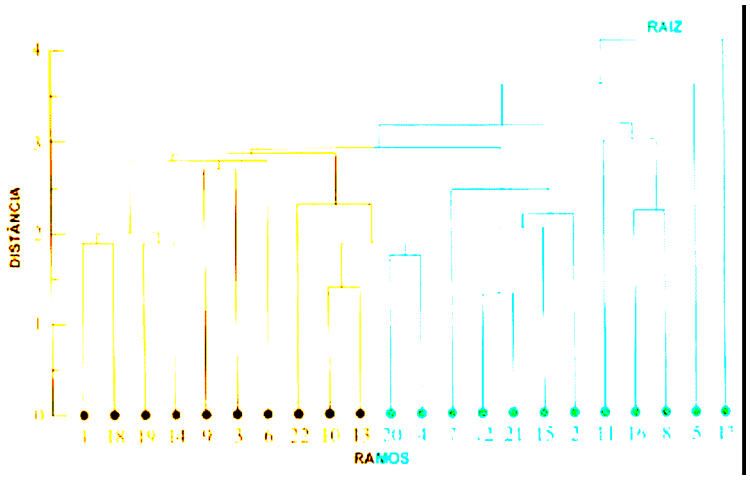

Figura 2.3: Exemplo da organização de um dendograma. Extraída de (DONI, 2004).

O agrupamento particional busca dividir o conjunto de dados em um conjunto de grupos distintos entre si, maximizando as dissimilaridades dos diferentes grupos. A ideia central é escolher uma partição inicial dos exemplos e, em seguida, alterar os grupos visando à maior similaridade entre os grupos. As técnicas que seguem o agrupamento particional, incluem o K-médias, que será descrito detalhadamente adiante e outros como o K-medóide, que utiliza o valor médio dos elementos em um grupo como um ponto referência, chamado de medóide. Esse é o elemento mais em central em um grupo e, a estratégia básica é encontrar *K* grupos em *N* exemplos e, arbitrariamente, encontrar um elemento representativo (medóide) para cada grupo.

Os métodos particionais apresentaram vantagens em aplicações com grandes conjuntos de dados, nos quais a construção de um dendograma era computacionalmente mais custosa (JAIN *et al*, 1999). No entanto, é necessário determinar o número *k* de grupos antes do processo.

Geralmente, os métodos por partição são baseados na otimização de uma função definida que envolve, por exemplo, a minimização do erro quadrático, e são de natureza combinatorial buscando formação dos grupos com elementos mais similares.

Na prática, a repetição dos algoritmos com diferentes “sementes” converge para a solução ótima (JAIN *et al*, 1999).

ALGORITMO K-MÉDIAS

Em função de ser o método de implementação mais simples e mais comum dentre as técnicas de *cluster*, K-médias, um método que utiliza a função quadrática como função de custo, é uma referência com a qual os outros algoritmos de análise de *cluster* são comparados (JAIN *et al*, 1999).

A técnica K-médias é o algoritmo de agrupamento particional proposto por MacQueen em 1967 (JAIN *et al*, 1999). Este algoritmo inicia com a escolha aleatória do centróide e formação dos grupos baseando-se na similaridade intra grupos até que um critério de convergência seja alcançado: não há mais alteração nos grupos ou o erro quadrático deixa de diminuir significativamente após as iterações. Este algoritmo requer que os dados sejam compostos de variáveis numéricas, pois uma parte do processo é baseada no cálculo das médias.

O processo de cálculo consiste basicamente das seguintes etapas:

a) selecionar *k* exemplos;

b) determinar as coordenadas destes exemplos como sendo os centróides dos grupos;

c) calcular a distância do próximo ponto aos *k* centróides (geralmente empregando a distância euclidiana);

d) incorporar o ponto ao grupo mais próximo;

e) recalcular o centróide deste grupo;

f) passar ao próximo exemplo – se terminarem, recomeçar do primeiro exemplo, revendo seu posicionamento;

g) encerrar o processo se não houver possibilidade dos dados mudarem de grupo ou retornar à etapa (c).

Embora simples e razoavelmente eficiente, o algoritmo K-médias tem algumas desvantagens. Um dos problemas apontados é a escolha das condições iniciais e do número de grupos.

O número de grupos *k* e a inicialização dos centróides (escolha dos primeiros *k* pontos) podem influenciar decisivamente nos resultados.

Esta técnica pode fornecer diferentes resultados devido à escolha aleatória do centróide inicial, mas tende a convergir para a solução ótima que minimiza a soma das distâncias dentro do grupo (JIN HWAN DO e DONG-KUG CHOI, 2007).

2.2.4 TÉCNICAS DE VALIDAÇÃO DE *CLUSTER*

A utilização de técnicas de *cluster* exige, muitas vezes, a definição do número de grupos a serem formados. Nestes casos, utilizam-se índices de validação dos agrupamentos a fim de avaliar a qualidade dos grupos, facilitando a determinação do melhor número de agrupamentos possível. Estes índices são baseados na alta similaridade intra grupos, garantindo grupos homogêneos e na dissimilaridade entre os grupos formados. A seguir, dois índices de validação são descritos mais detalhadamente: (A) o Índice de Davies – Bouldin e (B) o Índice da Silhueta.

a) Índice de Davies-Bouldin (DB)

Indica a similaridade entre os agrupamentos. O índice DB independe do número de agrupamentos e do método de partição utilizados. Uma das principais

características deste índice é sua adequabilidade para estruturas hiperesféricas, já que o mesmo usa o centróide como ponto de referência (BOLSHAKOVA e AZUAJE, 2003; PETRÓVIC, 2006). O índice é calculado utilizando o número de grupos *C* e as distâncias intra grupos e entre os grupos obtidos de acordo com a equação abaixo:

$$DB(U) = \frac{1}{c}\sum_{i=1}^{c} \max_{i \neq j} \left\{ \frac{\{\Delta(Xi) + \Delta(Xj)\}}{\delta(Xi, Xj)} \right\}$$

- c = número de agrupamentos

- Δ (X_i) = distância intra agrupamento (dispersão interna do agrupamento X_i), baseado na distância para o centróide

- δ (X*i*, X*j*) = distância entre os agrupamentos X*i* e X*j*, também baseado na distância entre os centróides.

Um bom agrupamento dos dados deve apresentar homogeneidade intra grupos e heterogeneidade entre os grupos. Logo, valores baixos indicam boa partição dos dados. A minimização deste índice parece indicar a partição natural dos dados (DAVIES e BOULDIN, 1979).

b) Índice silhueta de clusterização

O índice da Silhueta (*s(i)*) mostrado na fórmula abaixo calcula a largura da silhueta de cada objeto no grupo, podendo ser calculada a largura média nos grupos e a largura média no conjunto de dados. A maior média global da silhueta indica o melhor número de agrupamentos (BOLSHAKOVA e AZUAJE, 2003). Portanto, o número de grupos com largura máxima média global da silhueta corresponde ao número ideal de grupos.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- *a(i)* – diferença entre o objeto *i* e todos os outros objetos do mesmo *cluster*

- *b(i)* – diferença entre o objeto *i* e o centróide do *cluster* mais próximo

Valores próximos de (um) 1 indicam bom agrupamento, valores próximos de zero indicam ponto na delimitação do grupo e valores negativos indicam pontos mal agrupados, ou seja, o menor valor do índice indica que a distância média do objeto *i* a outros objetos de seu *cluster* é maior do que a distância do *i* a objetos de outros *clusters.*

2.2.5. ANÁLISE DE CONCORDÂNCIA – KAPPA

A análise de concordância revela a similaridade entre os resultados de dois métodos ou observadores distintos, como por exemplo, dois métodos de classificação. O Kappa mede o grau de concordância entre as diferentes técnicas além do que seria esperado pelo acaso.

Trata-se de uma estatística para avaliar a significância da concordância entre observações. A hipótese nula a ser testada é se o Kappa é igual a 0, o que indicaria concordância nula, ou se ele é maior do que zero, concordância maior do que o acaso (teste monocaudal: H_0: K = 0; H_1: K > 0). Um valor de Kappa com valor máximo (um) 1 indica total concordância, enquanto valores negativos sugerem que a concordância encontrada foi inferior àquela esperada por acaso.

O índice Kappa é calculado pela divisão da diferença entre a concordância esperada e a concordância

observada e a diferença entre a concordância absoluta e a concordância esperada. Como esta última diferença representa a maior concordância possível entre

a esperada e a observada, quanto maior é o índice Kappa, maior é a concordância entre as observações. O índice kappa pode ser definido matematicamente pela representação abaixo:

$$K = \frac{Ce - Co}{1 - Ce}$$

- *K* - Índice Kappa

- Co - Concordância observada

- Ce – Concordância esperada ao acaso

- Ca – Concordância absoluta (1)

Segundo LANDIS e KOCH (1977), a avaliação qualitativa da concordância sugerida aproxima-se da descrita na Tabela 2.1 abaixo.

Tabela 2.1: Interpretação do índice Kappa - Adaptada de (LANDIS e KOCH, 2003).

Valores de kappa	Interpretação
< 0	Concordância pobre
0,00 - 0,20	Ligeira concordância
0,21 - 0,40	Concordância limítrofe
0,41 - 0,60	Moderada concordância
0,61 – 0,80	Substancial concordância
0,81 – 1,00	Perfeita concordância

Capítulo 3

REVISÃO DE LITERATURA

Muitos estudos vêm sendo realizados visando a identificação ou previsão de posições determinantes na resistência do HIV-1 a partir dos testes genotípicos. A interpretação de resistência aos medicamentos a partir do genoma viral pode auxiliar na conduta clínica e escolha das drogas durante o processo terapêutico. Alguns métodos estatísticos, ML e algoritmos baseados em regras vêm sendo testados para auxiliar na predição de resistência às drogas antirretrovirais a partir do genoma do HIV-1.

Os métodos estatísticos mais utilizados consideram a regressão, entretanto, devem associar os efeitos entre as mutações para serem adequados (SAIGO *et al*, 2007), uma vez que esta associação entre as mutações é crucial na determinação da resistência. As mutações de resistência são divididas em primárias e secundárias: enquanto as primárias são diretamente responsáveis pela diminuição na sensibilidade à droga, as secundárias não são capazes de determinar resistência sozinhas, somente quando associadas a outras mutações. Logo, o método utilizado na determinação do modelo deve considerar esta associação entre as variáveis (mutações). SAIGO *et al* (2007) propuseram um modelo denominado *itemsetboosting* que é baseado em um modelo de regressão. A cada iteração uma nova combinação de mutação é encontrada e aplica-se o algoritmo de *branch-and-bound* para otimização dos resultados. Este método mostra claramente todas as associações possíveis entre as mutações e revelou alta acurácia na predição de resistência a inibidores nucleosídicos de transcriptase reversa (*Nucleoside Reverse Transcriptase Inhibitors*, NRTIs) (SAIGO *et al*, 2007).

Um algoritmo baseado em regras denominado CBA (*Classification Based Association*) foi utilizado por SRISAWAT e KIJSIRIKUL (2004) para construir um modelo preditor de resistência fenotípica a seis PIs. O desempenho da predição foi avaliado utilizando a técnica de validação cruzada 10-*fold cross validation,* apresentando uma acurácia entre 84,11% e 92,64%. Este algoritmo mostrou melhor desempenho avaliando a acurácia média quando comparados a outros baseados em regras como o HIVdb (HIV *database*), SVM e análise estatística de regressão linear múltipla (SRISAWAT e KIJSIRIKUL, 2004). Segundo os autores, há duas vantagens importantes no método de classificação utilizando o algoritmo CBA. Tratando-se de um conjunto de regras, facilita a compreensão e interpretação na previsão da resistência em um caso novo, além de permitir a utilização de atributos literais como padrões de entrada (SRISAWAT e KIJSIRIKUL, 2004). Dentre os métodos comparados, SVM e regressão linear não são flexíveis, exigindo a codificação dos atributos para dados numéricos, tais como atributos binários. A codificação dos padrões de entrada pode levar à perda de informação acarretando redução no desempenho da previsão (SRISAWAT e KIJSIRIKUL, 2004).

Além de métodos estatísticos e baseados em regras, as ML são as técnicas mais empregadas na predição do fenótipo a partir do genótipo. Muitos métodos supervisionados vêm sendo empregados com esta finalidade, como as árvores de decisão (*Decision Tree*, DT), SVMs e ANNs.

A árvore de decisão compreende um modelo prático de uma função recursiva que atribui valor a uma variável e, baseando-se neste valor é feita a escolha de outra variável ou a saída. As DT, sendo um método supervisionado, são treinadas de acordo com um conjunto de exemplos previamente classificados e, posteriormente, outros exemplos são classificados baseados nesta mesma árvore.

Antes da construção da árvore de decisão, BEERENWINKEL *et al* (2002) verificaram a significância estatística de cada posição da sequência na determinação

da resistência por meio da técnica de informação mútua. Segundo os autores, a construção da árvore baseada nesta técnica é um excelente método para auxiliar na obtenção do modelo. A técnica de árvore de decisão parece adequada, pois fornece a classificação e representa os efeitos da interação entre as diferentes mutações (BEERENWINKEL *et al,* 2002). Ainda assim, SAIGO (2007) considera as associações entre as mutações limitadas por esta técnica, sugerindo que a utilização de ANNs e SVMs que consideram todas as possíveis interações. A qualidade do modelo foi estimada pela técnica de validação cruzada *leave-one-out,* exibindo erros pequenos exceto nas drogas Zalcitabine, Didanosina e Estavudina que apresentaram erros de 25% - 32%. Além disso, algumas posições encontradas no modelo não são descritas como associadas à resistência como as posições 44 e 118 da RT que se mostraram importantes em relação à Zidovudina quando a única posição associada é a 30.

A maioria dos modelos de predição descritos na literatura leva em consideração a grande complexidade em virtude das diferentes mutações, interações entre elas e o amplo espectro de drogas envolvidas (BEERENWINKEL *et al,* 2002). WANG *et al* (2004) propuseram um modelo linear, a partir do modelo completo ("*backward*"), no qual cada mutação contribui de forma independente e quantitativamente na predição da resistência do HIV-1 aos fármacos. Apesar da facilidade de obtenção do modelo e da alta acurácia obtida, os autores já concluem que tal modelo preditor não é aplicado a casos clínicos devido às evidências experimentais que revelam interações entre as mutações.

Alguns modelos de predição obtidos a partir de dados clínicos e da história terapêutica analisados por regressão linear são comparados a modelos obtidos pela análise genotípica (ZOLOPA, 1999). Este estudo mostrou que a observação da sequência da enzima viral pode ser importante na predição de resistência a Saquinavir e Ritanavir em indivíduos com falha terapêutica em uso de PI. O modelo baseado em dados clínicos explicou 45% enquanto o modelo baseado no genótipo explicou 70% dos casos de resistência (ZOLOPA, 1999). As sequências foram comparadas às

sequências consenso do subtipo B (Los Alamos *data base*), sendo qualquer alteração considerada mutação. Este estudo, entretanto, apresenta limitações: quanto à modificação da conduta terapêutica após identificação de resistência, não garante melhor supressão da replicação viral no paciente em falha terapêutica e a complexidade quanto às associações mutacionais não foi levada em consideração (ZOLOPA, 1999).

A utilização de ANNs tem ampliado em diversas áreas por tratar problemas de alta complexidade (WANG e LARDER, 2003). O aprendizado das ANNs ocorre por um processo iterativo que ajusta os pesos das conexões na saída do modelo, ou seja, determinam a influência de cada variável na saída. Um modelo de predição de resistência ao PI Lopinavir foi construído utilizando o algoritmo de retropropagação (*back-propagation)*, no qual os pesos das conexões entre os neurônios são atualizados até a obtenção de um erro mínimo (WANG e LARDER, 2003). A validação utilizou a técnica de validação cruzada (10-*fold cross-validation).* Foram construídos dois modelos de rede utilizando dois grupos de variáveis (mutações): 28 posições na protease relacionadas potencialmente à resistência a Lopinavir e 11 posições na protease já descritas na resistência ao mesmo PI. Ambos os grupos foram treinados, validados e testados. O resultado mostrou que o grupo com 28 posições teve maior acurácia, revelando uma correlação de 88% e sugerindo novas mutações importantes na determinação de resistência ao Lopinavir (WANG e LARDER, 2003).

Outro fator que pode contribuir para a determinação de resistência é a conformação tridimensional da proteína. BONET *et al* (2007) demonstraram que a energia envolvida em um aminoácido ou entre aminoácidos pode auxiliar na representação do genótipo do HIV por se aproximar da configuração em 3D. Os autores utilizaram SVM, redes neurais Multicamadas (*Multilayers Perceptrons*, MLPs) e redes neurais recorrentes bidimensionais (*Bidirectional Recurrent Neural Networks*, BRNN) como métodos de classificação. No último, a acurácia obtida variou de 81,3 a 94,7% dependendo do PI considerado (BONET *et al,* 2007). Considerando a influência

da energia dos aminoácidos na conformação tridimensional (3D), existem duas formas de analisar a sequência. A primeira considera a sequência como um todo, encontrando aspectos gerais para descrever a seqüência completa, como no método aplicando MLP. Na segunda, trabalha-se com subunidades e faz-se a análise de cada parte verificando a influência de uma determinada região no resto da sequência (BONET *et al*, 2007).

Na rede BRNN utilizada por BONET (2007) foi aplicado o modelo mais simples, no qual a sequência foi dividida em três partes de tamanhos iguais, havendo influência entre si, onde a parte central é influenciada pelas outras duas e assim por diante. O treinamento aplicou o algoritmo *Backpropagation through time* (BPTT). A função alvo (*target*) utilizada foi entropia cruzada (*Cross-entropy*) e a função de ativação foi a Softmax. Ambas as redes utilizadas, MLP e BRNN, foram validadas aplicando validação cruzada *10-fold cross validation* e apresentaram resultados semelhantes, sugerindo que os dois tipos de rede podem ser utilizados.

O desenvolvimento de um modelo computacional híbrido Algoritmo Genético para a seleção de atributos e Classificador de Kernel de Fisher para a seleção de variáveis mostrou-se promissor na previsão de resistência, em sequências 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 (SILVA, 2009).

DEFORCHE *et al*. (2007) propuseram a aplicação de redes neurais bayesianas, considerando independência condicional entre subconjuntos de variáveis e utilizando a regra de probabilidades, visando verificar as relações entre as variáveis: tratamento, mutações de resistência e a presença de polimorfismos para os PIs Indinavir, Saquinavir e Nelfinavir. Os resultados obtidos permitiram identificar as posições de mutação de resistência 30, 88 e 90 para o Nelfinavir, 90 para Saquinavir e 82 para Indinavir.

DRAGHICI e POTTER (2003) analisaram a predição da resistência a dois PIs: Indinavir e Saquinavir, avaliando por duas perspectivas. Primeiramente, foi construído

um modelo baseado nas características estruturais do complexo protease do HIV – droga com ação inibidora. Estudos revelam que alterações estruturais em tais complexos estão associadas à resistência ao medicamento uma vez que a perda de contato entre a droga e a protease acarreta modificações no complexo e, consequente resistência à droga. Posteriormente, um classificador foi construído baseando-se na sequência genotípica com as mutações de resistência à droga. Em ambos os casos, o SOM foi utilizado para extrair características importantes e agrupar os padrões de entrada de forma não-supervisionada, seguido da identificação caracterizada pelo IC90, resultado do teste fenotípico que representa a quantidade de droga necessária para reduzir em 90% a replicação viral. Segundo DRAGHICI e POTTER (2003), as categorias consideradas foram alta, média e baixa resistência ao PI, além de outras duas classes, uma na qual não se conhecia previamente a classificação quanto à resistência e outra mais heterogênea, na qual houve mistura de sequências resistentes ao PI e não resistentes. Em trabalho anterior, POTTER e DRAGHICI (2002) realizaram um modelo de predição de resistência ao Saquinavir baseando-se SOM e experimentando diferentes dimensões de mapa, somente matrizes quadradas variando entre 3 x 3 e 12 x 12. Este modelo obteve 68% de acurácia na melhor rede individualmente e até 78% de acurácia ao combinar diferentes redes.

Em ambos os trabalhos, os autores citados utilizaram SOM na predição de resistência a determinadas drogas e avaliaram somente quanto à resistência, não levando em consideração os diferentes subtipos do vírus e suas diferentes respostas aos medicamentos. O uso do SOM no presente estudo busca a separação das sequências da protease do HIV-1 em grupos naturais, não considerando a droga do tratamento antirretroviral. Na análise, foram considerados tanto a resistência quanto o subtipo viral.

Capítulo 4

MATERIAL E MÉTODOS

O banco de dados utilizado é constituído da sequência de aminoácidos da enzima protease do gene *pol* (polimerase) do HIV-1 de 863 pacientes infectados pelo HIV-1, sendo 694 em falha terapêutica e 169 *naïve*, ou seja, que não tiveram contato prévio com qualquer tratamento antirretroviral. A subtipagem dos vírus mostra que 570 indivíduos foram infectados pelo subtipo B, 166 pelo C, 105 pelo F e apenas uma sequência pertence ao subtipo D. Os demais indivíduos, 21, estão infectados por CRFs (Tabela 4.1).

Os dados foram cedidos pelo Laboratório de Virologia do Centro de Saúde – CCS/UFRJ, Programa Nacional de DST/AIDS– MS.

Tabela 4.1: Distribuição dos dados.

Subtipo viral	Número de indivíduos em falha terapêutica	N de indivíduos não - tratados (naïve)	Total
B	471	99	570
C	117	49	166
F	86	19	105
Outros	20	2	22
Total	694	169	863

O desenvolvimento do trabalho foi dividido em sete etapas descritas detalhadamente a seguir:

1. Codificação dos aminoácidos para variáveis numéricas
2. Seleção de atributos - Eliminação de variáveis irrelevantes
3. Redução da dimensionalidade utilizando SOM: variando a dimensão do mapa para determinar o tamanho da matriz
4. Determinação do número de grupos: Testes de validação de *cluster* – Silhueta e Davies-Bouldin
5. Agrupamento pela técnica k-médias
6. Análise de concordância KAPPA entre os modelos completo e reduzidos
7. Escolha do modelo reduzido concordante
8. Análise dos grupos

As etapas 3, 4 e 5 foram realizadas com o modelo completo e repetidas com o modelo reduzido escolhido.

Todas as técnicas foram implementadas no software MATLAB R2007a (*The MathWorks* – U.S.A.).

A técnica de redes neurais escolhida para a redução de dimensionalidade está disponível na internet no SOM toolbox para o programa MATLAB. Os padrões de entrada na rede devem ser dados contínuos, logo, foi necessário converter o conjunto de dados, aminoácidos representados por letras, em dados numéricos. Para isto, foi utilizada a escala de hidrofobicidade de Eisenberg (EISENBERG *et al,* 1984). (Tabela 4.2).

Tabela 4.2 – Escala de hidrofobicidade: valores atribuídos a cada aminoácido pela escala de Eisenberg

Aminoácidos	**Símbolo**	**Escala de Eisenberg**
Alanina	A	0,62
Cisteína	C	0,29
Ác. Aspártico	D	-0,9
Ác. Glutâmico	E	-0,74
Fenilalanina	F	1,19
Glicina	G	0,48
Histidina	H	-0,4
Isoleucina	I	1,38
Lisina	K	-1,5
Leucina	L	1,06
Metionina	M	0,64
Asparagina	N	-0,78
Prolina	P	0,12
Glutamina	Q	-0,85
Arginina	R	-2,53
Serina	S	-0,36
Treonina	T	-0,05
Valina	V	1,08
Triptofano	W	0,81
Tirosina	Y	0,26

O modelo considerou somente as posições com maior variabilidade de aminoácidos, uma vez que estas são mais influentes na discriminação dos grupos. Esta variabilidade no resíduo de aminoácido depende fundamentalmente de dois fatores: ocorrência de mutações de resistência à terapia antirretroviral e polimorfismos característicos do subtipo viral.

Algumas posições possuem elevada taxa de mutação devido à terapia a qual o paciente foi submetido ou representam posições de polimorfismos que caracterizam os diferentes subtipos do HIV-1. Ambas são interessantes na determinação do modelo e na caracterização dos grupos formados, enquanto posições idênticas à sequência consenso possivelmente não são influenciadas pela ação do fármaco nem caracterizam subtipos.

Inicialmente, cada exemplo é representado pelos 99 atributos, aminoácidos da sequência da enzima protease. Visando à exclusão das posições com pequena variação, consideramos como modelo completo somente as posições com variação mínima de 1% dentre os exemplos, que é a taxa estimada de mutação no HIV-1 ao ano. Uma importante característica do HIV-1 é a 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 (quasispécie).

O mapeamento do conjunto de dados para um espaço bidimensional foi feito utilizando o SOM, onde cada neurônio *J* do mapa representa o conjunto de exemplos *N* vizinhos a *J.* O algoritmo disponível no pacote SOM toolbox para o Matlab utiliza a distância euclidiana para determinar o neurônio vencedor.

VESANTO e ALHONIEMI (2000) sugerem o cálculo do número de neurônios (unidades no mapa) de saída como $5\sqrt{n}$, na qual *n* representa o número de exemplos treinados. Uma alternativa na escolha do tamanho do mapa é considerar a qualidade do SOM baseando-se nos erros de quantização e topográfico (CÉRÉGHINO e PARK, 2009).

No presente estudo, foi determinada a dimensão da matriz da rede SOM baseando-se nos erros topográfico e de quantização, que avaliam a qualidade da rede e no número de neurônios sem ativação. O erro topográfico representa a proporção de vetores de dados nos quais os primeiro e segundo neurônios vencedores não são unidades adjacentes, enquanto o erro de quantização mostra a distância média entre cada vetor de entrada e o neurônio vencedor. Foram testadas matrizes quadradas de 3 x 3 a 20 x 20, sendo avaliados os erros topográfico e de quantização, além do percentual de neurônios ativados.

A técnica k-médias foi utilizada a fim de separar os J neurônios em C grupos (Figura 4.1). Variou-se o número de grupos (k) de 2 a 6 e a técnica foi repetida 10 vezes, após 50 replicações internas em cada. Esta técnica pode fornecer diferentes resultados devido à escolha aleatória do centróide inicial, mas tende a convergir para a solução ótima que minimiza a soma das distâncias dentro do grupo após algumas repetições (JIN HWAN DO e DONG-KUG CHOI, 2007).

Os resultados foram avaliados pelos índices de Davies-Bouldin e da Silhueta.

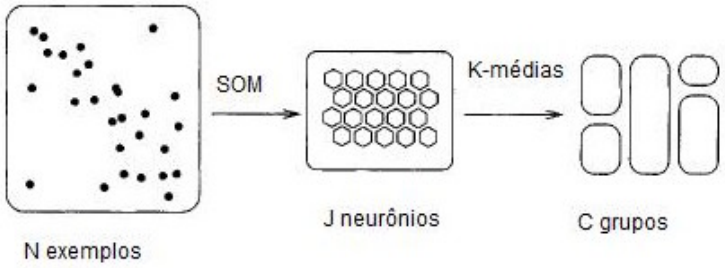

Figura 4.1: Esquema das etapas de agrupamento dos exemplos. Adaptado de (VESANTO e ALHONIEMI, 2000).

A técnica k-médias foi então repetida para o número *C (k)* de grupos indicado pelos índices.

A fim de reduzir o número de variáveis no modelo, estes *C* grupos foram obtidos repetidamente variando o número de atributos (posições). A determinação do número de atributos nestes modelos reduzidos considerou a variabilidade de aminoácidos nas posições. As freqüências de corte utilizadas partiram de 2,5% com intervalos de 2,5% (2,5%, 5%, 10%, 12,5%, 15%, 17,5%, 20%, 22,5%, 25%, 27,5%, 30%, 32,5%).

O processo foi repetido desde a obtenção do SOM até a obtenção dos C grupos pelo k-médias, utilizando sempre a mesma dimensão do mapa do SOM e o *C* determinado pelos índices. O índice Kappa foi calculado comparando os grupos obtidos com o modelo completo e os modelos reduzidos. O modelo reduzido escolhido mostrou-se concordante de acordo com o índice Kappa, apesar do menor número de atributos representativos de cada exemplo. A análise de concordância foi realizada baseando-se no índice kappa e na interpretação de LANDIS e KOCH (1977) (Tabela 2.1).

Após determinarmos o número de variáveis no modelo reduzido, a obtenção do SOM foi repetida e as técnicas de validação de *cluster* foram reconsideradas a fim de verificar se o número *C* de grupos escolhido com o modelo completo coincide com o *C* obtido com o modelo reduzido selecionado. A técnica k-médias foi repetida com o novo banco de dados resultante do SOM utilizando os *C* grupos indicado pelos índices. Finalmente, os grupos foram analisados quanto às informações de resistência e subtipos virais.

Quanto à resistência, os indivíduos podem estar em falha terapêutica devido à infecção por vírus resistentes ou não terem sido tratados (*naïve)*. Quanto ao subtipo viral, a infecção pode ser causada pelo subtipo B do vírus ou pelos HIV do grupo não-B, que inclui os subtipos C, D, F e CRFs. Como o número de sequências correspondentes ao subtipo D e os CRFs era pequeno, consideramos os dois como um tipo “outros” a fim de facilitar a interpretação dos resultados.

Em estudos de HIV/AIDS, os principais bancos de dados públicos existentes são o do Laboratório Nacional de Los Alamos e o da Universidade de Stanford, sendo este último mais voltado para o estudo das mutações associadas à resistência aos medicamentos. Desta forma, as posições da protease selecionadas no modelo foram comparadas às descrições encontradas na base Stanford.

Capítulo 5

RESULTADOS

A seleção de variáveis reduziu o banco de dados de 99 para 61 atributos (aminoácidos) representativos de cada exemplo, considerando pelo menos 1% de variabilidade. O treinamento deste conjunto de dados constituído pelas 583 sequências da protease do HIV-1 de indivíduos infectados descritas por 61 aminoácidos, convertidos em dados numéricos pela escala de Eisenberg, foi considerado como modelo completo.

A avaliação da dimensão do mapa auto-organizável a partir da matriz quadrada 3 x 3 até 20 x 20 mostrou que o aumento da dimensão do mapa leva à diminuição do erro de quantização (Figura 5.1), enquanto o erro topográfico mostrou-se variável com o aumento da dimensão da matriz (Figura 5.2). O mapa 20 x 20 do SOM apresentou menor erro de quantização, 1,056 e um erro topográfico de 0,022, entretanto, 143, de um total de 400 neurônios, não estavam ativados (35,75%). O mapa selecionado foi o de dimensão 15 X 15, com erro de quantização igual a 1,158, erro topográfico igual a 0,019 e apresentando 56 neurônios não ativados em um total de 225 neurônios (24,89%). A distribuição do percentual de neurônios inativos com a variação da dimensão da matriz revelou pequena variação, mas o aumento da dimensão mostra tendência ao aumento no percentual de neurônios inativos (Figura 5.3).

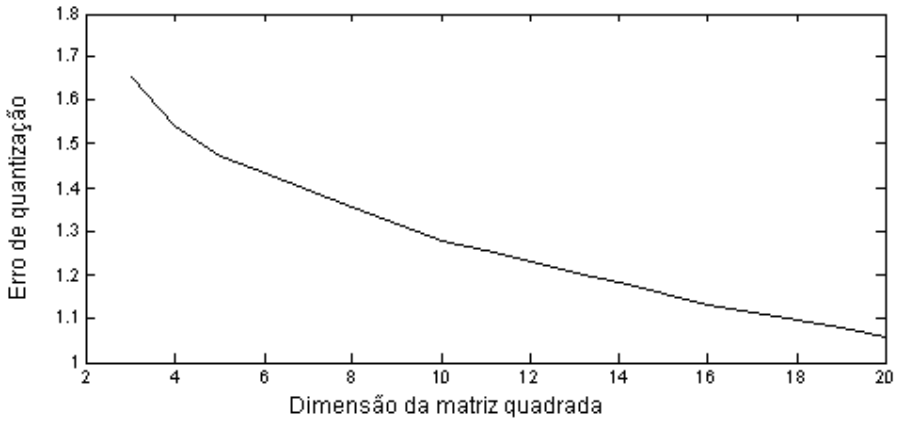

Figura 5.1: Distribuição do erro de quantização. No eixo x a dimensão do mapa variou de 3 x 3 a 20 x 20.

Figura 5.2: Distribuição do Erro Topográfico. No eixo x a dimensão do mapa varia de 3 x 3 a 20 x 20.

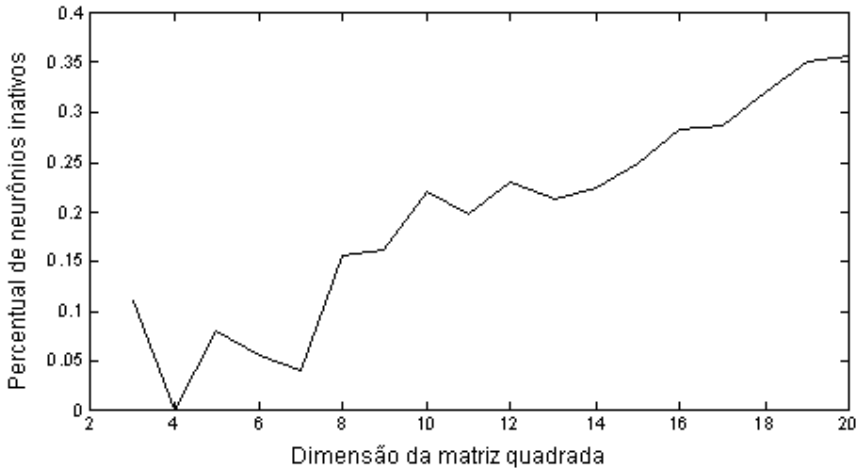

Figura 5.3: Distribuição do Percentual de neurônios inativos. No eixo x a dimensão do mapa varia de 3 x 3 a 20 x 20.

O mapa resultante do treinamento da rede mostra a existência de grupos de neurônios com ativação similar. Estes grupos podem ser visualizados na matriz U que mostra as distâncias entre unidades do mapa vizinhas (Figura 5.4).

As distâncias maiores indicam a delimitação de grupos enquanto os valores baixos revelam neurônios de ativação semelhantes. Os elementos de um mesmo grupo possuem distâncias pequenas entre si, logo, são indicados por áreas uniformes escuras, com valores baixos. A análise visual da matriz U obtida sugere quatro áreas escuras homogêneas, que indicam unidades pouco distantes entre si, porém não parecem bem delimitadas (Figura 5.4).

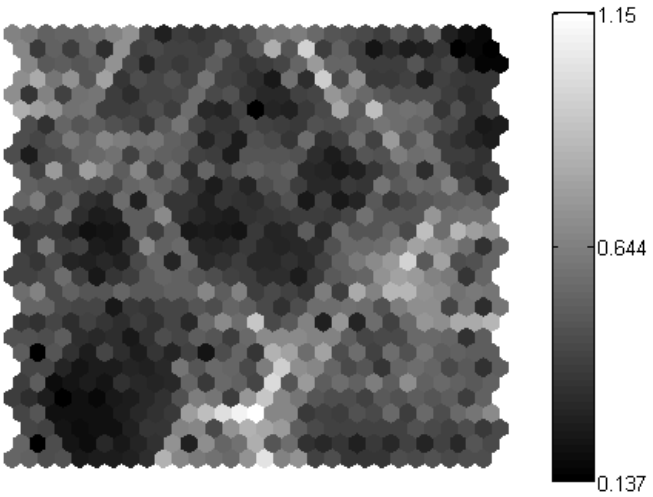

Figura 5.4 - Matriz U – Mapa 15 x 15 com 61 variáveis. A escala ao lado indica a distância entre as unidades do mapa (neurônios).

Os índices de avaliação de *cluster* foram obtidos para as dez (10) réplicas testando os diferentes números de grupos (k = 2:6). Os valores apresentados são os valores médios das réplicas de ambos os índices utilizando 61 variáveis (modelo completo) (Tabela 5.1). Observou-se que para ambos os índices a melhor partição é obtida com três grupos. Pelo índice de Davies-Bouldin, valores baixos indicam uma boa partição dos dados, enquanto que pelo índice da silhueta uma boa separação é indicada por valores altos.

O índice kappa indicou baixa concordância, baseando-se na interpretação de LANDIR e KOCH (1977), ao comparar o modelo completo com o modelo reduzido com 8 variáveis (32,5%) (Tabela 5.2). Foi considerado então o modelo com *n* = 10 variáveis (30% de variação) como o modelo reduzido concordante. A análise de concordância entre os modelos reduzidos e o completo comparou os três grupos obtidos em ambos após repetição da sequência de técnicas descrita com diferentes números de variáveis (intervalos de 2,5% de variação).

Tabela 5.1 – Índice de avaliação dos grupos. Os valores são as médias das 10 réplicas realizadas para cada índice variando o número de grupos k de 2 a 6.

K / Médias	2	3	4	5	6
DB	1,0213	**0,9737**	1,0851	1,1283	1,022
Silhueta	0,5783	**0,6006**	0,5201	0,4809	0,4904

DB – Davies-Bouldin

K – Número de grupos

As posições da protease consideradas no modelo reduzido apresentam variação em pelo menos 30% dos exemplos (Figura 5.5).

A dimensão do SOM 15 x 15 utilizando o banco de dados reduzido apresentou um erro topográfico de 0,027 e um erro de quantização de 0,397. O número de neurônios inativos foi igual a 99, o que representa 44 % dos 225 neurônios obtidos.

A análise visual da matriz U obtida no modelo reduzido sugere a existência de três a quatro grupos pela agregação de unidades do mapa, mais bem definidos do que no modelo completo. Observa-se a discriminação de três áreas escuras homogêneas delimitadas por regiões mais claras (Figura 5.6). A visualização do comportamento dos neurônios formados analisando cada variável mostra a influência de cada uma das dez posições na delimitação dos grupos (Figuras 5.7).

Tabela 5.2: Índices kappa comparando modelos com diferentes números de variáveis (*n*). O percentual representa a variação mínima de aminoácidos nos atributos considerados.

%	N	Kappa
2,5	50	0,9988
5	37	0,9977
7,5	32	0,9791
10	30	0,9965
12,5	27	0,9930
15	26	0,9930
17,5	22	0,9896
20	20	0,9722
22,5	16	0,9594
25	13	0,9699
27,5	11	0,9687
30	10	0,9815
32,5	8	**0,4287**

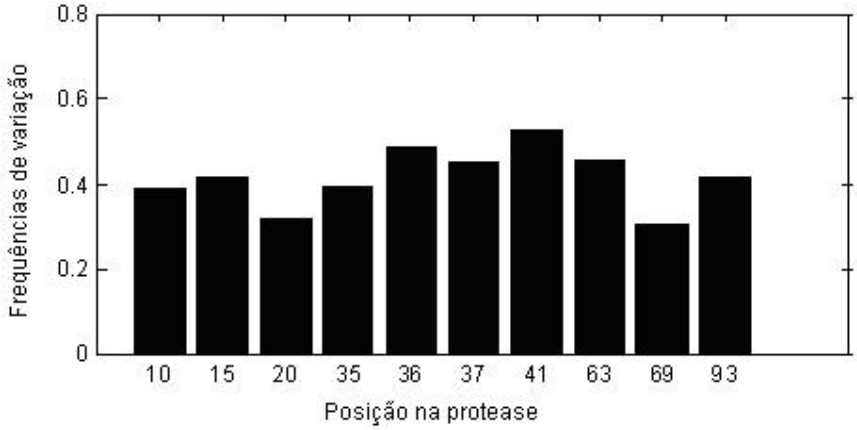

Figura 5.5: Seleção de variáveis. O eixo x mostra as posições selecionadas e o eixo y apresenta as frequências de variação considerando um mínimo de 30% (0,3) como corte na seleção de variáveis.

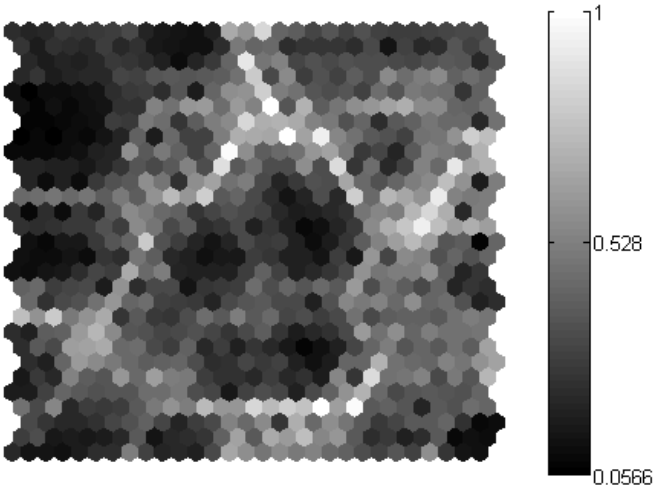

Figura 5.6: Matriz U do mapa 15 x15 com 10 variáveis.

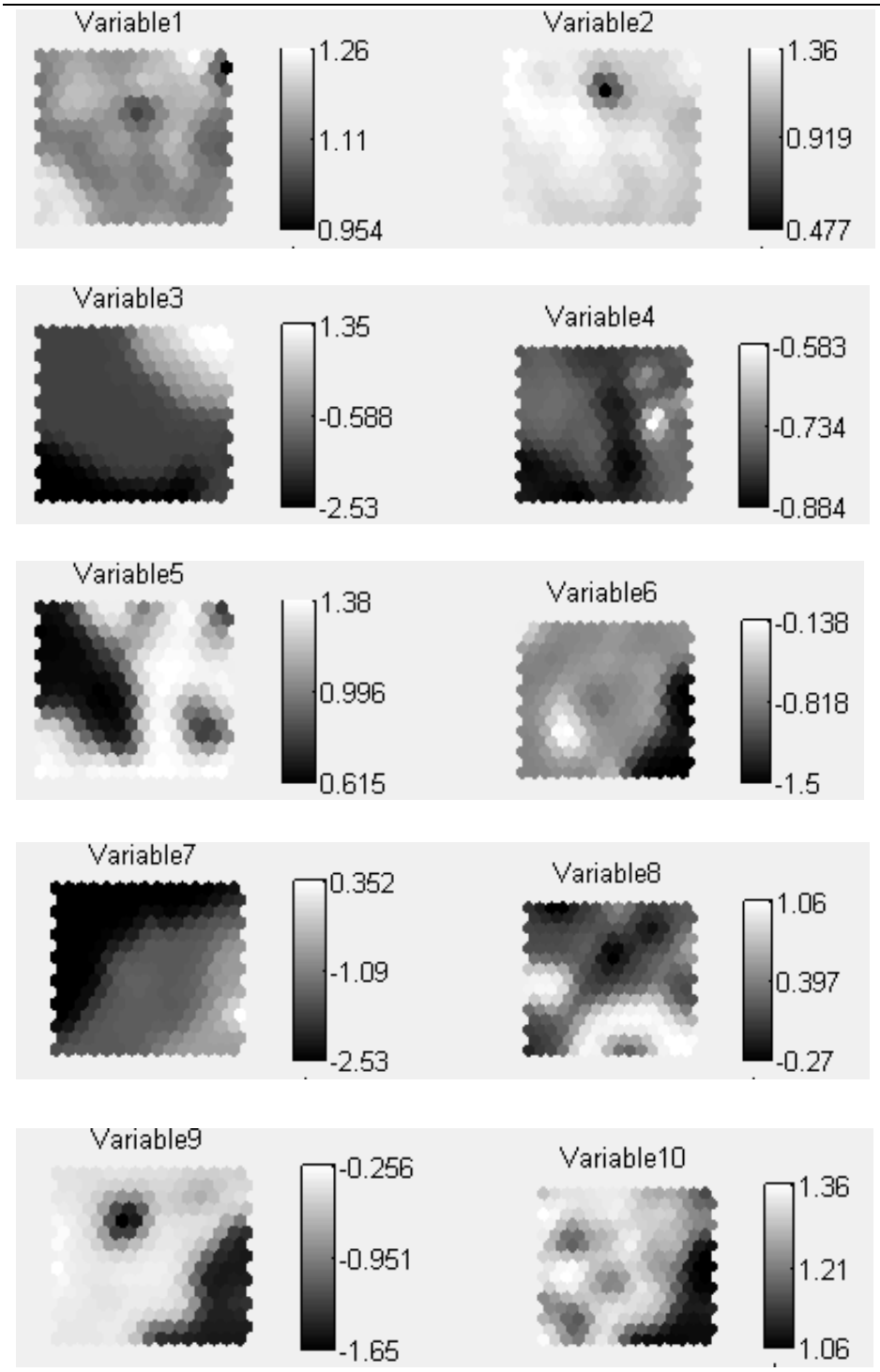

Figura 5.7: Representação da matriz de distâncias dos componentes do plano para cada variável. As variáveis 1 - 10 correspondem às posições 10, 15, 20, 35, 36, 37, 41, 63, 69 e 93, respectivamente. A escala ao lado representa a distância entre os

elementos, onde a tendência ao branco significa que os elementos são diferentes e as regiões mais escuras são mais homogêneas.

Os testes de validação de *cluster* foram recalculados para este novo conjunto de entrada. Os valores dos índices DB e Silhueta mostram-se concordantes novamente, indicando três grupos como o número de *clusters* mais apropriado (Tabela 5.4).

Tabela 5.4 – Índice de avaliação dos grupos. Os valores são as médias das 10 réplicas realizadas para cada índice variando o número de grupos k de 2 a 6.

K / Médias	2	3	4	5	6
DB	0,9319	**0,8397**	0,8995	0,9131	1,0081
Silhueta	0,6059	**0,6172**	0,5502	0,5386	0,5117

DB – Davies-Bouldin

K – Número de grupos

Os três grupos obtidos pela técnica k-médias foram analisados quanto às informações conhecidas de resistência e subtipos virais. Os indivíduos pertencentes a cada um dos grupos foram identificados a partir dos neurônios constituintes destes grupos.

A análise dos pacientes nos três grupos mostrou que o grupo 1, com 121 indivíduos, é constituído 100% por indivíduos resistentes, sendo 76,03% caracterizadas pelo subtipo B. O grupo 2, com 578 indivíduos, apresentou 79,06% de sequências do HIV resistentes à drogas e 81,49% com seqüências características do

subtipo B. Finalmente, o grupo 3, constituído por 164 indivíduos, apresentou 70,73% de indivíduos resistentes e foi majoritariamente preenchido por indivíduos infectados pelo HIV subtipo C, representando 93,29 % desses indivíduos (Tabelas 5.5 e 5.6).

Esta separação das sequências nos grupos obtidos, caracterizando-os pelos subtipos virais B, C, F e outros, que abrange os CRFs e subtipo D (Tabela 5.5) e a separação nos grupos quanto à resistência como resistentes (indivíduos em falha terapêutica) e *naïve* (indivíduos não-tratados) (Tabela 5.6) são mostrados a seguir.

Tabela 5.5 - Separação nos grupos por subtipos

Subtipos Grupo	B (%)	C (%)	F (%)	outros (%)	Total (%)
1	76,03	7,44	13,22	3,31	100,00
2	81,49	0,70	15,22	2,59	100,00
3	4,27	93,29	0,61	1,83	100,00

Tabela 5.6 – Separação nos grupos por resistência

Grupo	Resistentes (%)	Naïve (%)	Total (%)
1	100,00	0	100,00
2	79,06	20,94	100,00
3	70,73	29,27	100,00

O perfil predominante de cada grupo revelou posições com resíduos relacionados a polimorfismos de subtipos não-B e a resistência de PIs, caracterizando os grupos 1 e 3 associados, respectivamente, ao grupo de resistentes e ao subtipo C do HIV-1. Os resíduos I15V, M36I, R41N, L63P, H69K e I93L foram bem

caracterizados, observando-se os aminoácidos com maior taxa de ocorrência em cada posição selecionada nos três grupos (Tabela 5.7). Em seguida, são apresentados os percentuais de ocorrência dos outros aminoácidos nos três grupos a fim de permitir melhor análise destes resultados. (Tabelas 5.8; 5.9; 5.10).

Tabela 5.7 – Aminoácidos predominantes nas 10 posições selecionadas em cada um dos grupos.

Grupos **Posição**	**1**	**2**	**3**
10	L (36%)	L (63%)	L (72%)
15	V (54%)	I (69%)	V (71%)
20	T (36%)	K (83%)	K (65%)
35	D (46%)	E (60%)	E (76%)
36	I (73%)	M (57%)	I (83%)
37	N (66%)	N (66%)	K (80%)
41	R (59%)	R (58%)	N (79%)
63	P (69%)	P (56%)	L (47%)
69	H (81%)	H (86%)	K (98%)
93	I (55%)	I (73%)	L (91%)

Tabela 5.8 – Distribuição da ocorrência dos aminoácidos nas 10 posições no grupo 1.

Aminoácido/	Percentual no grupo 1 (%)									
Posição	10	15	20	35	36	37	41	63	69	93
R	0,83	-	-	-	-	7,44	**59,5**	0,083	0,083	-
K	-	-	-	-	-	10,74	33,06	-	12,40	-
D	-	-	-	**46,28**	-	0,083	-	-	-	-
Q	-	-	-	-	-	**66,12**	-	0,083	4,96	-
N	-	-	-	8,26	-	4,13	7,44	-	-	-
E	-	-	-	41,32	-	-	-	-	-	-
H	-	-	-	-	-	7,44	-	0,083	**80,99**	-
S	-	-	-	-	-	2,48	-	6,61	-	-
T	-	-	**36,36**	-	-	-	-	1,65	-	-
P	-	-	-	-	-	0,083	-	**68,60**	-	-
Y	-	-	-	-	-	-	-	-	0,083	-
C	-	-	-	-	-	-	-	0,083	-	-
G	-	-	-	4,13	-	-	-	-	-	-
A	-	-	0,83	-	-	-	-	5,79	-	-
M	-	-	25,62	-	23,97	-	-	-	-	-
W	-	-	-	-	-	-	-	-	-	-
L	**36,36**	-	1,65	-	0,83	-	-	11,57	-	42,98
V	18,18	**53,72**	2,48	-	2,48	-	-	1,65	-	1,65
F	11,57	-	-	-	-	-	-	-	-	-
I	33,06	46,28	33,06	-	**72,73**	-	-	0,083	-	**55,37**

Tabela 5.9 – Distribuição da ocorrência dos aminoácidos nas 10 posições no grupo 2.

Aminoácido/	Percentual no grupo 2 (%)									
Posição	10	15	20	35	36	37	41	63	69	93
R	-	0,17	16,96	-	-	0,35	**58,30**	-	1,56	-
K	-	0,17	**83,04**	0,17	-	0,52	41,00	-	1,38	-
D	-	-	-	39,10	0,17	10,73	-	0,17	-	-
Q	-	-	-	-	-	0,52	-	1,73	7,09	-
N	0,17	-	-	0,35	-	**66,44**	0,69	0,17	-	-
E	-	-	-	**60,38**	-	4,33	-	-	-	-
H	-	-	-	-	-	-	-	1,73	**86,51**	-
S	-	-	-	-	-	11,25	-	4,50	-	-
T	-	-	-	-	0,35	3,46	-	5,19	-	-
P	-	-	-	-	-	-	-	**56,57**	-	-
Y	-	-	-	-	-	0,52	-	-	3,11	-
C	-	-	-	-	-	1,38	-	1,21	0,17	-
G	-	-	-	-	-	-	-	-	-	-
A	-	-	-	-	-	0,52	-	3,98	-	-
M	-	-	-	-	**57,27**	-	-	-	-	0,17
W	-	-	-	-	-	-	-	-	-	-
L	**63,49**	0,17	-	-	1,73	-	-	23,70	0,17	26,47
V	9,81	29,93	-	-	2,42	-	-	1,04	-	-
F	3,81	-	-	-	-	-	-	-	-	-
I	22,66	**69,55**	-	-	38,06	-	-	-	-	**73,36**

Tabela 5.10 – Distribuição da ocorrência dos aminoácidos nas 10 posições no grupo 3.

Aminoácido/	Percentual no grupo 3 (%)									
Posição	10	15	20	35	36	37	41	63	69	93
R	-	-	28,05	-	-	-	-	-	0,61	-
K	-	-	**65,24**	-	-	**80,49**	10,98	-	**98,17**	-
D	-	-	-	23,17	-	1,22	0,61	-	-	-
Q	-	-	-	-	-	2,44	-	-	-	-
N	-	-	-	0,61	-	7,32	**79,27**	-	-	-
E	-	-	0,61	**75,61**	-	1,83	-	-	-	-
H	-	-	-	-	-	-	2,44	-	-	-
S	-	-	-	-	-	3,66	0,61	0,61	-	-
T	-	-	6,10	-	1,83	2,44	0,61	3,66	-	-
P	-	-	-	-	-	-	0,61	35,37	-	-
Y	-	-	-	-	-	-	0,61	-	-	-
C	-	-	-	-	-	-	-	1,22	-	-
G	-	-	-	-	-	-	0,61	-	-	-
A	-	-	-	-	0,61	-	-	1,83	-	-
M	1,22	-	-	-	13,41	-	-	1,22	-	0,61
W	-	-	-	-	-	-	-	-	-	-
L	**72,56**	-	-	-	-	0,61	-	**46,95**	-	**91,46**
V	6,10	**71,34**	-	-	1,22		0,61	6,71	-	-
F	5,49	-	-	-	-	-	-	0,61	-	-
I	14,63	28,66	-	0,61	**82,93**	-	3,05	1,83	1,22	7,93

Capítulo 6

DISCUSSÃO

A distribuição do HIV-1 no mundo revela grande diversidade dos subtipos virais, sendo o subtipo C o de maior ocorrência devido à grande incidência em regiões populosas, principalmente na África e na Índia (UNAIDS, 2009). O subtipo B é o responsável pela maioria das infecções na Europa e EUA, sendo este o mais estudado e compreendido quanto aos mecanismos de escape e resistência aos medicamentos.

O principal subtipo circulante na América do Sul é o B, mas o F e o C também são importantes, tornando-se relevantes estudos que permitam uma melhor compreensão de sua evolução e resistência ao tratamento nestes subtipos, a fim de auxiliar clinicamente a terapia antirretroviral (GONZALEZ *et al*, 2004; BESSONG 2008).

Como a maioria dos estudos envolvendo o complexo padrão de resistência no HIV-1 considera somente o subtipo B, a ocorrência de falhas terapêuticas em indivíduos infectados pelos subtipos não-B torna-se bastante comum (KANTOR e KATZENSTEIN, 2004), uma vez que a existência de diferentes subtipos virais pode levar à respostas variadas aos medicamentos devido a polimorfismos naturais relacionados a mutações de resistência no subtipo B (OHTAKA e FREIRE, 2005; KANTOR e KATZENSTEIN, 2004).

A maioria dos métodos propostos para interpretação dos testes genotípicos utiliza treinamento supervisionado e posições de mutação para classificar a sequência do HIV-1 em susceptível ou resistente (BEERENWINKEL *et al,* 2006; BONET *et al,* 2007, DEFORCHE *et al*. 2007, SRISAWAT e KIJSIRIKUL, 2004). Os métodos não

supervisionados permitem a obtenção de grupos sem o conhecimento prévio da classificação e, uma vez que os grupos são obtidos, torna-se possível predizer a alocação de um padrão externo ao banco de dados do treinamento em um dos agrupamentos determinados (DRAGHICI e POTTER, 2003).

A motivação do presente estudo foi a observação da separação dos pacientes com sequências semelhantes de HIV-1 em grupos naturais, de forma não-supervisionada, independente de haver ou não mutações ou estarem sob tratamento. Desta forma, foi possível encontrar três grupos com características distintas. Avaliando tanto o subtipo viral quanto a resistência aos PIs foram determinados dois grupos bem definidos: um majoritariamente constituído pelo subtipo C e outro integralmente de HIV resistente. DRAGHICI e POTTER (2003) também utilizaram a técnica do SOM e avaliaram os grupos somente quanto à resistência, identificando as categorias consideradas alta, média e baixa resistência ao PIs Indinavir e Saquinavir, além de outras duas classes mais heterogêneas.

A utilização de redes neurais é um alternativa na avaliação de resistência aos PIs uma vez que há um complexo padrão de resistência conferido pela combinação de múltiplas mutações, não sendo explicado por funções lineares (WANG e LARDER, 2003). A técnica de redes neurais considera a relação não-linear entre as variáveis, sendo utilizada na previsão de mutações associadas à resistência a vários medicamentos antirretrovirais visando predizer a resposta clínica do paciente (WANG e LARDER, 2003; DRAGHICI e POTTER 2003).

No presente estudo, a seleção de atributos, baseando-se na determinação da frequência de ocorrência dos aminoácidos em cada posição e na avaliação dos grupos obtidos reduziu o número de variáveis descritoras de 99 para 10. A utilização SOM permitiu a construção de vetores protótipos, agrupando os dados em 225 neurônios topologicamente arranjados. A escolha do mapa 15 x 15 considerou a minimização dos erros topográfico e de quantização aliada ao número de neurônios ativados, onde neurônios inativos indicam valores zero para seus pesos, não contribuindo para a

posterior separação nos grupos. Em contrapartida, os neurônios com pesos diferentes de zero são constituídos por indivíduos similares. Esta prévia organização dos indivíduos em neurônios facilitou a separação dos dados em grupos.

A análise visual obtida após o SOM apresenta vantagens para a análise dos dados e dos *clusters* obtidos (VESANTO e ALHONIEMI, 2000). A visualização da matriz-U possibilitou um uso mais eficiente do algoritmo k-médias, destacando a organização de três a quatro grupos de neurônios com ativação similares, além da prévia organização das sequências semelhantes em neurônios. A matriz U obtida com o modelo reduzido delimitou melhor os três grupos formados, quando comparada à matriz do modelo completo.

A determinação do número de grupos considerados no k-médias foi realizada utilizando dois índices de validação de *cluster:* o índice Davies-Bouldin (VESANTO e ALHONIEMI, 2000) e o índice comparado, Silhueta. O índice DB é empregado como padrão no SOM toolbox, enquanto o índice Silhueta é descrito como um índice mais robusto e amplamente utilizado (BOLSHAKOVA e AZUAJE (2003) e PETRÓVIC (2006)). Segundo PETRÓVIC (2006), o índice da Silhueta produz melhor acurácia e é mais robusto, no entanto, o índice de Davies-Bouldin é computacionalmente menos complexo. Apesar de BOLSHAKOVA e AZUAJE (2003) e PETRÓVIC (2006) descreverem casos onde os resultados pelos diferentes índices são distintos, no presente estudo, os dois índices mostraram-se concordantes, indicando como melhor resultado o agrupamento dos neurônios e, conseqüentemente, dos dados, em três grupos.

A análise destes agrupamentos obtidos de acordo com as informações de resistência e subtipagem do vírus conhecidas mostraram a boa separação dos dados. Em relação aos subtipos, a maioria das seqüências é do subtipo B, mais comum, e poucos indivíduos são caracterizados pelos subtipos não-B. A determinação de um grupo (grupo 3) constituído pelo HIV subtipo C em 93,29% dos integrantes mostra uma boa separação. Destas sequências subtipo C, que totalizam 153 indivíduos, 108 são

de indivíduos com HIV resistente e 45 *naïve.* Dos outros 11 indivíduos do grupo, 7 são caracterizados pelo HIV subtipo B, 1 pelo subtipo F resistentes e 3 CRFs resistentes. Os grupos 1 e 2 apresentaram a maioria das seqüências correspondentes aos subtipos B e F. O grupo 1, integralmente composto de indivíduos resistentes, é constituído por 76,03% subtipo B e 13,22% subtipo F. Dentre os resistentes do grupo 2, o subtipo B representa 82,06% e o F 15,10% e dentre os não-resistentes, 79,34% são B e 15,70% são F.

A análise do perfil encontrado em cada grupo foi realizada observando-se algumas posições descritas como mutações e/ou polimorfismos característicos de subtipos virais. Dentre as posições encontradas no modelo (10, 15, 20, 35, 36, 37, 41, 63, 69, 93), muitas se destacam como importantes na caracterização de resistência e polimorfismos em subtipos específicos. Em ausência de exposição às drogas, a sequência da protease do HIV-1 dos subtipos B e não-B é polimórfica em 30% dos aminoácidos (KANTOR e KATZENSTEIN, 2004). Algumas destas posições são associadas à resistência no subtipo B, como as substituições nas posições 10, 20, 36, 63, 71, 77 e 93 da protease do HIV-1 (KANTOR e KATZENSTEIN, 2004) e podem estar relacionadas à redução de atividade de PIs, com evolução de diferentes mutações primárias e resistência cruzada (KANTOR e KATZENSTEIN, 2004).

A posição 10 apresentou o aminoácido L como mais frequente nos três grupos. É comum encontrar substituições L10I/V/F/R/Y (http://hivdb.stanford.edu) como mutação secundária de resistência a muitos PIs no subtipo B. Os resíduos L10I/V são características de polimorfismos (KANTOR e KATZENSTEIN, 2004) e as mutações L10F/R foram encontradas em 45% de amostras brasileiras majoritariamente constituídas do subtipo B (COUTO-FERNANDEZ, J. C *et al*, 2005). O grupo 1 revelou um percentual de 36,36% de L, seguido de 33,06% de I e 18,18% de V nesta posição. Sendo o grupo 1 constituído em sua maioria por subtipo B e 100% de sequências do vírus resistente, este resultado é esperado, uma vez que podem apresentar mutações em tal posição. Estas variantes incluem os aminoácidos I, V (KANTOR e

KATZENSTEIN, 2004), R e F (COUTO-FERNANDEZ, J. C *et al*, 2005) já descritos , como verificados na tabela 5.8. Os grupos 2 e 3 mostraram um percentual maior de L, sendo no grupo 2, 63,49% de L seguido de 22,66% de I e no grupo 3, 72,56% de L seguido de 14,63% de I.

A posição 20 da protease é caracterizada pelo aminoácido K na sequência consenso do subtipo B, apresentando a variação para T, principalmente no subtipo F e as demais alterações descritas são K20R/M/I/T/V, sendo K20R/M/I regiões polimórficas e K20T/V não polimórficas (http://hivdb6.stanford.edu). O aminoácido K foi encontrado em 83,04% das sequências do grupo 2, seguido de 16,96% de R e o T representou 36,36% do grupo 1 seguido de 33,06% de I neste último grupo. Em ambos os grupos, o aminoácido F representou cerca de 15% dos casos. No entanto, como o número de padrões F no banco de dados original era muito pequeno comparado ao número de amostras do subtipo B (Tabela 4.1), a análise não é conclusiva. No grupo 3, constituído quase 94% do subtipo C, o aminoácido K foi encontrado em 65,24% dos casos, seguido da variante K20R.

O resíduo na posição 36 é caracterizado pela variante M36I, principalmente em seqüências consenso de subtipos não-B (COUTO-FERNANDEZ, J. C *et al*, 2005) e fracamente associada ao subtipo B resistente a PI (http://hivdb6.stanford.edu). No grupo 1, caracterizado em sua maioria pelo subtipo B resistente e no grupo 3 caracterizado pelo subtipo C, o aminoácido I foi encontrado em 72,73% e 82,93% das sequências, respectivamente. No grupo 2, mais heterogêneo quanto aos subtipos e à resistência, esta posição apresentou o aminoácido M em 57,27% dos casos e o I em 38,06%. Outras variações para a posição 36 são M36L/V/T (http://hivdb6.stanford.edu), porém a significância de tais mutações é desconhecida e foram encontradas no modelo em menores proporções.

A alteração L63P é um polimorfismo comum e que se torna ainda mais frequente com o uso de PIs (http://hivdb6.stanford.edu), sendo clara a associação principalmente no vírus resistente subtipo B (COUTO-FERNANDEZ, J. C *et al*, 2005).

A análise de amostras brasileiras, do Rio de Janeiro, revelou esta mutação secundária em 72,4% das sequências observadas, considerando um grupo de amostras constituído 91,2% de subtipos B, somente 0,4% de subtipos C e os demais, subtipos F ou CRFs B/F (COUTO-FERNANDEZ, J. C *et al*, 2005). No presente trabalho, os grupos 1 e 2 mostraram esta alteração na maioria das sequências, enquanto no grupo 3, característico do subtipo C, o aminoácido L foi verificado na maior parte das sequências analisadas (Tabela 5.7). A alteração L63P foi verificada em 68,60% dos casos do grupo 1, revelando o perfil de sequências do HIV-1 resistentes e 56,57% do grupo 2 (Tabela 5.9).

Alguns resíduos já descritos não são encontrados no Stanford *database*, uma vez que estão relacionados a polimorfismos de subtipos não-B. O resíduo H69K é um polimorfismo característico do subtipo C (GONZALEZ *et al*, 2004; BESSONG 2008) que foi bem destacado (Tabela 5.7). Na posição 69, os grupos 1 e 2 revelaram 80,99% e 86,51% das sequências com o aminoácido H, enquanto o grupo 3, apresentou um percentual de 98,17% com o aminoácido K, caracterizando o subtipo C.

A posição 93 é caracterizada por mutação secundária I93L fracamente associada a PI no subtipo B, sendo também uma região polimórfica. O grupo 1 apresentou percentuais próximos destes aminoácidos, 55,37% de I, seguido de 42,98% de L; o grupo 2 apresentou em sua maioria o aminoácido da sequência consenso do subtipo B, 73,36% de I. No grupo 3, representado pelo subtipo C, foi encontrado o aminoácido L em 91,46% das sequências em virtude de ser uma região polimórfica encontrada como consenso em muitos subtipos não-B (http://hivdb6.stanford.edu).

A posição 35 da protease não é uma região polimórfica, sendo caracterizada no subtipo B como uma mutação secundária após terapia com PIs com a variante E35G (http://hivdb.stanford.edu). No entanto, o resíduo E35D foi descrito como diferença entre os subtipos B e F em amostras brasileiras (WALERIA-ALEIXO *et al,* 2008). Analisando os grupos encontrados, no grupo 1, 46,28% apresentou o

aminoácido D, seguido de 41,32% de E. Nos grupos 2 e 3, os percentuais de E foram mais altos, 60,38% e 75,61% seguidos de 39,10% e 23,17% do aminoácido D, respectivamente. O aminoácido G sugerido no Stanford *database* foi encontrado somente no grupo 1 em apenas 4,13% das sequências.

Dentre as demais posições selecionadas no modelo, as posições 15 e 41 já foram citadas quanto às variantes comparando os subtipos B e não-B no Brasil (WALERIA-ALEIXO *et al*, 2008; GONZALEZ *et al*, 2004) e na África (BESSONG, 2008). As sequências consenso dos subtipos B diferem quando comparadas ao subtipo F nas posições I15V, E35D, M36I, R41K, R57K, Q61N, L63P e L89M e ao subtipo C nas posições I15V, M36I, R41K, H69K, L89M e I93L (BESSONG, 2008; GONZALEZ *et al*, 2004) e ainda nos resíduos T12S, L19I, K20R (BESSONG, 2008).

Analisando a posição 15 da protease em nosso modelo, o grupo 1 apresentou 53,72% de V e 46,28% de I. No grupo 2, o aminoácido I encontra-se em 69,55% das sequências e o grupo 3, claramente constituído de subtipo C, apresentou 71,34% do aminoácido V. Este resultado sugere a associação do variante I15V para subtipos não-B, incluindo o subtipo C, como já descrito (BESSONG, 2008, GONZALEZ *et al*, 2004).

A posição 41, os grupos 1 e 2, constituídos pela maior parte dos subtipos B e F da amostra, apresentaram o aminoácido R em 59,50% e 58,30%, seguido do aminoácido K em 33,06% e 41,00%, respectivamente. O grupo 3 não apontou uma grande variação no aminoácido desta posição, não revelando sequências com o aminoácido R, somente 10,98% do aminoácido K e a grande maioria, 79,27% apresentou o aminoácido N, sugerindo uma mutação R41N como característica do subtipo C, além da R41K já descrita (BESSONG, 2008, GONZALEZ *et al*, 2004).

A única posição considerada ainda não descrita nas bases referenciadas foi a posição 37, que mostrou diferença nos três grupos obtidos. O grupo 1, representado principalmente pelos vírus resistentes B e F apresentaram a 66,12% do aminoácido Q, o grupo 2, 66,44% de N e no grupo 3, característico do subtipo C, 80,49% de K.

A análise do perfil de cada grupo revelou que muitas posições selecionadas estão associadas a mutações secundárias na protease do HIV-1, como as posições 20, 36 e 63, que também foram encontradas como predominantes em amostras brasileiras analisadas por COUTO-FERNANDEZ, J. C. *et al* (2005). Observa-se a clara associação da mutação M36I ao subtipo não-B e L63P ao subtipo B, como também foi visto por COUTO-FERNANDEZ, J. C. *et al* (2005). Estas diferenças podem acarretar diferentes níveis de resistência aos medicamentos devido às barreiras genéticas entre os subtipos B e não-B do HIV-1, podendo influenciar na progressão da resistência aos antirretrovirais (BRINDEIRO *et al,* 2003).

Algumas posições relacionadas a polimorfismos de subtipos não-B e outras associadas a resíduos de mutações primárias a PI não foram encontradas no modelo, como a D30N (GONZALEZ *et al*, 2004, COUTO-FERNANDEZ, J. C. *et al*, 2005) e a L90M encontrada em baixa proporção das amostras brasileiras analisadas por COUTO-FERNANDEZ, J. C. *et al* (2005). Uma vez que foram selecionadas as posições de maior variabilidade dentre os exemplos, o modelo pode ter descartado variáveis importantes. Como não foi considerada a comparação com a sequência consenso nesta seleção, posições em que muitos exemplos modificaram o aminoácido não foram selecionadas.

Nesta análise, deve-se considerar a possibilidade de indivíduos em falha terapêutica semelhantes a indivíduos *naïve,* como consequência de uma possível transmissão do vírus mutado, e, portanto, com sequências da protease alteradas ainda que o indivíduo não tenha sido submetido à pressão das drogas antirretrovirais. A transmissão de HIV-1 resistentes aos medicamentos foi descrita para o subtipo B, na década de 90, principalmente nos EUA e Europa, onde a HAART é mais acessível (BODEN *et al*, 1999 *apud* KANTOR e KATZENSTEIN, 2004). Mutações associadas à terapia têm sido observadas em até 20% dos casos de infecção aguda pelo HIV-1 e no início da infecção (KANTOR e KATZENSTEIN, 2004).

Capítulo 7

CONCLUSÃO

A utilização da técnica SOM mostrou-se eficiente na análise dos dados ao ser empregado antes da técnica de aglomerados, apresentando vantagens no entendimento dos padrões de entrada: o conjunto de dados é representado por um pequeno grupo de vetores protótipos, reduzindo a dimensionalidade dos dados e facilitando a organização destes protótipos em grupos; as formas de visualização disponíveis no SOM toolbox permitem a observação da matriz-U de distâncias entre os elementos e a prévia identificação de grupos, além da análise da influência das variáveis na separação dos elementos topologicamente arranjados pela matriz de distâncias.

O SOM *toolbox* utilizou a distância euclidiana na determinação do neurônio vencedor, que foi mantida como padrão na técnica K-médias. A métrica sugerida como escolha tratando-se de aminoácidos e, portanto, variáveis categóricas, seria a distância de *Hamming*, que mostra o número de *bits* (aminoácidos) que diferem as sequências comparadas. Em trabalhos futuros, pode-se explorar o desenvolvimento do SOM *toolbox* utilizando esta distância para verificar se os resultados seriam superiores aos obtidos no presente estudo. A comparação entre os métodos de avaliação de *cluster* apontou para desempenhos semelhantes, uma vez que foram concordantes para o banco de dados utilizado. Desta forma, ambas as técnicas, Davies-Bouldin e Silhueta, mostraram-se eficientes na determinação do número k de grupos a ser utilizado no k-médias.

Portanto, a sequência de técnicas realizadas apresentou resultados satisfatórios, com grupos bem caracterizados e seleção de variáveis importantes na descrição de polimorfismos e mutações de resistência aos PIs. A caracterização de um grupo constituído pelo subtipo C permitiu a identificação de posições que diferenciam as sequências consenso deste subtipo e do subtipo B, cujo estudo é amplamente explorado.

As principais diferenças entre os subtipos precisam ser mais exploradas a fim de determinar o curso de evolução da infecção pelos subtipos não-B, seus mecanismos de resistência e que permitam avaliar a importância de vários polimorfismos presentes em cada isolado.

REFERÊNCIAS BIBLIOGRÁFICAS

BARRÉ-SINOUSSI, F., CHERMANN, J. C., REY, F. *et al*. 1983. “Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)”, ***Science***, v. 220 n.4599, pp.868–71.

BEERENWINKEL, N., SHIMIDT, B., WALTER, H. *et al* 2002. “Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype”. ***PNAS*** v.99, n.12 (jun), pp. 8271-8276.

BESSONG, P. O. 2008. “Polymorphisms in HIV-1 subtype C proteases and the potencial impact on protease inhibitors”. ***Tropical Medicine and International Health***, v.13, n.2 (fev) , pp.144-151.

BOLSHAKOVA, N., AZUAJE, F. 2003 “Cluster Validation Techniques for Genome Expression Data” ***Signal Processing***, v.83, pp. 825-833.

BONET, I., GARCIA, M. M., SAYES, Y. *et al*. 2007 “Predicting Human Imunnodeficiency Vírus (HIV) drug resistance using recurrent neural networks”. IWINAC Part I, ***LNCS*** v. 4527, pp. 234–243.

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

BUONAGURO, L., TORNESELLO, M. L., BUONAGURO, F. M. 2007 “Human Immunodeficiency Virus type 1 subtype B distribuition in the worldwide epidemic: pathogenetic and therapeutic implications”. ***Journal of Virology***, v.81, n.19 (out), pp. 10209-10219..

CARRION, G., EYZAGUIRRE, L., MONTANO, S. M. *et al* , V. 2004 “Documentation of subtype C HIV type 1 strains in Argentina, Paraguay, and Uruguay”. ***AIDS Research and Human Retroviruses***, v.20, n.9, pp.1022–1025..

CASTRO, E.,MORENO, M., DEIBIS, L. *et al* 2005. “Trends of HIV-1 molecular epidemiology in Venezuela: introduction of subtype C and 77 identification of a novel B/C mosaic genome”. ***Journal of Clinical Virology***, v.32, pp.257–258.

CÉRÉGHINO, R., PARK, Y.-S. 2009. “Review of the Self-Organizing Map (SOM) approach in water resources: Commentary ”. ***Environmental Modelling & Software***, v.24, (fev), pp.945-947.

COHEN, O. J., KINTER, A., FAUCI, A. S. 1997. “Host factors in the pathogenesis of HIV disease”. ***Immunolog. Rev.*** v.159, pp. 31-48.

COUTO-FERNANDEZ, J. C., SILVA-DE-JESUS, C., VELOSO, V. S. *et al.* 2005 “Human immunodeficiency virus type 1 (HIV-1) genotyping in Rio de Janeiro, Brazil: assessing subtype and drug-resistance associated mutations in HIV-1 infected individuals failing highly active antiretroviral therapy”, ***Mem Inst Oswaldo Cruz****, v. 100, n.*1 (fev), pp. 73-78..

DAVIES, D. L., BOULDIN, D. W. 1979 “A cluster separation measure”. ***IEEE Transactions on Pattern Recognition and Machine Intelligence***, v.1, n.2, pp.224-227.

DEFORCHE, K. SILANDER, T., CAMACHO, R., *et al*. 2006. “Analysis of HIV-1 pol sequences using Bayesian networks: implications for drug resistance”. ***Bioinformatics,*** v.22, n.15, pp. 2975-2979.

DRAGHICI, S., POTTER, B. 2003 “Predicting HIV drug resistance with neural networks”. ***Bioinform*atics**, v.19, n.1, pp. 98-107..

DONI, M., V. 2004. ***Análise de cluster: métodos hieráquicos e de particionamento.*** Disponível em: http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF
Acesso em 03. Fev. 2010.

EISENBERG, D., WEISS, R. M, TERWILLIGER, T. C. 1984. “The hydrophobic moment detects periodicity in protein” ***PNAS***, v.81,pp. 140-144.

GALLO, R. C., SARIN, P. S., GELMANN, E. P. *et al.* 1983. “Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS)”. ***Science***, v.220, n.4599, pp.865–867.

GAO, F., BAILES, E., ROBERTSON, D. L. *et al*. 1999 “Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes”. ***Nature***, v. 397, pp.436–441..

GONSALEZ, C. R., ALCALDE, R., NISHIYA, A. *et al*. 2007 “Drug resistance among chronic HIV-1-infected patients naïve for use of anti-retroviral therapy in Sao Paulo city”. ***Virus Research*** v. 129, pp. 87–90.

GONZALEZ, L.M.F., BRINDEIRO, R.M., AGUIAR, R.S. *et al.* 2004. “Impact of Nelfinavir Resistance Mutations on In Vitro Phenotype, Fitness, and Replication Capacity of Human immunodeficiency Virus Type 1 with Subtype B and C Proteases”. ***Antimicrobial agents and chemotherapy***. v.48, n.9 (set), pp. 3552–3555.

GONZALEZ, L.M.F., AGUIAR, R.S., AFONSO, A. *et al.* 2006. “Biological characterization of human immunodeficiency virus type 1 subtype C protease carrying indinavir drug-resistance mutations”. ***Journal of General Virology*** v. 87, pp.1303–1309.

HIRSCH, V. M., OLMSTED, R. A., MURPHEY-CORB, M*. et al.* 1989 “An african primate lentivirus (SIVsm) closely related to HIV-2”. ***Nature***, v. 339, n.6223, pp.389–392..

HOGAN, C. M., HAMMER, S. M. 2001 “Host determinants in HIV infection and disease. Part 2: genetic factors and implications for antiretroviral therapeutics”, ***Ann. Intern. Med***. V.134, pp.761-776..

JAIN, A. K., MURTY, M. N., FLYNN, P. J. 1999 “Data Clustering: A Review”. ***ACM Computing Surveys***, V.. 31, N. 3 (set)..

KANTOR, R., KATZENSTEIN, D. 2004. “Drug resistence in non-subtype B HIV-1”. ***Journal of Clinical Virology*** v. 29, pp. 152-159.

KASK, S. 1997. “Data exploration using self-organizing maps”. ***Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series*** n. 82, Finish Academy of Technology..

KOHONEN, T., OJA, E., SIMULA, O. et al. 1996. “Engineering Applications of the Self – Organizing Map”. ***Procedings of the IEEE***, v.84, n.10 (out).

KOUP, R. A., SAFRIT, J. T., CAO, Y. *et al.* 1994. “Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome” ***J. Immunol*** v. 68, pp.4650-4655.

LANDIS, J. R., KOCH, G. G. 1977. “The Measurement of Observer Agreement for Categorical Data”. ***Biometrics***, v. 33, n.1 (mar), pp. 159-174.

LEVY, J. A., HOFFMAN, A. D., KRAMER, S. M. *et al.* 1984. “Isolation of lymphocytopathic retroviruses from San Francisco patients with AIDS”. ***Science***, v. 225, n.4664, pp.840–842..

MITRA, P., MURTHY, C. A., PAL, S. K. 2002. “Unsupervised Feature Selection Using Feature Similarity”. ***IEEE,*** v.24, n3, pp. 301-312.

MORGADO, M.G.; GUIMARAES, M.L.; GRIPP, C.B.W.G *et al*. HEC AIDS Clin Res Group. 1998 “Molecular epidemiology of HIV in Brazil: High prevalence of B subtype and identification of an HIV-1 subtype D infection in Rio de Janeiro City”. ***J of AIDS & Human Retrov***, v. 18, p. 488-494..

MORGADO, M.G.; SABINO, E.; SPHAER, E. *et al*. 1994 “Polymorphism in the v3 Region Of The Envelope Protein Of HIV-1 in Brazil: Divergence From Prevalent North American/European Subtype B Strains and Identification of Newly Described F Subtype”. ***Aids Res. Hum Retrov****,* v. 10, pp.569-576..

OHTAKA H, FREIRE E. 2005.” Adaptive inhibitors of the HIV-1 protease - Review.” ***Prog Biophys Mol Biol.*** v. 88, n.2 (jun), pp.193-208.

PANTALEO, G., FAUCI, A. S. 1996 “Immunophatogenesis of HIV infection”. ***Annu Rev Microbiol. Review,*** v.50, pp. 825-54..

PÉTROVIC, S. 2006. “A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters”. ***Proceedings of the 11th Nordic Workshop on Secure IT-systems****, NORDSEC* pp.53-64.

POTTER, R. B., DRAGHICI, S. 2002. “A SOFM APPROACH TO PREDICTING HIV DRUG RESISTENCE”. ***Pacific Symposium on Biocomputing*** v.7, pp.77-87.

POTTS, K.E.; KALISH, M.L.; LOTT, T. *et al* 1993. “Genetic heterogeneity of the V3 region of the HIV-1 envelope glycoprotein in Brazil”. Brazilian Collaborative AIDS Research Group. ***AIDS*****,** v. 7, pp. 1191-1197.

RAMBAUT, A., POSADA, D., CRANDALL, A. K. *et al*. 2004. “The causes and consequences pf HIV evolution”. ***Reviews Nature*** v. 5 (jan).

RÖGNVALDSSON, S., YOU, L.:2004. Why neural networks should not be used for HIV-1 protease cleavage site prediction. ***Bioinformatics*** v. 20, n.11, pp.1702-1709.

SAIGO, H., UNO, T., TSUDA, K. 2007 “Mining complex genotypic features for predicting HIV-1 drug resistance”. ***Bioinformatics*** v.23, n.18, pp. 2455-2462..

SANCHES, M., KRAUCHENCO, S., MARTINS, N.H. et al. 2007. “Structural Characterization of B and non-B Subtypes of HIV-Protease: Insights into the Natural susceptibility to drug resistance”. ***J. Mol. Biol***. v. 369, pp.1029–1040.

SEVIN, A. D. 2000 “Methods for investigation of the relationship between drug-susceptibility phenotype and human imunnodeficiency virus type 1 genotype with applications to AIDS Clinical Trials Group” ***Journal of infections diseases*** v.1, n.182, pp. 59-67..

SILVA, R. M. 2009. ***Algoritmo genético e kernel discriminante de Fisher aplicado a identificação de mutações de resistência do HIV-1 aos inibidores antiretrovirais da protease.*** Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil..

SIMON, V., HO, D. D. 2003. “HIV-1 dynamics in vivo: Implications for therapy”. ***Nature Reviews - Microbiology,*** v.1, pp.181–190.

SMITH, L. I. 2002. “A tutorial on Principal Components Analysis”(fev). Disponível em: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Acesso em 10 de março de 2008

SRISAWAT, A., KIJSIRIKUL, B. 2004. “Using associative classification for predicting HIV-1 drug resistence”. In: ***Proceedings of the fourth International Conference on Hybrid Intelligence Systems, IEEE.*** (HIS’04) - 0-7695-2291-2/04..

UNAIDS, 2009. “AIDS *epidemic update*” Disponível em: http://www.unaids.org/en/HIV_data/Epidemiology/default.asp Acesso em 07 mar. 2009.

VESANTO, J., ALHONIEMI, E. 2000 “Clustering of the Self-Organizing Map”. ***IEEE Transactions on Neural Networks***, v.11, n.3 (mai)..

WALÉRIA-ALEIXO, A. MARTINS, A.N., ARRUDA, M.B. *et al*. 2008. “Drug resistance mutation profile and accumulation kinetics in Human Immunodeficiency Virus-positive individuals infected with subtypes B and F failing highly antiretroviral therapy are influenced by different viral códon usage patterns”. ***Antimicrobial Agents and Chemotherapy***, v.52, n.12 (dez), pp.4497-4502.

WANG, D., LARDER, B. 2003. “Enhaced prediction of Lopinavir resistance from genotype by use of artificial neural networks”. **Journal of infections diseases** v.1, n. 188, pp.653-60.

ZANINI, A. 2004. ***Regulação econômica no Setor Elétrico Brasileiro: Uma metodologia para definição de fronteiras de Eficiência e cálculo do fator X para empresas distribuidoras de Energia Elétrica.*** Tese de D.Sc., PUC/RJ, Rio de Janeiro, RJ, Brasil..

ZOLOPA, A. R., SHAFER, R, W., SHAFER, R. W. *et al*. 1999. “HIV-1 genotypic resistance predict response to Saquinavir – Ritonavir therapy in patients in whom previous protease inhibitor therapy had failed”. ***Ann Intern Med.*** v.131, pp.813-821.