REDES NEURAIS SEM-PESO APLICADAS NA CATEGORIZAÇÃO DE SUBTIPOS DO HIV-1.

Caio Ribeiro de Souza

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

Orientadores: Felipe Maia Galvão França
Robson Mariano da Silva

Rio de Janeiro
Junho de 2011

REDES NEURAIS SEM-PESO APLICADAS NA CATEGORIZAÇÃO DE SUBTIPOS DO HIV-1.

Caio Ribeiro de Souza

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 CIENCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____
Prof. Felipe Maia Galvão França, Ph.D.

_____
Prof. Robson Mariano da Silva, D.Sc.

_____
Prof. Flávio Fonseca Nobre, Ph.D.

_____
Prof. Ricardo Costa Farias, Ph.D.

_____
Prof. Paulo Costa Carvalho, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
JUNHO DE 2011

Souza, Caio Ribeiro de

Redes Neurais Sem-Peso Aplicadas Na Categorização De Subtipos Do Hiv-1./ Caio Ribeiro de Souza – Rio de Janeiro: UFRJ/COPPE, 2011.

X, 91 p.: il.; 29,7 cm.

Orientadores: Felipe Maia Galvão França

Robson Mariano da Silva

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2011.

Referências Bibliográficas: p. 80-87.

1. Resistência HIV-1. 2. Redes Neurais Sem Peso. 3. *Bleaching*. I. França, Felipe Maia Galvão, *et al*. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Titulo.

“Aos meus pais, Antônio Carlos e Adélia. Suas vidas sempre me serviram de exemplo para buscar meus objetivos e me deram força e determinação.”

Agradecimentos

Aos meus pais por me fornecerem todas as condições necessárias para a minha formação acadêmica e principalmente por tudo que representam na minha vida.

Às minhas irmãs por todo incentivo e motivação dado ao longo desse processo.

Ao professor Felipe Maia Galvão França por seu apoio e orientação segura, tranquila e sempre paciente.

Ao professor Robson Mariano da Silva por sua orientação segura e por sua pesquisa ter sido a fonte de inspiração para a realização deste trabalho.

A professora Priscila Machado Vieira Lima por seu apoio incondicional sempre que necessário.

Ao professor Flávio Fonseca Nobre pelo auxílio prestado em importantes momentos desta pesquisa.

Às amigas Teresa Costa Barros, Vanessa Carla Felipe Gonçalves, Janaina Marques e Miriam Rodrigues por toda a ajuda prestada.

Aos amigos da UFRJ com quem tive o prazer de conviver desde a graduação e que me incentivaram durante todo esse período.

À Fundação COPPETEC e a DAbM (Diretoria de Abastecimento da Marinha do Brasil) por terem possibilitado meu afastamento nos momentos necessários, permitindo meu aprimoramento acadêmico.

Ao professor Rodrigo Brindeiro, chefe do Laboratório de Virologia Molecular (Instituto de Biologia - UFRJ, Brasil) por ter cedido a base de dados utilizada e fundamental para o desenvolvimento deste trabalho.

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

REDES NEURAIS SEM-PESO APLICADAS NA CATEGORIZAÇÃO DE SUBTIPOS DO HIV-1.

Caio Ribeiro de Souza

Junho/2011

Orientadores: Felipe Maia Galvão França
Robson Mariano da Silva

Programa: Engenharia de Sistemas e Computação

Atualmente, uma das causas mais importantes da ocorrência de falha terapêutica em pacientes infectados com o HIV-1 e sob tratamento é o acúmulo de mutações genéticas de resistência aos antiretrovirais disponíveis. Estudos recentes mostram que a realização de testes genotípicos de resistência a esses medicamentos são muito importantes para o tratamento do HIV-1. Estes testes podem auxiliar na redução das falhas terapêuticas. Porém, a dificuldade na interpretação genética das mutações ainda é um fator limitante. O objetivo desta dissertação é desenvolver uma rede neural sem peso capaz de categorizar os diferentes subtipos do HIV-1 e também de identificar a existência de mutações de resistência a este tipo de droga. O conjunto de dados utilizado consiste de 1205 amostras da sequência genética da Protease do HIV-1, provenientes de pacientes de subtipos B, C e F sob falha terapêutica. Tais dados foram obtidos junto ao Laboratório de Virologia Molecular (UFRJ, Brasil). Diversos experimentos com diferentes configurações foram realizados, e os resultados encontrados mostraram que as redes sem peso possuem excelente desempenho para o reconhecimento dos subtipos.

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

WEIGHTLESS ARTIFICIAL NEURAL NETWORKS (WANN) APPLIED TO CATEGORIZATION OF SUBTYPES OF HIV-1.

Caio Ribeiro de Souza

June/2011

Advisors: Felipe Maia Galvão França
Robson Mariano da Silva

Department: Computer and Systems Engineering

Nowadays, one of the most important causes of ARV therapy failure, in patients with HIV-1 and under treatment is the accumulation of resistance mutations to antiretroviral available drugs. Recent researches show that genotypic resistances testing of these drugs are very important to treatment of HIV-1. These tests can reduce the therapy failure. However the difficulty in interpreting genetic mutations is still a limiting factor. The aim of this dissertation is develop a WANN (Weightless Artificial Neural Network) that should be capable to categorize the different subtypes of HIV-1 and also to identify the existence of antiretroviral drugs resistance mutations. The data set used consists of 1205 gene sequence of the HIV-1 protease from patients with subtypes B, C and F under treatment failure. This data were obtained from the Laboratory of Molecular Virology (UFRJ/BRAZIL). Various experiments, with different configurations, have been done. The results showed WANN was culpable of properly recognizing subtypes.

Sumário

1 Introdução

1.1 Motivação

Há mais de 30 anos, foi registrado o primeiro caso de óbito em decorrência da Síndrome da Imunodeficiência adquirida (SIDA, em inglês AIDS). Na última década, muito vem sendo feito para combater essa doença, mas ainda hoje ela é considerada uma pandemia mundial.

De acordo com o último boletim da UNAIDS, cerca de 34 milhões de pessoas estavam infectadas com o vírus da imunodeficiência humana (VIH, em inglês HIV) no final de 2010. Além disso, estima-se que 30 milhões de pessoas morreram de causas relacionadas à AIDS desde que o primeiro caso da doença foi reportado. Este boletim informa ainda que a taxa global de novas infecções foi reduzida em aproximadamente 25% entre os anos de 2001 e 2009. Na Índia, essa taxa caiu mais de 50%, enquanto na África do Sul, mais de 35%, sendo que tais países são os que possuem o maior número de pessoas convivendo com o HIV em seus continentes. Acredita-se que a redução do numero de pessoas infectadas, seja decorrente de uma maior conscientização e de comportamentos sexuais mais seguros nos últimos 10 anos (UNADIS, 2011).

Recentemente, houve progressos significantes na prevenção de novas infecções em crianças, em decorrência do aumento do número de mulheres vivendo com o HIV que tiveram acesso à profilaxia antiretroviral durante a gravidez, parto e amamentação. Assim, o número de novas crianças infectadas com o vírus em 2009 foi 26% menor do que o registrado em 2001. Porém, em números absolutos, é estimado que 16,6 milhões de crianças tenham perdido um ou ambos os pais por doença relacionada à AIDS, a grande maioria na África Subsaariana (UNADIS, 2011).

No Brasil, segundo o MINISTÉRIO DA SAÚDE (2011), no ano de 2000 eram registrados 31 mil novos casos de AIDS por ano e em 2009 esse número subiu para 38 mil. Embora muitos considerem isso um dado preocupante, o MINISTÉRIO DA SAÚDE (2011) alega que esse aumento se deve a uma incessante busca pelo diagnóstico da doença ainda em estado precoce. A estimativa é de que 630 mil pessoas vivam com o vírus no país e destas, pelo menos 255 mil não sabem de tal fato ou nunca fizeram o teste de HIV.

A expectativa do MINISTÉRIO DA SAÚDE (2011) é de que o crescimento do número de testes resulte em um aumento do número de casos de AIDS no país e de pessoas em tratamento, o que seria extremamente positivo pelo ponto de vista de que um maior número de pessoas diagnosticadas implica em uma maior prevenção, e maior abrangência do acompanhamento, possibilitando assim uma resposta mais efetiva ao tratamento. Consequentemente, haveria redução no número de mortos pela doença e um aumento na qualidade de vida das pessoas sob tratamento. Além disso, o Governo Brasileiro adotou desde 1991 uma política que visa garantir o acesso universal à terapia com antiretrovirais (ARV) para indivíduos portadores do HIV-1. Essa política tem causado um grande impacto na epidemia de HIV/AIDS, reduzindo a morbidade e a mortalidade do vírus. Corroborando com esta idéia, a própria Organização Mundial da Saúde (OMS, em inglês WHO) tem alertado sobre a importância e necessidade da diminuição dos preços dos medicamentos antiretrovirais (UNADIS, 2011).

Entretanto, o HIV-1 possui uma enorme variabilidade genética e antigênica, decorrente da sua elevada taxa mutação, que é estimada em 1% ao ano. Esta taxa possibilita que distintas variantes virais convivam no mesmo indivíduo infectado (MORGADO, 2000), permitindo o aparecimento de amostras resistentes aos medicamentos. Segundo SHAFER *et al. (*1998) e VELLA e PALMISANO (2000), esta é uma das principais causas das falhas terapêuticas e também o mais sério obstáculo da terapia antiretroviral (VELLA, 2002).

Segundo o MINISTÉRIO DA SAÚDE (2011), a resistência é uma consequência direta da diversidade do vírus, da não adesão ao tratamento e de problemas fármaco biológicos com os antiretrovirais (ARV). Conforme será exposto no Capítulo 2, diversos estudos complementam esta afirmação, alertando que o acúmulo de mutações de resistência e a replicação continuada do vírus fazem com que a suscetibilidade às drogas diminua. Desta forma a potência dos componentes do esquema terapêutico é progressivamente reduzida.

Torna-se necessário, portanto, a utilização de testes laboratoriais de avaliação de resistência do HIV-1 à terapia antiretroviral. Tais testes permitem verificar a presença de mutações, baseando-se na análise do genoma viral, que visa identificar mutações associadas à resistência (teste genotípico), ou na medida direta *in vitro,* da suscetibilidade do vírus aos ARV (teste fenotípico). Porém, segundo Hirsch *et al. (*2003) a interpretação da resistência a partir do genótipo ainda é um grande desafio.

1.2 Estudos Correlatos

Diversos métodos computacionais já foram utilizados para estudar o mecanismo de mutação de resistência às drogas, e também para desenvolver ferramentas de predição. Eles tem se mostrado muito úteis e ainda vem sendo desenvolvidos. Métodos de aprendizado estatístico, como redes neurais, máquinas de vetores de suporte e árvores de decisão também têm mostrado potencial promissor para a previsão de mutação de resistência (CAO *et al.,* 2005).

Em relação ao mecanismo de mutação podemos destacar os estudos de DIRIENZO *et al. (*2003), DEFORCHE *et al. (*2006) e SILVA (2009). Usando diferentes abordagens, esses trabalhos tentam relacionar o genótipo da protease com a existência de resistência aos antiretrovirais. DIRIENZO *et al*. (2003) foca no medicamento Amprenavir (APV) e usa um método não paramétrico que permitiu identificar oito códons capazes de caracterizar a existência da resistência para esta droga. DEFORCHE *et al. (*2006) faz uso de redes neurais bayesianas para identificar polimorfismos para as drogas Indinavir (IDV), Saquinavir (SQV) e Nelfinavir (NFV) e seu trabalho identificou como principais mutações de resistência as seguintes posições: 30N, 88S e 90M para o NFV, 90M para SQV e 82A/T para IDV. Usando um modelo computacional híbrido baseado na utilização de algoritmos genéticos (AGs) e no classificador *Kernel* Discriminante de Fisher (KDF), SILVA (2009) codifica os aminoácidos usando a escala de hidrofobicidade ponderada pelo peso molecular, visa predizer a resistência para Saquinavir, Nelfinavir e Lopinavir e identificar possíveis novas mutações de resistência. Tais experimentos obtiveram acurácia de respectivamente 88%, 81,25% e 84,93% e também foi capaz de selecionar as principais posições de mutações de resistência para os inibidores em estudo.

Além do trabalho de SILVA, (2009) os estudos de WANG e LARDER (2003), DRAGHICI e POTTER (2003), BEERENWINKEL *et al.* (2002 e 2003) e Wang *et al.* (2009) são muito importantes em relação ao reconhecimento da existência de resistência aos medicamentos. DRAGHICI e POTTER (2003) utilizaram redes neurais para categorizar amostras resistentes a Indinavir e Saquinavir. Ao trabalhar com dados estruturados eles obtiveram uma acurácia entre 60% e 70% e com dados seqüenciais o grau de reconhecimento chegou a 85%. Já o estudo de WANG e LARDER (2003) é focado na resistência em relação ao Lopinavir. Eles também utilizam redes neurais, porém fazem experimentos selecionando apenas 11 ou 28 posições da sequência de aminoácidos da protease. Para essas posições utilizadas eles conseguiram uma acurácia média de 85% e 88% respectivamente.

Nos trabalhos de BEERENWINKEL *et al.* (2002 e 2003) foram utilizadas tanto a protease quanto a transcriptase reversa do HIV-1. Nestes estudos utilizaram-se árvores de decisão e máquina de vetores de suporte e se buscava a detecção da resistência em relação a diversos medicamentos. Os resultados para os experimentos feitos com máquina de vetores de suporte obtiveram acurácia variando de 54% a 85% para os inibidores de transcriptase e de 78% a 89% para os inibidores de protease. Já no caso do experimento com árvores de decisão a taxa de acerto ficou entre 58% e 97% para os inibidores de transcriptase e de 82% a 90% para os de protease.

Wang *et al.* (2009) comparam 3 métodos, utilizam tanto a protease quanto a transcriptase reversa e usam a diferença média absoluta da mudança na carga viral e a correlação entre a previsão e a mudança na carga viral como métricas. Os algoritmos comparados são: máquinas de vetores de suporte, redes neurais e *"Random Forest",* método que se constitui na utilização de diversas árvores de decisão em paralelo. Em relação à diferença média absoluta, estes métodos apresentam respectivamente: 0.600, 0.543, e 0.607 log10 copias/ml. Já para a correção entre o observado e o previsto os resultados são de 62%, 68% e 70% respectivamente. Tal estudo informa ainda que uma combinação de rede neural com "*Random Forest*" será avaliada para utilização clínica, o que ilustra o potencial desta abordagem.

1.3 Objetivos

Essa dissertação abordará o problema da categorização do HIV-1, sendo o seu objetivo desenvolver uma rede neural sem peso, capaz de discriminar amostras de acordo com o subtipo do vírus e da existência ou não de resistência a algum medicamento. Para isso, será usada a WiSARD (Wilkes, Stonham, Aleksander Recognition Device) combinada com a técnica de *Bleaching*. A base de dados contem a sequência da protease do HIV-1 e inclui os subtipos B, C e F. Cada um desses subtipos possui amostras resistentes e *naives*, totalizando 6 categorias distintas.

Buscar-se-á também, avaliar os seguintes pontos: acurácia da rede em relação ao reconhecimento dos subtipos e em relação à existência de resistências, desempenho da técnica do *Bleaching*, aplicação da hidrofobicidade, da massa molecular e da combinação de ambas as informações na codificação dos aminoácidos da protease do HIV-1, capacidade da rede para trabalhar com grandes estruturas de dados e capacidade da rede para trabalhar com maior número de categorias, de modo a identificar também os medicamentos para os quais aquela amostra é resistente.

2 Conceitos Básicos

No Capítulo anterior mencionou-se que neste estudo será usada a WiSARD para a realização do reconhecimento de amostras do HIV-1. Será descrita a seguir, a estrutura e demais características do vírus, que são importantes para o entendimento do processo realizado, bem como, o funcionamento desta rede neural sem peso.

2.1 Fundamentos Biológicos

2.1.1 DNA

O ácido desoxirribonucléico (DNA) é uma molécula que armazena toda a informação genética necessária para coordenar o desenvolvimento e funcionamento dos seres vivos e de alguns vírus. Ela é formada por inúmeros nucleotídeos, organizados em sequencias e pareados, formando dois filamentos que se contorcem formando uma estrutura em dupla hélice. Cada nucleotídeo possui uma molécula de fosfato, uma de desoxi ribose e uma base nitrogenada. Essas bases podem variar entre adenina, timina, citosina e guanina (abreviadas A, T, C e G), e a ordem em que elas aparecem no DNA determina a informação armazenada (NELSON e COX, 2000).

2.1.2 Gene

Os genes são estruturas que fazem parte do DNA e contem a informação de proteínas que podem via a serem expressas. Neles estão presentes partes codificantes, que fornecem as instruções genéticas para construção de proteínas. Além dos genes, o DNA possui também partes intergênicas, capazes de “ligar” ou “desligar” os genes e também, de “frear” ou “acelerar” a atividade desses (NELSON e COX, 2000).

Para um melhor entendimento acerca desta estrutura, podem-se comparar as bases nitrogenadas que compõem o DNA com as letras do alfabeto. Seguindo essa analogia, o DNA seria um longo texto, enquanto os genes podem ser vistos como palavras. Tais palavras são formadas pelo agrupamento de letras e são responsáveis para dar significado ao texto (NELSON e COX, 2000).

2.1.3 RNA

O acido ribonucléico (RNA) é muito semelhante ao DNA, porém é formado por apenas um filamento de nucleotídeos. Nele, a base timina é substituída pela uracila (U). Existem diversos tipos de RNA e dentre eles podemos citar, por exemplo, o RNA mensageiro (mRNA), o de transporte ou transferência (tRNA) e o ribossomial (rRNA). Essa estrutura é importante, pois, embora os genes do DNA sejam os responsáveis por guardar a informação para gerar a proteína, para que esse processo ocorra é preciso sintetizar o RNA mensageiro (NELSON e COX, 2000).

2.1.4 Códon

Os códons são formados por três nucleotídeos. Cada mRNA possui diversos códons e durante o processo de síntese de uma proteína, são eles que codificam um determinado aminoácido que irá compor a proteína (NELSON e COX, 2000). Na figura 2.1 é possível verificar a relação entre códon e aminoácido. Nela estão destacados o códon de iniciação (AUG – Met) e os códons que determinam o fim da síntese de proteína (UAA, UAG e UGA – Fim). Além disso, pode-se observar também que cada códon codifica um único aminoácido. Porém, como existem 64 códons e apenas 20 aminoácidos conhecidos, diferentes códons podem levar a codificação de um mesmo aminoácido.

First letter of codon (5′ end)

Second letter of codon

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

Figura 2.1 "Dicionário" de palavras de código de aminoácidos em mRNAs. Os códons são escritos na direção 5'-> 3'. A terceira base de cada códon (em negrito) desempenha um papel menor na especificação de um aminoácido que os dois primeiros. Os três códons de terminação e o de inicio são sombreados na cor cinza. Todos os aminoácidos, exceto Metionina e Triptofano têm mais de um códon. Na maioria dos casos, códons que especificam o mesmo aminoácido diferem apenas na terceira base. (adaptado de NELSON e COX, 2000).

2.1.5 Síntese de Proteínas

A síntese de proteínas é um procedimento que pode ser iniciado quando um gene é transcrito. Num primeiro momento, ocorre a transcrição do DNA em RNA heterogêneo, que após ser processado, passa a conter somente as partes codificantes dos genes[1]. Em seguida, o mRNA originário do processamento do RNA heterogêneo é traduzido para proteína (NELSON e COX, 2000).

Conforme disposto na figura 2.1, os códons determinam os aminoácidos a serem codificados, bem como, o momento de parada da síntese de proteína, que é atingido quando um códon de "fim" é encontrado. Já a figura 2.2, exibe o dogma central da biologia molecular que contribui para o entendimento desse processo.

[1] As partes codificantes dos genes são denominadas "exons", enquanto as não codificantes são os "introns".

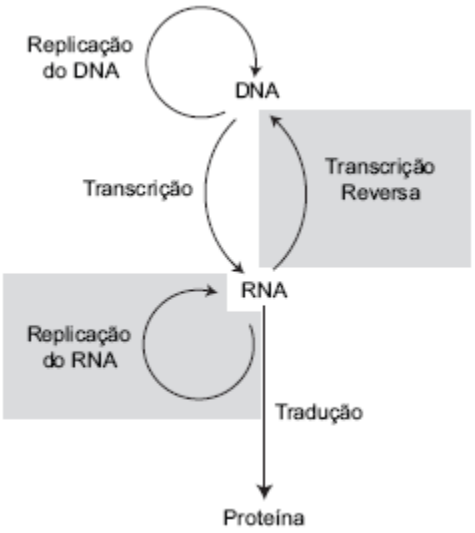

Figura 2.2 Dogma central da biologia molecular (adaptado de NELSON e COX, 2000).

A figura 2.3 exibe um trecho de DNA com o seu respectivo mRNA e também o aminoácido codificado por cada códon desse mRNA. Nesta imagem observa-se a troca da timina pela uracila, ocorrida na tradução do mRNA.

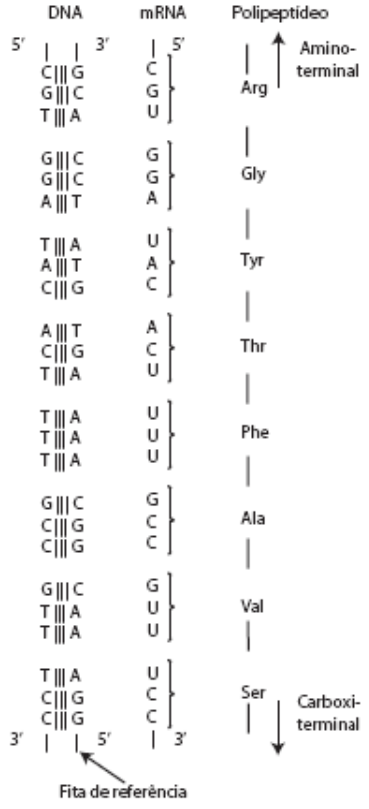

Figura 2.3 Relação entre DNA, mRNA gerado por este DNA, e aminoácidos gerados por cada códon desta parte do mRNA (adaptado de NELSON e COX, 2000).

2.1.6 Mutações

As mutações são alterações que ocorrem na sequência dos nucleotídeos, existindo diversos tipos distintos de mutações. Dentre elas, podemos destacar, por exemplo, as mutações pontuais, as de inserção e remoção. A pontual é a mais simples, pois envolve alteração de uma única posição na sequência de um gene. Neste caso, ocorre a substituição de um nucleotídeo por outro, e as demais informações não são alteradas. Seus efeitos são observados somente localmente, não se propagando para o restante da estrutura (NELSON e COX, 2000).

Nos casos de mutação por inserção ou remoção, não há substituição, e sim, ganho ou perda de um nucleotídeo. Diferente da mutação pontual que afeta um único códon, essas mutações podem afetar também todos os códons seguintes, pois alteram o quadro de leitura (i.e. *frameshift*). Com isso, todos os códons posteriores podem ser alterados (NELSON e COX, 2000). Na figura 2.4, demonstra-se a ocorrência desses tipos de mutação, bem como, seus efeitos nos códons seguintes.

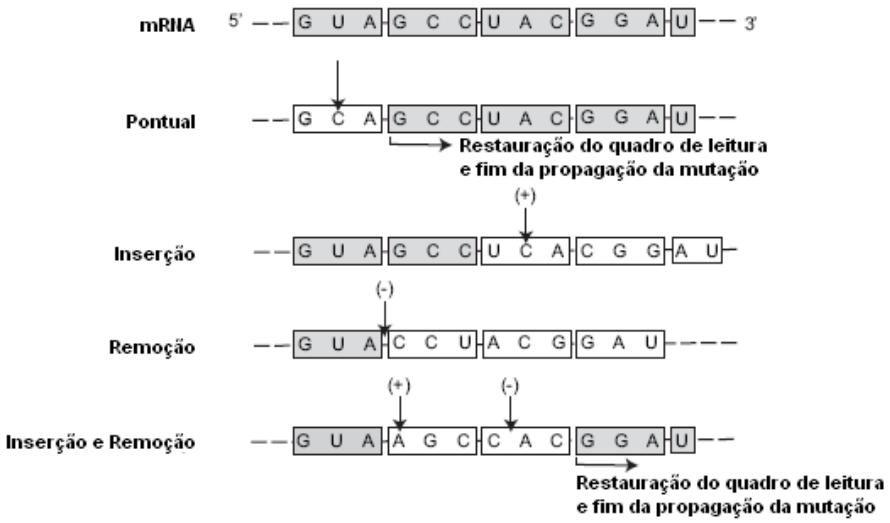

Figura 2.4 Exemplos dos diferentes tipos de mutações (adaptado de NELSON e COX, 2000).

Essas formas de mutação mencionadas podem afetar ou não a proteína gerada pelo RNA. As proteínas são formadas pelos aminoácidos, o que como relatado, são determinados pelos códons do RNA. Quando um nucleotídeo é substituído por outro, mas a mutação ocorrida não causa impacto algum na proteína, dá-se o nome de mutação "*missense*". Mutações deste tipo ocorrem, por exemplo, quando o códon "CGA" sofre mutação tornando-se "CGC". Embora a última base

tenha passado de adenina para citosina, na figura 2.1 observa-se que esses dois códons são responsáveis pela síntese da Arginina (NELSON e COX, 2000).

As mutações que alteram a função da proteína produzida são chamadas de "*nonsense*". Seus efeitos dependem da extensão dessa alteração. Tais mutações podem causar uma terminação precoce do processo de tradução do RNA, resultando assim na formação de uma proteína menor. Certamente, esta proteína terá sua função prejudicada, podendo inclusive ser não funcional (NELSON e COX, 2000). Na figura 2.5 verificam-se exemplos da ocorrência destes dois tipos de mutação.

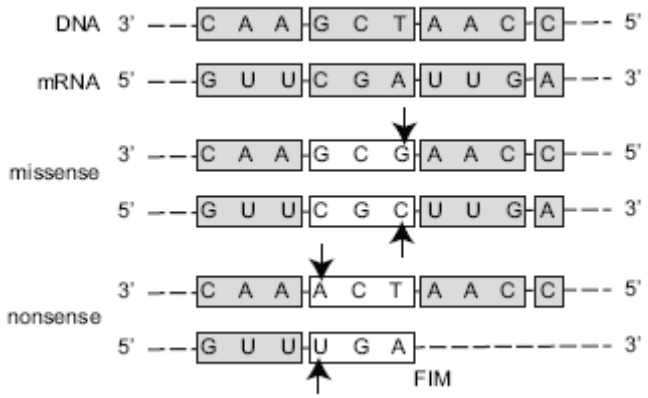

Figura 2.5 Exemplo de mutação *missense* e *nonsense* (adaptado de NELSON e COX, 2000).

2.2 O HIV-1

2.2.1 Origem

O vírus da imunodeficiência humana do tipo 1 (HIV-1) é classificado como pertencente a família *Retroviridae* e do gênero *Lentiviridae*. Diversos estudos apontam que tanto o HIV-1 quanto o HIV-2 tiveram como origem o vírus da imunodeficiência de símios (VIS em inglês SIV). GAO *et al* (1992) comparam amostras filogenéticas do SIV com o HIV e identificaram alto grau de similaridade. Além desta análise, HIRSCH *et al*. (1989) apontam que determinadas linhagens de SIV e HIV-2, que possuem homologia de cerca de 80%, foram encontrados em seres que habitam a mesma região geográfica. Para o HIV-1 este mesmo fato foi observado anos depois nos estudos de GAO *et al*. (2001). HAHN *et al.* (2000) acredita ainda que esta elevada similaridade é decorrente de múltiplas infecções entre primatas não-humanos e humanos.

2.2.2 Características

O HIV é constituído por uma fita simples de RNA e utiliza a transcriptase reversa para se replicar. Assim como os demais lentivírus, possui período de incubação prolongado, infecta as células do sangue e do sistema nervoso e age suprimindo o sistema imunológico. Desta forma ele reduz a capacidade de defesa da pessoa infectada e possibilita a entrada de doenças tidas como “oportunistas”.

Porém, sua característica mais marcante e ao mesmo tempo mais preocupante é a grande diversidade genética e antigênica. Tal diversidade pode ser superior a 10% em um individuo infectado (PINTO e STRUCHINER, 2006) e atualmente está organizada em três grupos: M, O e N. Deste, o grupo M é o principal e inclui os subtipos A, B, C, D, F, G, H, J e K. (LABORATÓRIO VIRCO, 2011). Segundo PEETERS (2000), esta organização favorece o estudo a cerca do HIV-1. Na figura 2.6 vê-se a proximidade genética entre eles.

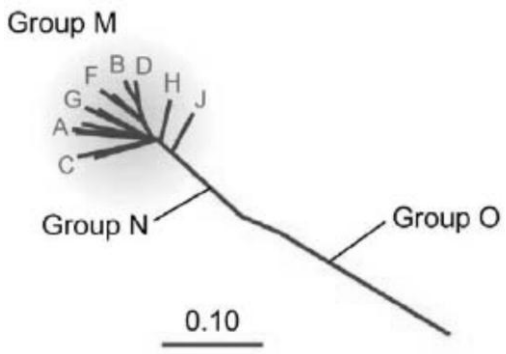

Figura 2.1 Classificação da diversidade do HIV-1. A barra representa uma diferença de 10% na sequência do nucleotídeo (adaptado de LABORATÓRIO VIRCO, 2011).

SPIRA *et al.* (2003) relatam que os subtipos do HIV-1 estão espalhados ao redor do mundo, porém em um determinado país, é comum a predominância de um subtipo em relação aos demais, ou mesmo a existência de apenas alguns deles. No Brasil, os estudos de MORGADO *et al*. (1998), TANURI *et al*. (1999) e BRINDEIRO *et al*. (2003) relatam a ocorrência do subtipo B, com taxa superior a 60%, o que o torna predominante. Além deste, foram encontrados também os subtipos C e F, com taxas inferiores a 30% e a 12% respectivamente. Porém, em seu estudo, SPIRA *et al.* (2003) ressalta que mais subtipos estão sendo constantemente descobertos, e as migrações das populações podem mudar os padrões de infecção, fato este, que, no Brasil, já foi

mencionado por SOARES *et al*. (2003). A figura 2.7, exibe os dados encontrados por BRINDEIRO *et al. (*2003).

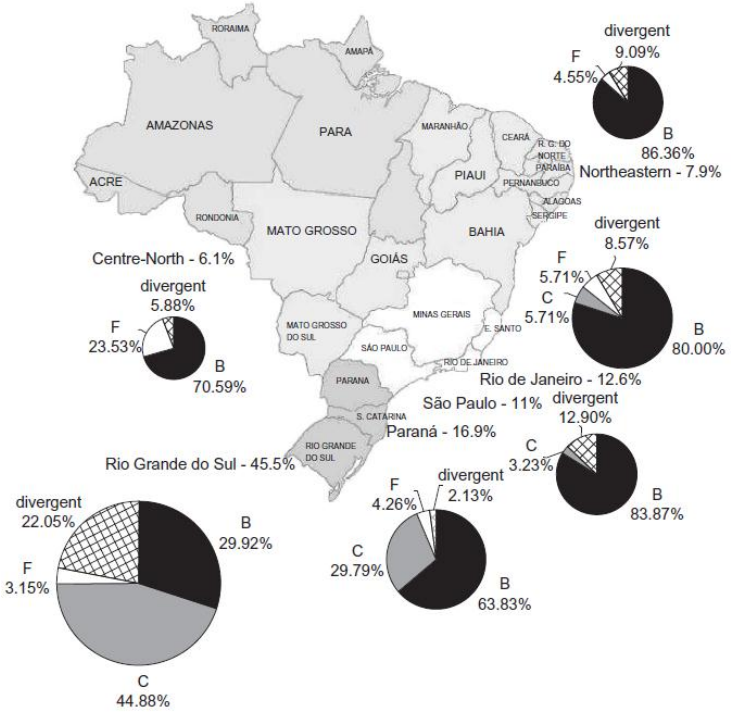

Figura 2.7 Mapa do Brasil mostrando a distribuição dos subtipos. Baseado na análise de PR e RT. As amostras que apresentam discordância entre PR e RT foram consideradas como genomas divergentes. As áreas dos gráficos são proporcionais a quantidade de amostras de cada local analisado (BRINDEIRO *et al.,* 2003).

2.2.3 Ciclo de Replicação e Terapia Antiretroviral

Os vírus normalmente são combatidos pelas células de defesa do próprio organismo infectado. Durante uma infecção viral algumas células têm seu funcionamento prejudicado, e quando o hospedeiro parasitado percebe que está sendo atacado, reage aumentando a quantidade de anticorpos produzidos. Por conta disto, os tratamentos antivirais consistem em amenizar os sintomas apresentados de modo a dar maior conforto ao paciente, que aguarda até que seus anticorpos atuem efetivamente e consigam eliminar o vírus. Porém, conforme exposto anteriormente, no caso do HIV, são as próprias células de defesa que sofrem os ataques, sendo este o motivo da dificuldade em combater a infecção. Consequentemente foi preciso buscar novas estratégias de combate ao vírus.

O ciclo reprodutivo do HIV-1 apresenta diversas etapas, que uma vez interrompidas afetam toda a sua replicação. (PEÇANHA, 2002) Atualmente os tratamentos anti HIV atuam inibindo as principais enzimas virais e também a fusão do vírus com o linfócito T. Desta forma, retarda-se o desenvolvimento da deficiência imunológica, a imunidade do paciente pode ser restabelecida, e sua qualidade de vida melhorada. (SOUZA e ALMEIDA, 2003).

Inicialmente, o vírus é levado até uma célula compatível com seus sítios de ligação e lá, se fundirá à membrana celular da célula hospedeira (CHAN e KIM, 1998). Os inibidores de fusão (IF) e de acoplamento buscam impedir que o vírus penetre nos linfócitos ou monócitos, não permitindo assim que uma nova célula seja infectada (DOMS, 2004). O primeiro inibidor de fusão e o de acoplamento datam respectivamente de 2003 e de 2007. Embora sejam medicamentos novos a expectativa é de que em pouco tempo novas drogas com ação semelhante sejam reconhecidas. (LABORATÓRIO VIRCO, 2011).

Concluída a fusão, o RNA e as enzimas virais entram na célula e logo em seguida a transcriptase reversa do vírus sintetiza o DNA viral a partir do RNA do vírus. Esta enzima é constituída por uma cadeia de 560 códons e uma segunda cadeia de resíduos iniciais da *p66* (SEVIN *et al.*, 2000). Pela alta relevância dessa etapa, ela foi o primeiro alvo no desenvolvimento da terapia ARV, que ainda hoje é composta principalmente de inibidores de transcriptase reversa (POCH *et al.*, 1989).

Após a síntese do DNA viral, a integrase atua unindo-o ao DNA celular, formando um pró-vírus. Com esse DNA integrado, a célula dá inicio ao processo de retrotranscripção e passa a produzir RNA e proteínas virais. Os estudos de ADESOKAN *et al*. (2004) e de CRAIGIE (2001) já mencionavam a utilização de drogas capazes de inibir esta enzima, porém somente em 2007, o primeiro inibidor de integrase foi aprovado. (LABORATÓRIO VIRCO, 2011).

Na fase seguinte, as partículas virais passam por um processo de maturação. Nele, a protease atua quebrando as poliproteínas virais *gag* e *gag-pol,* habilitando-as a juntar o RNA em novas partículas, formando os virions (CAO *et al*., 2005 e GONDA *et al*., 1986). Esta enzima pode ser definida por uma sequência de 99 aminoácidos (WLODAWER e GUSTCHINA, 2000). Os inibidores da protease (IPs) se ligam a protease impedindo que ela quebre as poliproteínas em proteínas menores. Com isso, as partículas virais produzidas não são infecciosas (LABORATÓRIO VIRCO, 2011).

O processamento das poliproteínas encerra a replicação do HIV. Nesta etapa os vírions amadurecem e tornam-se capazes de infectar uma nova célula.(CHAN e KIM, 1998).

2.2.4 Resistência aos Antiretrovirais

Além da dificuldade natural de se combater um vírus que afeta o próprio sistema de defesa, o grande número de mutações do HIV limita bastante a terapia da AIDS (SHAFER, 2002). Conforme será visto a seguir isto é encarado tanto como causa, quanto como consequência da replicação do vírus na presença das drogas antiretrovirais.

Quando um paciente é submetido ao tratamento com medicamentos antiretrovirais embora grande parte da população viral seja eliminada, uma porcentagem dessa população consegue sobreviver. O uso contínuo de drogas antiretrovirais possibilita uma seleção natural das amostras resistentes, tornando-as predominantes. Desta forma o surgimento de resistências aos antiretrovirais pode ser considerado uma consequência da terapia (SHAFER, 2002).

Por outro lado, em seus estudos, RICHMAN *et al*. (2003) e PETROPOULOS (2000) relatam que a replicação do vírus em esquemas terapêuticos pouco eficientes podem causar um acúmulo de mutações de resistência. Desta forma a suscetibilidade às drogas é progressivamente reduzida. Nesta situação, a terapia é vista como sendo a causa do surgimento de mutações de resistência.

SHAFER (2002) também relata que a troca de aminoácidos em determinadas posições da cadeia peptídica, tanto na protease viral, quando na transcriptase reversa, estão associadas à resistência às drogas ARV. JOHNSON (2008) complementa essa informação exibindo as mutações já conhecidas que estão associadas á resistência aos medicamentos já pesquisados. Na tabela 2.1 podemos observar tais posições de mutação, que conforme será exposto no Capítulo 3, serviram de base para o primeiro experimento realizado neste trabalho.

Em consequência do surgimento das resistências aos fármacos que combatem o HIV, começou-se a estudar modos de identificação da existência de mutação que resultasse em falha terapêutica. Conforme exposto no Capítulo 1, nos últimos anos têm sido desenvolvidas metodologias que permitem avaliar fenotipicamente a susceptibilidade/resistência do HIV-1 a partir do perfil mutacional dos dados de genotipagem.

Tabela 2.1. A primeira letra refere-se ao aminoácido encontrado no tipo selvagem do vírus, em seguida está o número da posição onde ocorre a mutação e logo após, o aminoácido de substituição.

Medicamento	Tipo	Mutações de Resistência
Efavirenz	Inibidor de RT	L100I, K103N, V106M, V108I, Y181C/I, Y188L, G190S/A, P225H
Etravirine	Inibidor de RT	V90I, A98G, L100I, K101E/P, V106I, V179D/F/T, Y181C/I/V, G190S/A
Nevirapine	Inibidor de RT	L100I, K103N, V106A/M, V108I, Y181C/I, Y188C/L/H, G190A
Abacavir	Inibidor de RT	K65R, L74V, Y115F, M184V
Didadosine	Inibidor de RT	K65R, L74V,
Emtricitabine	Inibidor de RT	K65R, M184V/I
Lamivudine	Inibidor de RT	K65R, M184V
Stavudine	Inibidor de RT	M41L, D67N, K70R, L210W, T215Y/F, K219Q/E
Tenofovir	Inibidor de RT	K65R, K70E
Zidovudine	Inibidor de RT	M41L, D67N, K70R, L210W, T215Y/F, K219Q/E
Atazanavir/Ritonavir	Inibidor de PT	L10I/F/V/C, G16E, K20R/M/I/T/V, L24I, V32I, L33I/F/V, E34Q, M36I/L/V, M46I/L, G48V, I50L, F53L/Y, I54L/V/M/T/A, A71V/I/T/L, G73C/S/T/A, V82A/T/F/I, I84V, I85V, N88S, L90M, I93M
Darunavir/Ritonavir	Inibidor de PT	V11I, V32I, L33F, I47V, I50V, I54M/L, G73S, L76V, I84V, L89V
Fosamprenavir/Ritonavir	Inibidor de PT	L10I/R/V, V32I, M46I/L, I47V, I50V, I54L/V/M, G73S, L76V, V82A/F/S/T, I84V, L90M
Indinavir/Ritonavir	Inibidor de PT	L10I/R/V, K20M/R, L24I, V32I, M36I, M46I/L, I47V, I54V, A71V/T, G73S/A, L76V, V77I, V82A/F/T, I84V, L90M
Lopivanir/Ritonavir	Inibidor de PT	L10I/R/V, K20M/R, L24I, V32I, L33F, M46I/L, I47V, I50V, F53L, I54V/L/A/M/T/S, L63P, A71V/T, G73S, L76V, V82A/F/T/S, I84V, L90M
Nelfinavir	Inibidor de PT	L10F/I, D30N, M36I, M46I/L, A71V/T, V77I, V82A/F/T/S, I84V, N88D/S, L90M
Saquinavir/Ritonavir	Inibidor de PT	L10I/R/V, L24I, G48V, I54V/L, I62V, A71V/T, G73S, V77I, V82A/F/T/S, I84V, L90M
Tipranavir/Ritonavir	Inibidor de PT	L10V, I13V, K20M/R, L33F, E35G, M36I, K43T, M46L, I47V, I54A/M/V, Q58E, H69K, T74P, V82L/T, N83D, I84V, L90M
Enfuvirtide	Inibidor de Fusão	G36D/S, I37V, V38A/M/E, Q39R, Q40H, N42T, N43D
Raltegravir	Inibidor de Integrase	Q148H/K/R, N155H

Segundo BAXTER (2000), os testes de genotipagem são mais utilizados que os de fenotipagem, pois, tem uma maior disponibilidade, menor custo, e tempo inferior de duração. Além disso, estas análises são motivadas também por estudos que

demonstram que o acesso aos dados genotípicos são uma importante ferramenta para os médicos. No que pese todo o descrito, BAXTER (2000) ressalva que a existência de resistência cruzada pode dificultar a interpretação dos resultados. O que já não pode ser considerado uma tarefa simples pelo alto grau de similaridade entre os subtipos do HIV.

2.3 Redes Neurais

As redes neurais artificiais são ferramentas de Inteligência Artificial que possuem a capacidade de se adaptar e de aprender a realizar certa tarefa ou comportamento a partir de um conjunto de exemplos dados (OSÓRIO e BITTENCOURT, 2000). Os estudos sobre as redes neurais tiveram inicio na década de 40. Warren Mc Culloch, psiquiatra e neuroanatomista, e Walter Pitts, matemático, sugeriram a construção de uma máquina inspirada e baseada no cérebro humano. Esta máquina foi denominada *Psychon*, e partia de um modelo matemático (artificial) do neurônio biológico.

Importantes avanços ocorreram na década seguinte, com a criação do *Snark***,** por Marvin Minsky, em 1951 e com a criação do *Mark I Perceptron***,** por Frank Rosenblatt, Charles Wightman *et al.* em 1958. Entretanto, em 1969 os modelos baseados no *Perceptron* receberam uma forte crítica feita por Minsky e Papert através de sua obra *"Perceptrons: An Introduction to Computational Geometry",* onde estes mostravam matematicamente as limitações da rede de um único nível (OSÓRIO e BITTENCOURT, 2000, FREIMAN e PAMPLONA 2005).

Em consequência desta crítica, os estudos com redes neurais só voltaram a ganhar força na década de 80, quando foram impulsionados pelo algoritmo de retro propagação (*backpropagation*), pelo modelo de Hopfield, pela máquina de Boltzmann e pelos grandes avanços tecnológicos que tornaram os computadores mais velozes. Tais avanços permitiram realizar melhores simulações das redes neurais, bem como, resolver os problemas apontados por Minsky e Papert (OSÓRIO e BITTENCOURT, 2000, FREIMAN e PAMPLONA 2005).

Conforme dito anteriormente, tais redes foram inspiradas no sistema nervoso humano e podem ser separadas em duas categorias: redes neurais clássicas, ou com peso, e redes neurais sem peso, também conhecidas como redes baseadas em memória RAM. Neste primeiro momento serão explicadas as redes com peso, para em

seguida se detalhar as da outra espécie, que foram as efetivamente utilizadas neste trabalho.

2.3.1 Redes Neurais Com Peso

As redes neurais com peso são compostas por neurônios artificiais que se conectam de modo a formar uma teia entre si. Cumpre esclarecer que neurônios artificiais são autômatos simples, capazes de receber uma informação, efetuar um processamento e repassar os dados captados aos outros neurônios ligados a ele. A função de ativação da rede determina o processamento e comportamento destas estruturas, que são exemplificadas na figura 2.82 (OSÓRIO e BITTENCOURT, 2000).

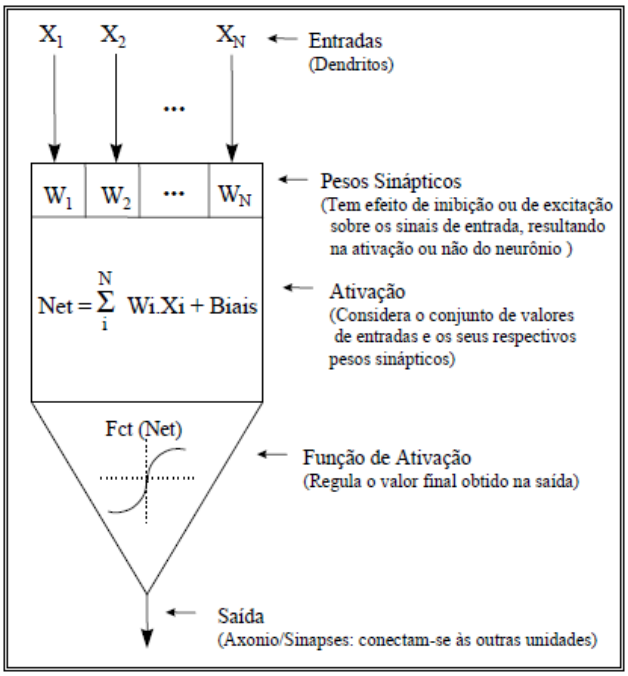

Figura 2.8 Exemplo de neurônio artificial (OSÓRIO e BITTENCOURT, 2000).

Enquanto a função de ativação tem efeito local, a estrutura da rede neural e os pesos entre os neurônios determinam o comportamento global da rede. Cada ligação entre dois neurônios possui um peso, que influenciará na ativação ou não do neurônio seguinte, afetando assim, o fluxo da informação através da rede neural. Portanto, o conhecimento aprendido pela rede, fica armazenado nas ligações entre os neurônios, que podem ser organizados de diversas maneiras diferentes, de acordo com o modelo

que melhor se adequar ao problema. A figura 2.9 exemplifica diferentes arquiteturas de redes.

De maneira geral, as camadas da rede podem ser separadas em três tipos: a camada de entrada, onde os dados são apresentados, as camadas intermediárias, onde é feita a maior parte do processamento, e a camada de saída, onde o resultado é apresentado. Como se vê na figura 2.9, a organização dos neurônios na rede pode ser feita de inúmeras formas diferentes. Podem por exemplo existir uma ou mais camadas intermediárias, como no caso das redes com múltiplas camadas, além disso, os neurônios de uma mesma camada podem estar ligados ou não entre si (OSÓRIO e BITTENCOURT, 2000).

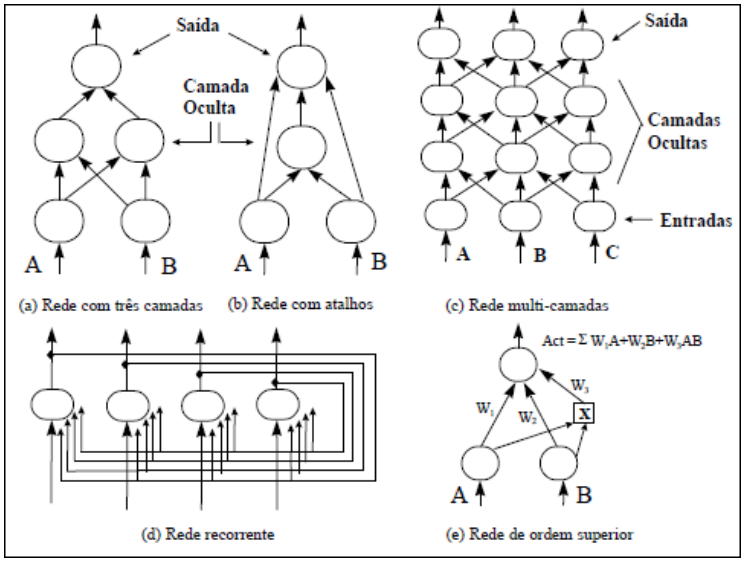

Figura 2.9 Exemplos de arquiteturas de redes neurais artificiais (OSÓRIO e BITTENCOURT, 2000).

A rede começa com todos os pesos atribuídos de maneira aleatória, necessitando assim, passar por um processo de aprendizagem. Nesta etapa, os pesos serão alterados de acordo com o algoritmo de cognição da rede, até que esta se estabilize. Depois de concluído o treinamento, os pesos das ligações inter-neuronais guardarão as informações aprendidas, e a rede estará pronta para a etapa de reconhecimento (OSÓRIO e BITTENCOURT, 2000).

2.3.2 Redes Neurais Sem Peso

As redes neurais sem peso eram originalmente conhecidas como "*redes N-tuplas*", cujos métodos foram desenvolvidos por Bledsoe e Browning, em 1959. Durante a década de 60, estas redes tiveram seus estudos interrompidos, mas voltaram a ser pesquisadas no inicio da década de 70 por Igor Aleksander, que juntamente Wilkes e Stonham, desenvolveu a WiSARD (Wilkes,Stonham, Aleksander Recognition Device), rede neural amplamente estudada até os dias de hoje (AUSTIN, 1998).

Embora ainda sejam inspiradas no sistema nervoso humano, essas redes seguem uma abordagem diferente das redes com peso. Nelas prioriza-se a emulação da topologia das conexões entre dendritos e axônios, ou seja, a árvore dendrítica (GRIECO, LIMA, GREGORIO, *et al.*, 2009). Além disso, seus neurônios artificiais não executam processamento nem possuem ligações com pesos sinápticos, de modo a guardar informação nas suas conexões, e por isso passaram a ser conhecidas como redes sem peso.

A informação aprendida é armazenada em memórias RAM, que funcionam como os neurônios artificiais. Assim, estas redes são conhecidas também como redes neurais baseadas em memórias RAM. Além disso, do mesmo modo como poderia ser modificada a arquitetura dos neurônios nas redes com peso, foram desenvolvidas diferentes formas de utilizar as memórias RAM. Tais formas impactam não só no tempo de resposta, como também no grau de generalização e especificidade. As principais redes sem peso existentes hoje são: WiSARD, G-RAM (Generalization RAM), VG-RAM (Virtual Generalization RAM) e GSN (Goal Seeking Neuron).

2.3.3 WiSARD – (Wilkes, Stonham, Aleksander Recognition Device)

A WiSARD foi a primeira rede neural artificial a ser patenteada e produzida comercialmente, sendo também o modelo de rede neural sem peso (WNN) mais representativo (GRIECO, LIMA, GREGORIO, *et al.*, 2010). Originalmente foi toda implementada em hardware, por Bruce Wilkie, mas posteriormente seu algoritmo foi reproduzido na forma de um programa em C++ que rodava em UNIX. Sendo assim, um laptop pôde ser transformado em uma rede neural com considerável poder cognitivo e flexibilidade, o que levou a um nível de rápida prototipagem do sistema anteriormente inconcebível ao se usar somente hardware (ALEKSANDER, 1998).

Nesta rede a entrada é separada em tuplas. Cada uma dessas tuplas será relacionada com uma memória RAM, e conforme mencionado anteriormente, o conhecimento estará armazenado nessas memórias. Além disso, cada tupla da entrada será utilizada para endereçar uma posição da memória e acessar assim o conhecimento armazenado nesta RAM. Pelo fato das tuplas corresponderem a um endereço de RAM, a entrada deve possuir valores binários ou ter passado por algum pré-processamento que converta o valor original em bits (ALEKSANDER, 1998).

O treinamento de um neurônio é feito mudando-se o conteúdo da posição de memória endereçada pela tupla, que é iniciado com "0", para "1". Desta forma, cada neurônio é extremamente eficiente em reconhecer um padrão apresentado anteriormente, mas não é capaz de generalizar a informação de modo a reconhecer padrões semelhantes. Porém, conforme mencionado, cada tupla representa somente parte da entrada, de modo que um padrão apresentado será dividido em M partes. Dentro da rede, cada categoria é armazenada em um discriminador, que é composto por M RAM's e estes conjuntos de memórias serão responsáveis pela generalização da rede e tolerância a ruídos (FRANÇA, SILVA, LENGERKE, *et al.,* 2009).

Importante ressaltar que uma tupla com N bits endereçará 2^N posições. Portanto o número de bits da tupla dependerá do tamanho de memórias que está sendo utilizada na rede. Considerando que a entrada possui M x N bits, será preciso M RAM's para cada discriminador utilizado.

> *"The advantages of this are that a high degree of discrimination can be obtained with a high n and that a training set for one class can contain a variety of differing patterns giving correct recognition if the unknown is nearer to any element of that set than of any other set. The disadvantages are that the total memory cost of the system are* $Dk2^n$ *(where D is the number of discriminators) introducing a penalty for high levels of discrimination both the discrimination and the cost being exponential functions of n."*
>
> *(Igor Aleksander, 1998)*

Na figura 2.10 pode-se observar uma rede que utiliza memórias de 8 bits (2^3) e necessita portanto de quatro memórias por discriminador para trabalhar com uma entrada de 12 bits. Neste exemplo está sendo armazenado um padrão apresentado para a categoria "F".

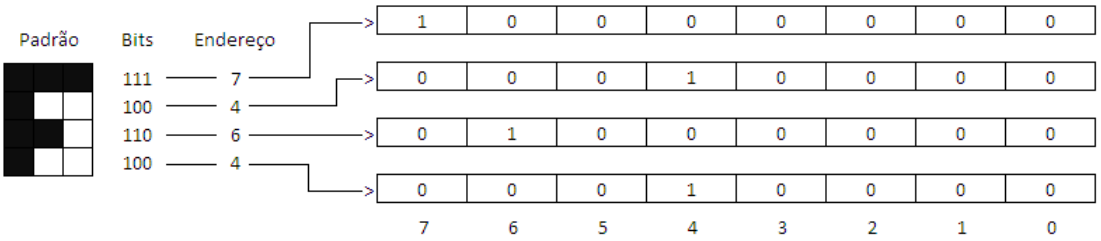

Figura 2.10 Discriminador da categoria "F" de uma WiSARD com memórias de 8 bits e entrada de 12 bits.

Nesta rede, o reconhecimento é feito de maneira muito similar ao treinamento. Porém, a entrada é apresentada a todos os discriminadores, cada RAM lê o endereço de memória apontado pela tupla correspondente e retorna o valor armazenado, que pode ser 0 ou 1. Feito isso, cada discriminador soma o número de memórias que retornou 1 e chega ao seu total de pontos. Em seguida, a categoria do discriminador que tiver obtido a maior pontuação é apresentada como resposta àquele reconhecimento (ALEKSANDER, 1998).

Embora esta seja uma maneira muito eficiente e com bons resultados, possui tal método um aspecto negativo muito relevante: caso ocorra empate, uma das categorias que empataram será escolhida de maneira aleatória. A figura 2.11 exemplifica o processo de reconhecimento ocorrido dentro de um discriminador com N RAM, que neste caso obteve r pontos. Pode-se observar que neste exemplo, as posições da entrada que compõem as tuplas de cada RAM foram agrupadas de maneira completamente aleatória.

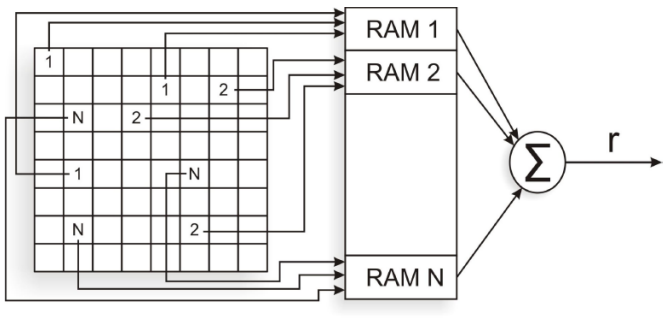

Figura 2.11 Cálculo de pontos de um discriminador durante processo de reconhecimento (França, Silva, Lengerke, *et al.*, 2009).

Segundo ALEKSANDER (1998), além do cálculo de pontos dos discriminadores, este sistema fornece um nível de confiança relativa ao resultado do reconhecimento realizado. Este parâmetro é calculado pela fórmula: $C = (R_{MAX} –$

R_{2MAX}) /R_{MAX} onde R_{MAX} é a maior pontuação encontrada para os discriminadores e R_{2MAX} a segunda maior pontuação. Em caso de empate, como os dois discriminadores obtiveram a mesma pontuação, a confiança será igual a zero, o que pode ser usado como um alerta a possíveis categorizações errôneas. A figura 2.12 exibe o resultado do reconhecimento em uma WiSARD com múltiplos discriminadores. Neste exemplo R_1 obteve a maior pontuação e R_N a segunda maior.

É fácil perceber que tanto o treinamento quanto o reconhecimento não depende do número de padrões aprendidos pela rede. Eles podem ser executados em tempo constante, pois dependem apenas do tamanho da entrada e no caso do reconhecimento, do número de categorias. Esta é a grande vantagem dessa rede, pois permite que ambas as etapas sejam realizadas quase instantaneamente. Além disso, o grau de generalização da rede pode ser facilmente controlado, dependendo da relação entre o tamanho da memória e o número de memórias do discriminador. Quando maior a memória, menor a capacidade de generalização da rede (França, Silva, Lengerke, *et al.*, 2009).

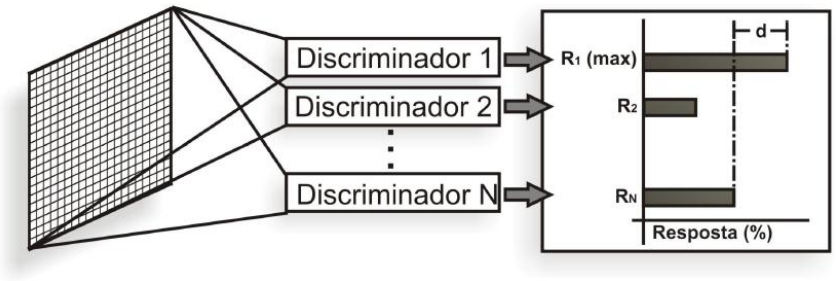

Figura 2.12 Reconhecimento com N discriminadores (França, Silva, Lengerke, et al., 2009).

2.3.4 VG-RAM – (Virtual Generalizing RAM)

A VG-RAM recebe este nome por ter como objetivo tornar virtual o uso das memórias RAM. A seguir será mostrado que a forma como as memórias são alocadas e utilizadas nesta rede é bem diferente da WiSARD. Porém, segundo ALEKSANDER (1998), o desempenho da VG-RAM como um reconhecedor de padrões é idêntico ao algoritmo da WiSARD.

O treinamento é feito armazenando-se nas RAM's o padrão apresentado e a categoria associada a ele. As memórias formam uma tabela que relaciona estas duas informações. Cada padrão treinado funciona de maneira equivalente a um discriminador da WiSARD e para cada novo treinamento, é alocada uma nova

memória, sendo que as já existentes não são alteradas. Desta forma, a quantidade de memórias alocadas cresce ao longo do processo de treinamento e de acordo com a necessidade. Assim a VG-RAM dispensa a prévia alocação de espaço necessária na WiSARD (ALEKSANDER, 1998).

ALEXSANDER (1998) afirma também que o custo de memórias do sistema passa a ser TKN, onde T é o número de treinamentos feitos, K continua sendo o número de memórias usadas e N o tamanho das memórias. Enquanto na WiSARD esse custo variava exponencialmente em relação a N, na VG-RAM ele varia de maneira linear. Porém, o ajuste dos discriminadores em relação a N continua sendo exponencial. Além disso, diferente do que ocorre na WiSARD, não precisa ser feito qualquer tipo de mapeamento dos bits da entrada em endereços de memórias.

Para realizar o reconhecimento, o padrão apresentado é comparado com todos os padrões armazenados na rede, sendo calculada a distância de Hamming para todos eles. A resposta ao reconhecimento será a categoria associada ao padrão que possuir a menor distância. Assim como ocorre na WiSARD, caso ocorra empate, a escolha é feita de maneira aleatória entre as categorias empatadas (ALEKSANDER, 1998).

Embora o custo de alocação tenha passado a crescer linearmente, ALEKSANDER (1998) alerta para uma desvantagem da VG-RAM em relação à WiSARD. Agora, o reconhecimento não depende do número de discriminadores, e sim do número total de treinamentos realizados. Consequentemente, ele tende a ser T vezes mais lento que o reconhecimento realizado pela WiSARD.

Para exemplificar o funcionamento da VG-RAM pode-se analisar o caso em que se apresentam à rede os padrões ‘110110’ e ‘011011’, ambos da categoria ‘A’ e o padrão ‘100100’ da categoria B. Neste exemplo, a VG-RAM precisará de memórias de 6 bits, e depois de realizados esses treinamento, estará utilizando três memórias, cada uma associada ao respectivo padrão.

Ao tentar reconhecer o padrão ‘101101’, será preciso calcular a distância de Hamming para os três padrões armazenados. Ambos os padrões da categoria ‘A’ terão 4 bits diferentes, enquanto que o padrão da categoria ‘B’ será diferente em apenas 2 bits, e por isso, será a categoria dada como resposta a esse reconhecimento.

2.3.5 *Bleaching*

Na descrição da WiSARD, foi mencionado que uma de suas principais desvantagens é o fato de realizar escolha aleatória entre os discriminadores de maior pontuação quando ocorre empate. A técnica do *Bleaching* foi elaborada de modo a minimizar esse problema, buscando uma resposta determinística para os casos de empate, aperfeiçoando-se assim esta rede neural. Tal método consiste em armazenar nas memórias da WiSARD o número de vezes que o padrão foi apresentado, para que na etapa de reconhecimento, essa informação possa ser usada como critério de desempate. Com isso, ao invés de serem armazenados apenas valores booleanos (0's e 1's) dentro das memórias, estas possuirão números inteiros.

Na fase de treinamento a alteração é muito simples. Cada vez que aquele padrão for apresentado, ele irá incrementar em uma unidade o valor que estiver guardado naquela posição de memória. Isto praticamente não altera, nem a complexidade, nem o tempo de resposta dessa etapa.

Já no reconhecimento, existirá agora, um valor de *Bleaching*. Este valor será usado para verificar se a memória pontuou ou não. Enquanto anteriormente, as memórias pontuariam caso possuíssem valor 1 gravado, com o uso desta técnica, passa-se a pontuar somente aquelas que tiverem valor armazenado maior que o valor do *Bleaching* que está sendo usado.

O *Bleaching* inicia com zero, de modo que da primeira vez em que as memórias são verificadas a pontuação encontrada seja exatamente igual ao caso onde esta técnica não era usada. Caso duas ou mais categorias tenham empatado, o valor do *Bleaching* é acrescido de uma unidade, e os pontos das memórias serão novamente calculados. Este procedimento deve ser repetido até que um discriminante seja eleito, ou até que todos eles parem de pontuar. Somente neste segundo caso é que a escolha deve ser feita de modo aleatório. O único ponto negativo dessa técnica é a possibilidade do aumento no tempo de resposta do reconhecimento.

No Capítulo 4 poderá ser mais bem observado que em nenhum dos experimentos realizados neste trabalho, duas ou mais categorias ficaram empatas com o uso do *Bleaching*. Ao analisar esses experimentos, pode-se ver com clareza como esta técnica foi fundamental para a melhoria do reconhecimento em praticamente todos os experimentos realizados. Percebe-se ainda, que embora o número de etapas do reconhecimento possa aumentar bastante, o tempo de resposta não aumentou de maneira proporcional, pois algumas estratégias de aperfeiçoamento puderam ser utilizadas. Tais estratégias serão mais bem detalhadas no tópico 4.3.

3 Metodologia

Neste Capítulo será descrito de que forma o problema de reconhecimento de padrão do HIV-1 foi tratado com a utilização da WiSARD. Será abordada também a base de dados utilizada e os procedimentos realizados para converter tais informações em estruturas que pudessem ser apresentadas à rede. Por fim, serão explicados os experimentos realizados, o que os motivou e suas características.

3.1 Base Teórica Aplicada ao Problema

No Capítulo 2 foi abordada, de forma descritiva, a estrutura do HIV-1, seu processo de replicação e as estratégias utilizadas para combatê-lo. Da descrição das terapias antiretrovirais, se extraiu que os medicamentos utilizados atuam inibindo as enzimas do vírus ou inibindo a fusão do vírus com o linfócito T. Mencionou-se também que quando o vírus sofre mutação, pode haver resistência ao medicamento utilizado. Porém já existem pesquisas relatando que uma forma de aumentar a eficiência dos tratamentos é realizar o teste genotípico do vírus do paciente infectado, de modo a melhor direcionar seu tratamento.

Tal fato levou a necessidade de reconhecimento e categorização genética do HIV-1. No Capítulo 1, resta claro que já existem diversos estudos que buscam relacionar o sequenciamento genético de uma enzima viral com o medicamento para o qual ela deve apresentar resistência. Neles foram usadas a transcriptase reversa ou a protease do vírus, onde ambas as enzimas apresentaram bons resultados.

Diversos algoritmos de categorização já foram utilizados, inclusive as redes neurais com peso, porém não foi encontrado nenhum experimento que utilizasse a WiSARD. Portanto, é interessante validar sua aplicabilidade em relação ao problema. Além disso, certamente esta rede conseguirá resultados muito bons, com grande estabilidade e com pouco tempo de processamento.

Inicialmente busca-se confirmar que esta metodologia será capaz de trabalhar com dados dessas proporções de forma eficiente. Dentre as enzimas listadas anteriormente, a protease é a que possui menor número de aminoácidos, o que consequentemente, acarretaria em um resultado mais rápido. Sendo assim, optou-se por utilizar esta informação acerca do HIV.

Caso os resultados obtidos por esta representação não fossem suficientemente bons, mas a rede demonstrasse ser capaz de lidar eficientemente com dados dessas proporções, poder-se-ia substituir a protease pela transcriptase reversa. Uma vez que, conforme foi visto no Capítulo 2, esta possui uma cadeia de aminoácidos mais longa, sendo inclusive biologicamente mais complexa, certamente apresentará resultados mais precisos.

Ao descrever a rede neural, foi mencionado que ela armazena informações em memórias RAM de acordo com o endereço dessas memórias, e portanto, trabalha basicamente com dados em binário. Sendo assim, torna-se imprescindível converter a protease do HIV em uma estrutura binária.

SILVA (2009) codificou a protease de maneira a obter uma sequência de bits. Em sua forma de codificação, ordenou os aminoácidos de acordo com a escala de hidrofobicidade e atribuiu para cada um deles um número binário de 20 bits. Desta forma, a protease que anteriormente possuía 99 aminoácidos, passou a ser tratada como 99 sequências de 20 bits, tendo um total de 1980 bits. Neste trabalho será usada esta mesma codificação, porém, outras formas de representar os aminoácidos foram elaboradas. Tais codificações serão mais bem detalhadas no tópico a seguir.

Além da necessidade dessa conversão para binário, foi exposto no Capítulo 2 que na WiSARD a entrada dos dados é matricial. Conforme mencionado anteriormente, após o embaralhamento das posições, cada linha se relaciona diretamente com uma memória de cada discriminador. Inicialmente, escolheu-se utilizar memórias de 8 bits, e por isso, a entrada passou a ser dividida em blocos de tamanho 8. Mas, nos casos em que o tamanho da entrada não é completamente divisível por 8 foi necessário complementar essa representação.

3.2 Representações

A codificação usada por SILVA (2009) mencionada anteriormente, consiste em representar cada aminoácido por um valor em binário de tamanho 20, de modo que possua apenas um bit igual a 1 e o restante seja 0. Na base decimal significa dizer que serão utilizadas as potências de 2,variando de 2^0 até 2^{19} (1, 2, 4, 8, 16, ..., 524288), onde cada aminoácido estaria representado por um desses valores. Para isso, foi usado o grau de hidrofobicidade, ordenando os aminoácidos de modo que o menos hidrofóbico estaria ligado ao menor número, no caso 2^0, e o mais hidrofóbico seria

representado pelo maior número, ou seja, 2^{19}. Tal codificação pode ser observada na tabela 3.1.

Tabela 3.1 Codificação BIN 20.

Aminoácido	Símbolo	Valor em Binário
Isoleucina	I	10000000000000000000
Valina	V	01000000000000000000
Leucina	L	00100000000000000000
Fenilalanina	F	00010000000000000000
Cisteina	C	00001000000000000000
Metionina	M	00000100000000000000
Alanina	A	00000010000000000000
Glicina	G	00000001000000000000
Treonina	T	00000000100000000000
Serina	S	00000000010000000000
Triptofano	W	00000000001000000000
Tirosina	Y	00000000000100000000
Prolina	P	00000000000010000000
Istidina	H	00000000000001000000
Glutammina	Q	00000000000000100000
Asparagina	N	00000000000000010000
Acido glutammico	E	00000000000000001000
Acido aspartico	D	00000000000000000100
Lisina	K	00000000000000000010
Arginina	R	00000000000000000001

Esta codificação, embora seja a mais simples, não traz fortes relações com as características químicas dos aminoácidos. Considerando-se a distância de Hamming, todos eles são equidistantes e diferem entre si em apenas 2 bits. Fora isso, ao trabalhar com uma rede neural é importante que dados muito semelhantes sejam representados por valores próximos, enquanto que dados muito diferentes sejam representados por valores distantes. Buscaram-se então novas codificações que atendessem melhor a esse critério.

Embora a codificação Bin 20 seja baseada na escala de hidrofobicidade, ela apenas alinha os aminoácidos em sequência, espalhando-os uniformemente. Assim, mostra-se que no caso da WiSARD seria mais interessante, arrumá-los de modo não uniforme, através do uso dos valores absolutos desta escala com o fim de se chegar a uma nova codificação. A tabela 3.2 exibe os valores de hidrofobicidade da escala KYTE e DOOLITTLE (1982), também denominada escala KD.

Tabela 3.2 Escala Kyte e Doolittle (1982) – KD.

Aminoácido	Símbolo	Escala KD	Categoria
Isoleucina	I	4,5	Hidrofóbico
Valina	V	4,2	Hidrofóbico
Leucina	L	3,8	Hidrofóbico
Fenilalanina	F	2,8	Hidrofóbico
Cisteina	C	2,5	Hidrofóbico
Metionina	M	1,9	Hidrofóbico
Alanina	A	1,8	Hidrofóbico
Glicina	G	-0,4	Neutro
Treonina	T	-0,7	Neutro
Serina	S	-0,8	Neutro
Triptofano	W	-0,9	Neutro
Tirosina	Y	-1,3	Neutro
Prolina	P	-1,6	Neutro
Istidina	H	-3,2	Hidrofílico
Glutammina	Q	-3,5	Hidrofílico
Asparagina	N	-3,5	Hidrofílico
Acido glutammico	E	-3,5	Hidrofílico
Acido aspartico	D	-3,5	Hidrofílico
Lisina	K	-3,9	Hidrofílico
Arginina	R	-4,5	Hidrofílico

Na tabela 3.2 pode-se observar que os valores variam de 4,5 a -4,5. O fato desta escala possuir números com casas decimais e menores que zero não é um grande problema para se criar uma codificação em binário. Eles podem ser facilmente escalonados. Porém, é importante observar que os aminoácidos “Q”,”N”, “E” e “D” possuem o mesmo valor. Isto resultaria em uma mesma codificação e consequentemente impediria a WiSARD de distingui-los. Logo, tal fato poderia prejudicar em muito os experimentos.

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

Equação 3.1 Cálculo das escalas KD normalizadas pela massa molecular.

SILVA (2009) afirma que a ponderação da escala de Kyte e Doolitte se fez necessária para evitar que o método não incorporasse aminoácidos com valores de hidrofobicidade equivalentes, apesar da presença de mutação. Para isso, utiliza-se a tabela 3.3, que é resultado da normalização da escala KD pela massa molecular. Os

valores dessa tabela podem ser obtidos usando-se a equação 3.1, que resolve o problema dos aminoácidos com valores repetidos na medida em que apresenta novas opções de escalas, de acordo com a massa molecular de cada aminoácido. Embora todas essas escalas apresentem valores muito próximos, quando for escalonada para um binário de 20 bits, mesmo as menores diferenças ficarão detectáveis pela rede neural. Além disso, esses valores irão manter o viés biológico, pois continuam fortemente atrelados às propriedades dos aminoácidos.

Tabela 3.3 Escala KD normalizada pela massa molecular (SILVA, 2009).

		I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
		5,895	4,914	4,978	4,620	3,025	2,831	1,602	-0,300	-0,833	-0,840	-1,836	-2,353	-1,840	-4,960	-5,145	-5,110	-4,655	-4,620	-5,694	-7,830
I	5,895	0,000	4,914	0,064	4,556	-1,531	4,362	-2,760	2,460	-3,293	2,453	-4,289	1,936	-3,776	-1,184	-3,961	-1,149	-3,506	-1,114	-4,580	-3,250
V	4,914	0,981	0,000	0,064	-0,294	-1,889	-2,083	-3,312	-5,214	-5,747	-5,754	-6,750	-7,267	-6,754	-9,874	-10,059	-10,024	-9,569	-9,534	-10,608	-12,744
L	4,978	0,917	-0,064	0,000	-0,358	-1,953	-2,147	-3,376	-5,278	-5,811	-5,818	-6,814	-7,331	-6,818	-9,938	-10,123	-10,088	-9,633	-9,598	-10,672	-12,808
F	4,620	1,275	0,294	0,358	0,000	-1,595	-1,789	-3,018	-4,920	-5,453	-5,460	-6,456	-6,973	-6,460	-9,580	-9,765	-9,730	-9,275	-9,240	-10,314	-12,450
C	3,025	2,870	1,889	1,953	1,595	0,000	-0,194	-1,423	-3,325	-3,858	-3,865	-4,861	-5,378	-4,865	-7,985	-8,170	-8,135	-7,680	-7,645	-8,719	-10,855
M	2,831	3,064	2,083	2,147	1,789	0,194	0,000	-1,229	-3,131	-3,664	-3,671	-4,667	-5,184	-4,671	-7,791	-7,976	-7,941	-7,486	-7,451	-8,525	-10,661
A	1,602	4,293	3,312	3,376	3,018	1,423	1,229	0,000	-1,902	-2,435	-2,442	-3,438	-3,955	-3,442	-6,562	-6,747	-6,712	-6,257	-6,222	-7,296	-9,432
G	-0,300	6,195	5,214	5,278	4,920	3,325	3,131	1,902	0,000	-0,533	-0,540	-1,536	-2,053	-1,540	-4,660	-4,845	-4,810	-4,355	-4,320	-5,394	-7,530
T	-0,833	6,728	5,747	5,811	5,453	3,858	3,664	2,435	0,533	0,000	-0,007	-1,003	-1,520	-1,007	-4,127	-4,312	-4,277	-3,822	-3,787	-4,861	-6,997
S	-0,840	6,735	5,754	5,818	5,460	3,865	3,671	2,442	0,540	0,007	0,000	-0,996	-1,513	-1,000	-4,120	-4,305	-4,270	-3,815	-3,780	-4,854	-6,990
W	-1,836	7,731	6,750	6,814	6,456	4,861	4,667	3,438	1,536	1,003	0,996	0,000	-0,517	-0,004	-3,124	-3,309	-3,274	-2,819	-2,784	-3,858	-5,994
Y	-2,353	8,248	7,267	7,331	6,973	5,378	5,184	3,955	2,053	1,520	1,513	0,517	0,000	0,513	-2,607	-2,792	-2,757	-2,302	-2,267	-3,341	-5,477
P	-1,840	7,735	6,754	6,818	6,460	4,865	4,671	3,442	1,540	1,007	1,000	0,004	-0,513	0,000	-3,120	-3,305	-3,270	-2,815	-2,780	-3,854	-5,990
H	-4,960	10,855	9,874	9,938	9,580	7,985	7,791	6,562	4,660	4,127	4,120	3,124	2,607	3,120	0,000	-0,185	-0,150	0,305	0,340	-0,734	-2,870
E	-5,145	11,040	10,059	10,123	9,765	8,170	7,976	6,747	4,845	4,312	4,305	3,309	2,792	3,305	0,185	0,000	0,035	0,490	0,525	-0,549	-2,685
Q	-5,110	11,005	10,024	10,088	9,730	8,135	7,941	6,712	4,810	4,277	4,270	3,274	2,757	3,270	0,150	-0,035	0,000	0,455	0,490	-0,584	-2,720
D	-4,655	10,550	9,569	9,633	9,275	7,680	7,486	6,257	4,355	3,822	3,815	2,819	2,302	2,815	-0,305	-0,490	-0,455	0,000	0,035	-1,039	-3,175
N	-4,620	10,515	9,534	9,598	9,240	7,645	7,451	6,222	4,320	3,787	3,780	2,784	2,267	2,780	-0,340	-0,525	-0,490	-0,035	0,000	-1,074	-3,210
K	-5,694	11,589	10,608	10,672	10,314	8,719	8,525	7,296	5,394	4,861	4,854	3,858	3,341	3,854	0,734	0,549	0,584	1,039	1,074	0,000	-2,136
R	-7,830	13,725	12,744	12,808	12,450	10,855	10,661	9,432	7,530	6,997	6,990	5,994	5,477	5,990	2,870	2,685	2,720	3,175	3,210	2,136	0,000

Para a criação da nova escala, representou-se o menor valor por 1 e o maior por $2^{20} – 1$, de modo que os valores extremos fossem representados pelos extremos em binário de 20 bits. Os demais números foram escalonados mantendo-se as proporções das distâncias entre seus valores, não sendo mais distribuídos uniformemente, como era feito na codificação BIN 20.

Além das quantias decorrentes da hidrofobicidade, outras propriedades dos aminoácidos podem ser usadas para representá-los. A própria massa molecular usada na formula 3.1 pode ser um parâmetro interessante. Neste trabalho, optou-se por utilizar uma das escalas KD ponderadas, a massa molecular e uma combinação dessas duas informações, para a geração das codificações. No caso da massa molecular, usou-se o mesmo raciocínio da escala anterior, espalhando-se os valores de 1 até $2^{20} – 1$ e obtendo-se assim, 20 números binário de 20 bits. Porém, para a combinação dessas duas informações, os valores foram espalhados somente de 1 até $2^{10} – 1$, obtendo-se então valores de 10 bits. Posteriormente, foram juntados os 10 bits do valor da massa molecular, com os 10 bits do valor da escala da KD ponderada pela

massa molecular que foi selecionada, de modo a formarem também 20 números de 20 bits. Os valores das massas moleculares podem ser observados na tabela 3.4.

Tabela 3.4 Massa Molecular.

Amminoácido	Símbolo	Massa (g/mol)
Isoleucina	I	131
Valina	V	117
Leucina	L	131
Fenilalanina	F	165
Cisteina	C	121
Metionina	M	149
Alanina	A	89
Glicina	G	75
Treonina	T	119
Serina	S	105
Triptofano	W	204
Tirosina	Y	181
Prolina	P	115
Histidina	H	155
Glutamina	Q	146
Asparagina	N	132
Ácido glutâmico	E	147
Ácido aspártico	D	133
Lisina	K	146
Arginina	R	174

A representação binária desses valores, a principio, obedecia à regra exposta anteriormente, aproximando aminoácidos com propriedades semelhantes e afastando os de comportamento diferente. Porém, ao se analisar a codificação gerada, é de se observar que esta apresentava alguns saltos na sua representação. Tais saltos podem ser observados, por exemplo, entre os números sete e oito, que em binário são representados respectivamente por 0111, 1000. Embora sejam valores próximos, diferem em todos os quatro bits, possuindo grande distância de Hamming.

Por causa desta característica da codificação binária, passa a se usar também codificações com Gray code. Nesse sistema binário os números sempre diferem em um bit do seu antecessor e do seu sucessor. No mesmo exemplo, os números sete e oito são representados por 0100, 1100, diferindo em um único bit. Embora Gray code seja melhor que a codificação binária para representar valores semelhantes, ela possui uma característica que não é adequada para a pretendida representação.

Neste sistema, os extremos também só diferem de um único bit. Os números 0 e 15 por exemplo são representados por 0000 e 1000 respectivamente. Sendo assim, o aminoácido mais hidrofílico teria uma distância de Hamming muito pequena em relação ao mais hidrofóbico.

Para solucionar este problema da escala em Gray code passa-se a escalonar os números somente até dois terços de 2^{20} e de 2^{10}. Nestes valores, o número é representado por todos os bits em 1. Desta forma, o máximo e o mínimo terão a maior distância de Hamming possível. Essa nova forma de escalonamento combina as características positivas tanto do sistema binário, quando do código de Gray, bem como elimina suas respectivas falhas.

Voltando aos exemplos anteriores, ao invés de se espalhar os valores até 15, seriam escalonados os números somente até 10. A representação de 10 usando Gray code é 1111, que possui todos os bits diferentes de 0. No caso do exemplo de números próximos, o número sete e o oito continuariam sendo representados por 0100 e 1100.

Tabela 3.5 Resumo das Codificações

Sigla	Descrição
BIN20	Vetor de 20 bits contendo apenas um bit em 1
HIDROBIN	Hidrofobicidade em binário usando valores de 1 a 2^{20}-1
HIDROGRAY	Hidrofobicidade em Gray code usando valores de 1 a 2^{20}-1
HIDROGRAY23	Hidrofobicidade em Gray code usando valores de 1 a (2^{20}-1)*(2/3)
MMBIN	Massa molecular em binário usando valores de 1 a 220-1
MMGRAY	Massa molecular em Gray code usando valores de 1 a 220-1
MMGRAY23	Massa molecular em Gray code usando valores de 1 a (220-1)*(2/3)
HIDROMMBIN	Massa molecular seguido pelo valor da hidrofobicidade ambos em binário usando valores de 1 a 210-1
HIDROMMGRAY	Massa molecular seguido pelo valor da hidrofobicidade ambos em Gray code usando valores de 1 a 210-1
HIDROMMGRAY23	Massa molecular seguido pelo valor da hidrofobicidade, ambos em Gray code e usando valores de 1 a (210-1)*(2/3)

Por tratar-se o caso em tela apenas de especulação sobre a eficiência das escalas, não é possível o descarte de nenhuma delas, e portando, deve ser usado um total de 10 codificações diferentes nos experimentos aqui realizados. Na tabela 3.5 é possível ver as codificações usadas e as siglas pelas quais estas poderão ser facilmente identificadas nos Capítulos seguintes. Além disso, no ANEXO I há uma tabela com os aminoácidos e os valores binários de cada uma dessas codificações.

3.3 O Banco de Dados

Os dados utilizados nesses experimentos foram cedidos pelo Laboratório de Virologia Molecular da Universidade Federal do Rio de Janeiro (UFRJ/Brasil), integrante da rede de laboratórios de genotipagem do Ministério da Saúde (RENAGENO). Essa base de dados é composta por 1205 sequências do gene da protease provenientes de isolados séricos de pacientes portadores do HIV-1.

Cada amostra possui uma sequência de 99 aminoácidos, que compõem a protease, o subtipo[2] do vírus e a informação acerca de ser esse vírus é proveniente de um paciente que nunca foi submetido à terapia com uso de medicamentos antiretrovirais (*naive*) ou proveniente de um paciente que apresentou falha terapêutica por desenvolver resistência à droga que lhe estava sendo administrada.

A base de dados utilizada, contem amostras de três subtipos do HIV-1: B, C e F. Todos esses subtipos possuíam amostras *naive* e resistentes a medicamentos. Desta forma, serão trabalhados seis grupos. As quantidades das amostras em cada um dos subtipos podem ser vistas abaixo, na tabela 3.6.

Tabela 3.6 Distribuição das Amostras.

Subtipo	Resistente		*Naive*		Total	
	Amostras	%	Amostras	%	Amostras	%
B	747	61,99%	99	8,22%	846	70,21%
C	155	12,86%	49	4,07%	204	16,93%
F	136	11,29%	19	1,58%	155	12,86%
Total	1038	86,14%	167	13,86%	1205	100,00%

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

Conforme pode ser observado na tabela 3.6, as amostras utilizadas estão distribuídas de maneira extremamente desequilibrada. Embora tal desequilíbrio esteja de acordo com a dispersão do vírus encontrada na população brasileira, tal característica da base de dados foi bastante prejudicial ao processo de reconhecimento da rede neural, sendo inclusive necessário o balanceamento desses grupos, de modo que a quantidade em cada um deles ficasse compatível. Essa discrepância entre a quantidade de amostras fica bem visível no gráfico 3.1. Tal procedimento será mais bem detalhado no tópico 3.4, onde será descrito o processo de treinamento.

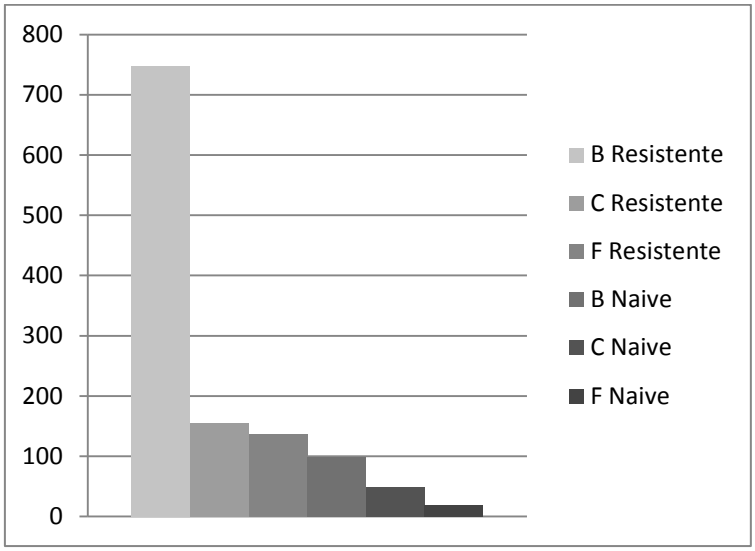

Gráfico 3.1 Distribuição das Amostra.

3.4 Os Experimentos

Inicialmente, os experimentos tinham como objetivo diferenciar as amostras de acordo com os seus grupos presentes na base de dados. Porém, também foram feitos experimentos visando diferenciar apenas o subtipo do vírus, e em um último momento, até mesmo visando diferenciar apenas o grupo B resistente do grupo B *naive*. Em consequência da busca pelo melhor resultado possível, foi necessário realizar experimentos variando: as posições da protease usadas, o tamanho das memórias da rede (8 ou 16 bits) e até mesmo fazer um balanceamento da base de dados. Além disso, em todos os casos, os experimentos foram realizados com todas as formas de codificações descritas no tópico 3.2.

3.4.1 Representação dos Dados

Conforme descrito anteriormente, foram criadas 10 codificações diferentes para representar os dados e cada nova configuração elaborada era experimentada com todas elas. Nos primeiros testes realizados, observou-se que em média o grau de acerto dos resultados variava pouco em relação à codificação usada. Para os três primeiros conjuntos de experimentos a maior diferença encontrada entre as diferentes representações criadas foi de apenas 5,6%.

Além disso, nenhuma era predominantemente melhor, e a cada nova configuração da WiSARD, não havia como prever qual apresentaria um melhor resultado. Nem mesmo era possível garantir que esta variação continuaria pequena. Sendo assim, com a não utilização de alguma codificação, corria-se o risco de acabar não realizando um experimento que porventura viesse a ter o melhor dos resultados. Portanto, optou-se por sempre utilizar todas as codificações criadas.

3.4.2 Posições Selecionadas

No Capítulo 2 foi exposto que na literatura acerca das mutações do HIV-1, já se conhece algumas posições importantes para a determinação da existência de resistente a um dado medicamento. Além disso, algumas posições tidas como “assinatura” do vírus, indicam o subtipo ao qual ele pertence.

Os primeiros experimentos foram realizados usando-se apenas algumas das posições da protease já conhecidas pela literatura. No Capítulo 4 será exposto que ao se observar os valores do *Bleaching,* foi perceptível a existência de uma grande semelhança entre as categorias. Em decorrência deste fato e da busca por resultados melhores realizou-se também uma análise na base de dados.

Tal análise visou descobrir quais posições deveriam ser mais importantes no reconhecimento. Foram descartadas as posições que possuíam taxa de mutação inferior a 0,02% e removeram-se também aquelas onde a mutação ocorria de maneira homogênea para mais de um subtipo, ou seja, onde existia uma alta taxa de mutação, sem que isso determinasse o subtipo da amostra. Na posição 3, por exemplo, quase todas as amostras apresentam mutação de V para I, logo, esta posição não ajudaria na categorização, o que levou ao seu descarte.

Além disso, era fundamental utilizar também todos os 99 aminoácidos da entrada. Portanto, foram realizados experimentos com 22 das posições clássicas de

mutação, com as 27 posições mais significativas (porcentagem de relevância acima de 20%), e com todas as 99 posições da protease.

3.4.3 Configuração da WiSARD:

Dentre os parâmetros da WiSARD descritos no tópico 2.2 foi alterado o tamanho da memória e o número de memórias utilizadas. Neste último, a mudança se deu em decorrência da forma como a entrada foi utilizada. Por ser crível que o embaralhamento da entrada não causaria impactos relevantes no resultado, em todos os experimentos foram feitos de modo randômico.

Num primeiro momento, escolheu-se trabalhar com memórias de 8 bits para programar a rede neural, pois se julgou que este seria o número ideal de equilíbrio entre grau de especificação e de generalização da WiSARD. Posteriormente foram refeitos alguns testes usando memórias de 16 bits, de modo a verificar se havia grandes diferenças no grau de acerto da rede. A análise entre os resultados obtidos nesses casos será descrita no Capítulo 4. Também foram feitos testes usando memórias de 4 e 32 bits, porém não apresentou resultados bons o suficiente para que se entendesse interessante refazer os demais experimentos usando esses tamanhos de memória.

Conforme explicado no Capítulo 2, o número de memórias em cada discriminador da WiSARD é obtido dividindo-se o tamanho da entrada pelo tamanho de cada memória. No tópico anterior, viu-se que em alguns experimentos foram selecionadas posições específicas da entrada para a realização dos reconhecimentos. Em decorrência disto, o tamanho da entrada, e consequentemente o número de memórias da rede, não foi o mesmo para todos os experimentos.

Ao se trabalhar com 22 posições da entrada, e com memórias de 8 bits, chegava-se ao número exato de memórias em cada discriminador, pois em binário a entrada passou a ter 440 bits, o que poderia ser armazenado em 55 memórias de 8 bits. Nos demais casos, foi preciso arredondar o número de memórias para cima e completar o final da entrada com 0’s, para que nenhuma informação da entrada fosse perdida. Toda essa etapa de tradução dos aminoácidos da entrada em sequência binária e o acréscimo de 0’s, quando necessário, foram feitos antes do embaralhamento.

3.4.4 Balanceamento dos Dados:

No tópico 3.3 descreve-se a base de dados utilizada. Na tabela 3.6 é de se observar que a grande maioria das amostras, aproximadamente 62%, pertence ao grupo de subtipo B e é resistente a algum medicamento. Em contrapartida, há também um grupo com cerca de apenas 1,5% das amostras, no caso, o subtipo F *naive*.

Num primeiro momento, os experimentos foram feitos com a base de dados mantendo esta proporção, sem qualquer preocupação com a necessidade de se equilibrar a quantidade de amostras dos grupos. Porém, depois de análise dos primeiros resultados, que serão mais bem detalhados no próximo Capítulo, viu-se que era muito importante balancear os dados durante o processo de aprendizado da rede. Tal fato foi constatado, pois a rede estava conseguindo reconhecer com um grau de acerto muito alto as amostras do grupo predominante, mas praticamente não reconhecia as amostras dos demais grupos. Apesar dos bons resultados, não foi satisfatório ao estudo uma rede capaz de reconhecer somente um dos seis grupos treinados. Caso fosse mudada a base de dados a ser reconhecida, passando, por exemplo, a reconhecer mais amostras do subtipo F *naive*, o desempenho certamente cairia bastante.

As primeiras estratégias pensadas para resolver este problema consistiam em igualar a quantidade de amostras de todos os grupos, pelo máximo, ou pelo mínimo. Limitar pelo mínimo consistia em selecionar aleatoriamente apenas 19 amostras de cada um dos demais grupos, e usar todas as do grupo F *naive*. Desta forma, se chegaria a uma mesma quantidade para todos os subgrupos. Porém, passariam a ser utilizados apenas 9,5% dos dados disponíveis.

Na opção de igualar o número de amostras dos grupos pelo máximo, os grupos menores foram replicados até que todos possuíssem a mesma quantidade. Assim, todos teriam 747 amostras, que é o número existente no grupo B resistente. Essa abordagem permitiu utilizar 100% da base de dados, implicando, porém, em trabalhar com muitas amostras repetidas.

Embora tenham contribuído bastante para equilibrar a taxa de acerto entre os grupos, tais abordagens não representam boa utilização dos dados. O balanceamento pelo mínimo implica na não utilização de mais de 90% das amostras, o que impossibilita uma comparação adequada com os experimentos anteriores. Não há como precisar se uma mudança das amostras selecionadas resultará num aumento ou mesmo numa diminuição significativa dos resultados.

Em contrapartida, o balanceamento pelo máximo utiliza toda a base de dados. Porém, implica na existência de dados repetidos num mesmo experimento. O grupo F *naive*, por exemplo, precisou ser replicado 40 vezes para atingir a quantidade necessária. Como as amostras de treinamento e reconhecimento são sempre escolhidas aleatoriamente, existe uma grande probabilidade de que uma mesma amostra tenha sido selecionada tanto para reconhecimento como para treinamento. Além disso, neste ultimo caso, muitas tiveram que ser usadas mais de uma vez durante o processo de reconhecimento. Isto certamente afeta o grau de acerto do experimento, pois dá um peso maior para a resposta dessas amostras em relação às demais.

Portanto, acredita-se ser necessário buscar uma nova maneira de equilibrar o tamanho dos grupos. Outra forma de balanceamento possível é o loteamento das amostras, consistindo esta estratégia em separar a base em lotes distintos. Cada lote deve ter a mesma quantidade de amostras, sem repetições. Os experimentos são realizados para cada lote independentemente e depois é calculada a média dos resultados obtidos.

Assim, seriam utilizadas todas as 19 amostras do grupo F *naive*, sendo escolhidas aleatoriamente outras 19 de cada um das demais categorias. Cada lote possuiria 54 no total. Esse processo seria repetido até que todo o grupo B resistente fizesse parte de algum lote. Tal estratégia de balanceamento utilizaria toda a base de dados e possibilitaria comparar esses resultados com os obtidos anteriormente.

3.4.5 Validação Cruzada:

A validação cruzada 10-FOLD consiste em um processo onde a base de dados é separada em 10 conjuntos distintos, contendo aproximadamente o mesmo número de amostras. Depois de criados os conjuntos, nove deles são usados para o treinamento e o 10º é usado para o reconhecimento. Esse processo é feito 10 vezes, de modo que todos os conjuntos sejam reconhecidos independentemente.

Ao final, o grau de acerto desse experimento é a média dos graus de acerto dos 10 conjuntos reconhecidos. Desta forma, não se corre o risco do resultado do experimento estar vinculado às amostras que foram usadas para a etapa de treinamento e às que foram separadas para o reconhecimento. Porém, como podem ocorrer grandes variações entre os 10 resultados obtidos, é importante calcular também o desvio padrão entre esses conjuntos.

3.4.6 Objetivo do Reconhecimento:

Nos trabalhos mencionados no Capítulo 1 geralmente busca-se reconhecer a existência de resistência a um determinado medicamento dentre amostras de um mesmo subtipo. Neste trabalho buscou-se Inicialmente, reconhecer os seis grupos listados na base de dados: B resistente, B *naive*, C resistente, C *naive*, F resistente, F *naive*. Em um segundo momento, pensou-se ser interessante uma abordagem por etapas, na qual primeiro se reconheceria o subtipo do vírus, diferenciando somente entre B, C e F, e em uma segunda etapa, estes seriam separados entre resistentes e selvagens. Ao final destas duas etapas haveria novamente as respostas separadas nos seis grupos, sendo assim possível se analisar e comparar os resultados. Esse procedimento não só poderia dar resultados melhores que os obtidos inicialmente, como serviria de base para futuros experimentos nos quais se buscaria o reconhecimento dos medicamentos aos quais o vírus apresentasse resistência, independente de se conhecer a priore o subtipo.

O reconhecimento por subtipo citado acima poderia ser feito de duas formas diferentes. Na primeira, os seis grupos seriam treinados e reconhecidos da mesma maneira como estava sendo feito anteriormente. Porém, ao analisar as respostas, só a informação do subtipo seria considerada. A segunda forma consiste em agrupar as amostras resistentes e *naives* de acordo com o subtipo antes de treiná-las. Deste modo, na rede neural, só a informação de subtipo estaria armazenada.

Embora a segunda forma aparentemente fosse a mais adequada, os resultados da primeira se apresentaram de forma muito mais rápida. Por já se ter os resultados dos experimentos com os seis grupos, não seria preciso nenhum novo treinamento. Seria necessário apenas a análise e interpretação dos resultados obtidos com os experimentos anteriores de maneira diferente. Isso daria uma previsão dos resultados que deveriam ser encontrados ao se realizar os experimentos agrupando-se as amostras resistentes com as *naives*. Além disso, seria um bom indicativo para seguir adiante ou não.

Por fim, realizou-se também um experimento somente com as amostras do subtipo B onde se buscou diferenciar B *naive* e B resistente. Dentre os subtipos presentes na base de dados, este é o que possui maior conhecimento biológico, e entender como a rede neural atua no reconhecimento dele poderia servir de base para evoluir os conhecimentos acerca dos demais subtipos.

3.4.7 Resumo dos Experimentos Realizados:

Nos tópicos anteriores, foram abordadas todas as características pensadas no momento da realização dos experimentos, bem como, a importância e a forma nas quais elas foram trabalhadas. Na tabela 3.7 há um quadro resumo onde é possível observar de que forma essas características foram combinadas. A única peculiaridade que não foi incluída nesta síntese, foi a forma de codificação utilizada, pois, conforme dito, sempre todas foram testadas. Sendo assim, cada linha da tabela 3.7 representa um conjunto de 10 experimentos, onde foram numeradas as redes desenvolvidas de maneira a simplificar a exibição dos resultados apresentados futuramente.

Tabela 3.7 Resumo dos Experimentos realizados

Número da WISARD	Grupos Usados	Balanceamento dos Dados	Posições	Tamanho da Memória
1	B resistente, B *naive,* C resistente, C *naive,* F resistente, F *naive.*	Nenhum	22 clássicas	8 bits
2	B resistente, B *naive,* C resistente, C *naive,* F resistente, F *naive.*	Nenhum	99 (toda a entrada)	8 bits
3	B resistente, B *naive,* C resistente, C *naive,* F resistente, F *naive.*	Nenhum	27 mais significativas	8 bits
4	B resistente, B *naive,* C resistente, C *naive,* F resistente, F *naive.*	Maior grupo (B resistente)	99 (toda a entrada)	8 bits
5	B resistente, B *naive,* C resistente, C *naive,* F resistente, F *naive.*	Menor grupo (F selvagem)	99 (toda a entrada)	8 bits
6	B resistente, B *naive,* C resistente, C *naive,* F resistente, F *naive.*	Lote	99 (toda a entrada)	8 bits
7	B resistente, B *naive,* C resistente, C *naive,* F resistente, F *naive.*	Lote	99 (toda a entrada)	16 bits
8	B (B resistente + B *naive),* C (C resistente + C *naive),* F (F resistente + F *naive)*	Lote	99 (toda a entrada)	8 bits
9	B (B resistente + B *naive),* C (C resistente + C *naive),* F (F resistente + F *naive)*	Lote	99 (toda a entrada)	16 bits
10	B resistente, B *naive*	Lote	99 (toda a entrada)	8 bits
11	B resistente, B *naive*	Lote	99 (toda a entrada)	16 bits

4 Análise dos Resultados

No Capítulo 3 as diferentes características que poderiam influenciar os experimentos foram abordadas, e listaram-se todos os experimentos realizados neste trabalho. Neste Capítulo, será mais bem detalhada a evolução da utilização destas características, o que será feito na medida em que os resultados obtidos forem demonstrados e analisados. Ao final, será realizada ainda, uma análise da técnica do *Bleaching*, já descrita no Capítulo 2, que foi utilizada em todos os experimentos. Será devidamente avaliada a sua importância e implicações nos resultados obtidos.

4.1 Análise em Relação ao Problema (Classificação de HIV-1)

A principal métrica usada nesta pesquisa foi o grau de acerto da WiSARD em relação aos grupos apresentados. Cada experimento foi analisado e serviu de base para os seguintes, de modo a aprimorar o desempenho da rede. A seguir serão descritos cada experimento, seus resultados e o que foi observado. É importante lembrar que para cada conjunto de parâmetros escolhido, todas as 10 codificações mencionadas no Capítulo 3, foram alvos de experimentos.

4.1.1 Experimentos Iniciais

Os três primeiros experimentos foram realizados com todas as 1205 amostras disponíveis, sem haver qualquer tipo de tratamento prévio na base de dados e utilizando na rede neural memórias de 8 bits. Este tamanho de memória foi escolhido pela crença de que seria o mais adequado, havendo a expectativa de que em relação ao tamanho dos dados, este traria um melhor equilíbrio entre generalização e especificidade.

Tais experimentos diferenciavam-se apenas em relação às posições da protease lidas pela WiSARD. No primeiro foram usadas 22 das posições clássicas de assinatura de subtipo ou de mutação de resistência já conhecidas na literatura. Em

seguida, todas as 99 posições da protease foram usadas, e posteriormente, foram selecionadas 27 posições, escolhidas com base em uma análise das amostras disponíveis na base de dados.

4.1.1.1 Posições de Mutação já Conhecidas

Havia a necessidade de se verificar inicialmente o desempenho da rede em relação ao tamanho dos dados utilizados, buscava-se também obter bons resultados. Por isso, decidiu-se pela seleção de algumas das posições clássicas descritas na literatura, para a realização do primeiro experimento. Esta seleção contempla algumas posições de assinatura de subtipo e também posições reconhecidamente importantes para identificar a resistência a medicamentos.

Este primeiro conjunto de parâmetros apresentou resultados que variaram entre 64,6% até 68,9%, e calculando-se a média de todas as codificações utilizadas, obteve-se um valor de 66,3% de acerto. Embora os resultados não tenham sido excelentes, apresentavam um desvio padrão muito baixo, sendo sempre inferior a 2,4%. Portanto, o reconhecimento teve uma grande estabilidade. Além disso, fora usado pouco mais de um quinto das informações disponíveis, e a rede não apresentou qualquer tipo de problema em relação ao tamanho dos dados. Sendo assim, viu-se que era possível executar experimentos usando todas as 99 posições dos aminoácidos, o que seguramente levaria à melhores resultados.

4.1.1.2 Toda a Protease (99 posições)

Conforme previsto no Capítulo 3, o uso de todas as informações da entrada melhorou em muito os resultados. Esses experimentos obtiveram grau de acurácia variando de 77,8% até 82%. Da análise do desempenho médio de todas as codificações, obtive-se 79,6% de acerto. Em relação ao experimento anterior, isso representa um aumento de 13,3 pontos percentuais.

Esta melhora foi obtida sem alterar de forma significativa o desvio padrão, que se manteve sempre abaixo de 3,9%. Porém, o aumento do número de memórias por discriminador implicou em um maior tempo de execução. Além disso, causou

diminuição da relevância de cada memória em relação à pontuação de cada categoria, o que por sua vez, diminuiu bastante a confiança[3] do resultado.

4.1.1.3 Posições Mais Significativas (27 posições)

Da observação dos dados acessíveis, percebeu-se que havia posições com taxa de mutação muito baixa, inferior a 2%. Portanto, tais posições não seriam relevantes para a categorização das amostras presentes na base de dados. Foram selecionadas então, somente as posições onde houvesse mutações significativas. Deste modo seria obtido o melhor resultado, com a melhor confiança, em menor tempo possível. Tal abordagem se assemelha a realizar uma clusterização prévia dos dados, selecionando-se em seguida as posições mais representativas de cada cluster. Assim, buscou-se a obtenção das melhores características das duas configurações anteriores.

Para a realização destes experimentos foi analisada a frequência da ocorrência de mutações em todas as posições da protease. Na primeira etapa foram removidas todas as posições que apresentavam frequência de mutações inferior a 2%. Tais posições não seriam relevantes na diferenciação entre os seis grupos. Em seguida retiraram-se também as posições onde a mutação era homogênea ou ocorria para mais de um grupo. Como exemplo, pode-se citar a terceira posição, na qual mais de 97% das amostras mudavam de Valina para Isoleucina. Mutações como esta não iriam contribuir para diferenciar os grupos, e talvez, pudessem até estar prejudicando o reconhecimento. Ao final deste processo foram selecionadas 27 posições.

Ao executar esses experimentos, foi verificado um grau de acurácia ainda melhor do que o esperado. Os resultados obtidos variavam de 79,8%, até 85,3%, tendo o desvio padrão mantido o mesmo patamar anterior. Isto representa uma média 3,5% acima que a encontrada usando toda a protease. Além disso, o tempo e a confiança desses experimentos foram tão bons quanto o dos experimentos usando as posições conhecidas, conforme se objetivou.

[3] Tal trecho refere-se à confiança explicada no Capítulo 2, que é calculada com base nos dois discriminadores de maior pontuação da rede neural.

4.1.1.4 Observações Acerca dos Experimentos Iniciais:

Na tabela 4.1 é possível observar os resultados obtidos nos experimentos descritos acima. Apesar dos bons resultados encontrados até o momento, era preciso descobrir o que estava impedindo a WiSARD de ter resultados ainda melhores. Foi feita então uma análise mais detalhada, na qual se percebeu que a acurácia dos experimentos realizados até o momento variava muito entre os grupos. Em todos os testes, as amostras de subtipo B resistente apresentavam resultados excelentes enquanto que os demais grupos tinham resultados muito ruins.

Tabela 4.1 Resumo dos experimentos iniciais.

		CONHECIDAS (22)		TODAS (99)		MAIS SIGNIFICATIVAS (27)	
CÓDIGO	N AMOSTRAS	MÉDIA	DESVIO	MÉDIA	DESVIO	MÉDIA	DESVIO
BIN20	1205	64,7%	1,8%	78,3%	2,9%	79,8%	3,1%
HIDROBIN	1205	67,6%	2,0%	80,4%	3,3%	85,2%	2,7%
HIDROGRAY	1205	68,9%	2,0%	81,6%	2,5%	85,3%	2,4%
HIDROGRAY23	1205	67,2%	1,9%	79,8%	3,1%	84,7%	3,2%
MMBIN	1205	64,8%	0,7%	82,0%	2,4%	85,1%	2,7%
MMGRAY	1205	64,8%	1,3%	78,5%	3,9%	80,2%	3,3%
MMGRAY23	1205	65,1%	1,5%	80,5%	3,1%	83,1%	2,8%
HIDROMMBIN	1205	67,2%	2,2%	77,8%	2,4%	82,6%	2,6%
HIDROMMGRAY	1205	64,6%	1,8%	78,3%	2,1%	82,7%	3,1%
HIDROMMGRAY23	1205	67,6%	2,4%	79,1%	3,3%	83,4%	3,8%
MEDIA	1205	66,3%	1,8%	79,6%	2,9%	83,2%	3,0%

Considerando os 30 experimentos realizados até o momento, o grupo do subtipo B resistente obteve taxa de acerto entre 97,5% e 99,5%. Em contrapartida, o grupo do subtipo F *naive* não conseguiu grau de acerto maior que 20% em nenhum momento. Além disso, este grupo não apresentou acerto algum em qualquer um dos experimentos com as posições clássicas. Tais resultados podem ser vistos com mais detalhes na tabela 1 do ANEXO II.

Esse fato tornou-se extremamente preocupante, posto que não é aceitável que um reconhecedor de padrões seja eficiente somente para um padrão e ineficiente para os demais. Seria necessário, portanto, estratégias que equilibrassem melhor o grau de acerto entre os grupos usados.

Outra análise importante, feita em relação aos experimentos já realizados, é que dificilmente se chegaria a bons resultados considerando apenas as posições clássicas conhecidas na literatura. No primeiro experimento foram selecionadas

algumas delas, de modo a ter um conjunto pequeno, mas com alta relevância. Da comparação dessas posições com as 27 mais significativas, nota-se que havia muitas posições diferentes. Portanto, dentre as posições clássicas selecionadas muitas possuíam taxa de mutação baixa e dificilmente resultariam em uma boa caracterização para os grupos.

Além disso, embora as 27 posições mais significativas tivessem apresentado os melhores resultados, estas estavam muito ligadas à análise feita com essa base de dados. Possivelmente, em experimentos futuros, este intervalo poderia não apresentar resultados tão bons com um novo conjunto de amostras. O mesmo poderia ocorrer para as posições clássicas. Em vista disso, todos os que se seguiram utilizaram as 99 posições da protease. Porém era preciso pensar em possíveis melhorias para a confiança e para diminuir o tempo de execução.

4.1.2 Experimentos Balanceados Pelo Máximo e Pelo Mínimo.

A grande diferença entre o número de amostras por grupo certamente era a responsável pela enorme discrepância entre os graus de acertos dos experimentos anteriores. Uma categoria com muito mais amostras que as demais pode não ser um fator preocupante na WiSARD tradicional. Porém, com o uso do *Bleaching*, isto se mostrou muito relevante.

Na WiSARD descrita por ALEXANDER (1998), quando há o treinamento excessivo de um dos grupos, pode haver um maior espalhamento desta categoria em relação às demais. Isto pode aumentar a probabilidade de um padrão ser reconhecido erroneamente, porém dificilmente implica em uma incapacidade de se reconhecer todas as demais categorias.

Já quando se usa a técnica de *Bleaching* percebe-se que isto é extremamente preocupante. No caso deste estudo, foram treinadas 672 amostras do subtipo B resistente e apenas 17 do subtipo F *naive*. Observou-se que quando o valor do *Bleaching* passava de 17, o último grupo já era descartado, e quando passava de 139, que equivalia ao número de treinamentos realizados com o subtipo C resistente, somente o grupo B resistente passava a ser possível como resposta. Diante disto, viu-se que era preciso aproximar ao máximo o número de treinamentos de cada grupo, sendo inicialmente cogitado usar ou o máximo, ou o mínimo. Conforme mencionado no item 3.4.3, foram realizados experimentos de ambas as formas.

Os testes feitos com estas estratégias de balanceamento diminuíram em muito as diferenças entre a média de acertos dos grupos. No caso do balanceamento pelo subtipo B resistente, o grupo com menor taxa de acerto foi o B *naive* que variou de 78% a 87,5% e o de maior taxa foi o F *naive*, que variou de 94,8% a 100%. Dentre todos os casos, o que apresentou maior variação possuía taxas entre 80% e 100%, ou seja, uma tinha uma diferença máxima de 20% entre dois grupos. Na tabela 4.2 essas diferenças podem ser observadas com detalhe.

Tabela 4.2 Comparação dos grupos com balanceamento pelo grupo B resistente.

CÓDIGO	B RESISTENTE	B *NAIVE*	C RESISTENTE	C *NAIVE*	F RESISTENTE	F *NAIVE*
BIN20	81,6%	78,0%	80,0%	85,0%	87,0%	94,8%
HIDROBIN	91,6%	79,0%	85,0%	89,6%	95,8%	94,8%
HIDROGRAY	91,0%	81,0%	87,8%	91,0%	95,6%	94,8%
HIDROGRAY23	91,3%	87,5%	81,8%	88,7%	96,1%	100,0%
MMBIN	90,1%	84,9%	88,2%	83,0%	95,3%	94,8%
MMGRAY	91,3%	80,6%	88,2%	85,3%	97,1%	100,0%
MMGRAY23	89,8%	83,5%	88,2%	86,7%	94,4%	100,0%
HIDROMMBIN	89,1%	81,9%	84,7%	83,7%	95,0%	100,0%
HIDROMMGRAY	91,7%	83,7%	82,7%	83,9%	97,2%	98,4%
HIDROMMGRAY23	87,7%	83,3%	85,5%	85,8%	93,4%	100,0%

Para os experimentos usando apenas 19 amostras de cada grupo, foram obtidos resultados semelhantes, mas não tão bons. O grupo de maior acurácia foi o B resistente, cuja taxa variou de 80% a 100% e o de menor foi o grupo C *naive*, que obteve variações de 50% a 70%. Neste caso, a maior variação encontrada foi de 35%, o que foi observado em três experimentos.

Ao usar menos amostras, cada erro de reconhecimento ocorrido possui um peso maior. Portanto, essa variação observada nos treinamentos que foram balanceados pelo mínimo pode ser vista como uma consequência natural desta forma de balanceamento. Isto também pode ser visto ao comparar os desvios padrão e as taxas de acerto médio dos experimentos. Na tabela 4.3 estão dispostos os resultados para este balanceamento.

Tabela 4.3 Comparação dos grupos com balanceamento pelo grupo F *naive*.

CÓDIGO	B RESISTENTE	B *NAIVE*	C RESISTENTE	C *NAIVE*	F RESISTENTE	F *NAIVE*
BIN20	100,0%	70,0%	85,0%	65,0%	80,0%	80,0%
HIDROBIN	95,0%	65,0%	80,0%	60,0%	85,0%	80,0%
HIDROGRAY	90,0%	70,0%	70,0%	60,0%	75,0%	85,0%
HIDROGRAY23	90,0%	60,0%	75,0%	70,0%	85,0%	90,0%
MMBIN	85,0%	80,0%	75,0%	55,0%	85,0%	80,0%
MMGRAY	90,0%	60,0%	75,0%	55,0%	90,0%	70,0%
MMGRAY23	95,0%	85,0%	85,0%	65,0%	85,0%	70,0%
HIDROMMBIN	85,0%	80,0%	85,0%	50,0%	75,0%	85,0%
HIDROMMGRAY	80,0%	60,0%	70,0%	55,0%	85,0%	85,0%
HIDROMMGRAY23	85,0%	70,0%	85,0%	65,0%	80,0%	85,0%

Comparando o grau de acerto para os treinamentos realizados com essas duas formas de balanceamento vislumbrou-se que os realizados com 747 amostras por grupo variaram entre 84,4% a 90,9%, e desvios padrão de 0,7% a 1,7%. No caso dos experimentos com 19 amostras por grupo a taxa de acerto variou de 73,7% a 80,8% e o desvio padrão ficou entre 6,7% a 11,7%. Embora neste último caso os desvios padrão ainda sejam considerados baixos, são proporcionalmente muito maiores que no caso anterior. A comparação desses experimentos pode ser vista na tabela 4.4 e em maior detalhe na tabela 2 do ANEXO II.

Tabela 4.4 Comparação entre resultados balanceados pelo máximo e pelo mínimo.

	BALANCEAMENTO F *NAIVE*			BALANCEAMENTO B RESISTENTE		
CÓDIGO	N AMOSTRAS	MÉDIA	DESVIO	N AMOSTRAS	MÉDIA	DESVIO
BIN20	114	79,8%	7,3%	4482	84,4%	1,5%
HIDROBIN	114	77,3%	6,8%	4482	89,3%	1,2%
HIDROGRAY	114	74,7%	6,9%	4482	90,2%	0,7%
HIDROGRAY23	114	78,3%	6,7%	4482	90,9%	1,5%
MMBIN	114	75,6%	10,2%	4482	89,4%	1,1%
MMGRAY	114	74,6%	8,3%	4482	90,4%	1,3%
MMGRAY23	114	80,8%	9,1%	4482	90,4%	1,3%
HIDROMMBIN	114	76,6%	11,7%	4482	89,1%	1,7%
HIDROMMGRAY	114	73,7%	6,7%	4482	89,6%	1,5%
HIDROMMGRAY23	114	78,9%	7,5%	4482	89,3%	1,6%

Pelos dados apresentados nas tabelas 4.2 e 4.3, foram confirmadas as suspeitas de que a grande diferença entre o número de amostras por grupos estava prejudicando o reconhecimento. Porém, conforme descrito no tópico 3.4.3, não foi

completamente satisfatório se balancear os dados destas formas. Por isto foram realizados novos experimentos com o balanceamento por lotes.

4.1.3 Balanceamento Por Lote

Neste tópico, serão analisados somente os experimentos com lotes que visavam reconhecer os seis grupos já descritos. Os demais casos que também utilizaram balanceamento por lote serão abordados nos tópicos 4.2 e 4.3.

Conforme mencionado no Capítulo 3, foram necessários 40 lotes para a cobertura de toda a base de dados. Foi considerada como resultado destes experimentos a média de todos os acertos em relação ao total de amostras. Não foi calculada primeiramente a média dos lotes, pois matematicamente, não haveria diferença. Fazendo a média entre os lotes, ter-se-ia também um desvio padrão entre eles, porém esta medida não seria relevante em relação ao resultado final. Além disso, pode se considerar esta forma de balanceamento como um meio termo entre as duas anteriores. Por isso, espera-se que a diferença da taxa de acerto entre os grupos neste caso seja um valor intermediário entre os valores observados anteriormente.

A partir deste momento, passou-se a realizar experimentos com WiSARD de memórias de 16 bits. Embora a convicção inicial fosse que 8 bits seria o melhor tamanho a ser usado, seria interessante ao estudo testar memórias maiores, tendo em vista que em um momento futuro poderia ser utilizada a transcriptase reversa (RT) do vírus. Outro ponto interessante da utilização de memórias maiores é que tal fato diminuiria o total de memórias por discriminador, o que tornaria a WiSARD mais rápida.

Na tabela 4.5 demonstram-se os resultados dos grupos tanto nos experimentos com memórias de 8 bits, quanto de 16 bits. Pode-se observar que a diferença da acurácia entre os grupos para estes experimentos variaram entre 17% e 31%. Além disso, enquanto nos experimentos anteriores um único grupo apresentava as maiores taxas de acerto para todas as codificações, neste momento isso deixa de ser uma regra.

Tabela 4.5 Comparação dos grupos com balanceamento por lote para WiSARD com 8 e 16 bits.

	CÓDIGO	B RESISTENTE	B *NAIVE*	C RESISTENTE	C *NAIVE*	F RESISTENTE	F *NAIVE*
MEMÓRIAS DE 8 BITS	BIN20	73,0%	63,0%	70,5%	51,1%	72,1%	57,6%
	HIDROBIN	73,9%	61,0%	68,4%	52,5%	75,6%	62,8%
	HIDROGRAY	72,0%	61,3%	68,9%	53,4%	72,5%	73,3%
	HIDROGRAY23	74,0%	63,9%	68,5%	56,4%	78,0%	73,5%
	MMBIN	73,5%	57,6%	70,8%	51,6%	79,8%	68,4%
	MMGRAY	68,3%	58,8%	61,1%	51,3%	69,8%	62,8%
	MMGRAY23	73,0%	62,5%	69,1%	51,4%	76,4%	73,9%
	HIDROMMBIN	67,0%	62,8%	59,8%	49,8%	73,6%	64,0%
	HIDROMMGRAY	72,8%	60,8%	61,3%	52,8%	73,9%	59,9%
	HIDROMMGRAY23	68,3%	62,1%	64,9%	53,9%	71,3%	65,1%
MEMÓRIAS DE 16 BITS	BIN20	75,1%	61,3%	73,4%	54,4%	70,1%	58,6%
	HIDROBIN	69,5%	68,4%	71,0%	52,1%	78,3%	72,4%
	HIDROGRAY	72,4%	66,8%	68,4%	54,9%	77,6%	70,8%
	HIDROGRAY23	73,1%	66,0%	70,0%	54,8%	73,3%	75,0%
	MMBIN	74,3%	63,4%	72,3%	51,4%	80,0%	73,6%
	MMGRAY	67,8%	65,3%	63,8%	55,9%	73,8%	71,8%
	MMGRAY23	70,6%	68,6%	69,8%	54,1%	76,8%	73,1%
	HIDROMMBIN	66,4%	66,4%	60,6%	49,5%	73,0%	81,1%
	HIDROMMGRAY	68,6%	65,0%	65,6%	55,6%	74,1%	69,3%
	HIDROMMGRAY23	72,3%	64,8%	67,9%	55,8%	77,3%	65,0%

Ao realizar o balanceamento por lote, foram atingidos resultados equilibrados entre os grupos, não tendo sido preciso repetir amostras nos treinamentos/ reconhecimentos, sendo utilizada toda a base de dados disponível. Porém, apesar do alcance dos objetivos almejados ao procurar uma nova forma de balanceamento, tais experimentos possuem acurácia inferior a quase todas obtidas até o momento. Além disso, em relação ao desvio padrão pode-se dizer que estes experimentos apresentaram valores intermediários, entre os apresentados pelos testes usando balanceamento pelo máximo e os obtidos com balanceamento pelo mínimo.

Na tabela 4.6 podem ser verificados os resultados gerais para estes experimentos. Utilizando memórias de 8 bits foram obtidos no máximo uma taxa de acerto de 69%. Este resultado é superior ao encontrado em qualquer um dos experimentos usando somente as posições já conhecidas na literatura. Porém, também apresentou acertos de 61,9%, 62,8% e 63,5%, taxas estas inferiores a quaisquer experimentos obtidos até o momento. Por conta disso, sua média é de 65%, abaixo inclusive da média obtida entre os experimentos com as posições clássicas que é de 66%.

Tabela 4.6 Resultados dos experimentos com lote para WiSARD de 8 e 16 bits.

		Memórias de 8 bits		Memórias de 16 bits	
CÓDIGO	N AMOSTRAS	MÉDIA	DESVIO	MÉDIA	DESVIO
BIN20	4560	64,6%	4,6%	65,5%	3,4%
HIDROBIN	4560	65,7%	4,9%	68,6%	5,1%
HIDROGRAY	4560	66,9%	4,0%	68,5%	5,7%
HIDROGRAY23	4560	69,0%	4,5%	68,7%	7,0%
MMBIN	4560	66,9%	5,1%	69,1%	5,9%
MMGRAY	4560	62,0%	7,2%	66,4%	8,0%
MMGRAY23	4560	67,7%	5,4%	68,8%	7,1%
HIDROMMBIN	4560	62,8%	5,2%	66,2%	5,3%
HIDROMMGRAY	4560	63,5%	5,0%	66,4%	6,9%
HIDROMMGRAY23	4560	64,3%	4,7%	67,1%	5,8%
MÉDIA	4560	65,3%	5,1%	67,5%	6,0%

Os resultados alcançados com memórias de 16 bits ficaram entre 65,5% e 69,1%, obtendo-se média de 67,5%. Foram, portanto, um pouco melhores que o resultados com 8 bits balanceados por lote e também os que usavam somente as posições clássicas. Além dos resultados com memórias de 16 bits terem sido levemente maiores que os obtidos com memórias de 8 bits, de fato, foram realizados em menor tempo. Eles também confirmaram que a WiSARD tem capacidade para trabalhar com amostras maiores, que contenham, por exemplo, a sequência dos aminoácidos da transcriptase reversa (RT).

4.1.4 Resumo das Comparações Usando os Seis Grupos

Foram analisadas até o momento as experiências realizadas com os grupos de amostras: B resistente, B *naive*, C resistente, C *naive*, F resistente e F *naive*. O estudo iniciou-se verificando qual seria a melhor a entrada. Em seguida foi observada qual a melhor forma de balancear os dados, sendo feito na sequência experimentos usando não só memórias de 8, mas também, com 16 bits.

Em relação à escolha das posições da protease, concluiu-se no tópico 4.1.1.4 que seria preciso utilizar todas as 99 posições para que os resultados fossem completamente independentes da base de dados. Embora o uso de toda a protease tenha resultado em um primeiro momento numa confiança muito baixa, nos tópicos acerca do *Bleaching* será descrito como esta técnica melhorou este parâmetro. O balanceamento dos dados mostrou-se fundamental, e embora os melhores resultados

tenham sido obtidos ao usar 747 amostras nos treinamentos de cada grupo, pelas considerações feitas no tópico 3.4.4 devem ser considerados os valores obtidos com o balanceamento por lote.

Sendo assim, os experimentos que se seguiram foram sempre realizados com o uso de toda a protease e do balanceamento por lote. Embora o tamanho e o número de lotes possam mudar de acordo com o foco do experimento, a lógica de como são criados foi sempre mantida. Em relação ao uso de memórias de ou 8 bits ou 16 bits, os resultados são muito parecidos, sendo a única grande diferença a velocidade dos experimentos com 16 bits, que é cerca de 50% maior. Continuaram sendo usadas, portanto, ambos os tamanhos de memórias.

4.1.5 Experimentos Por Subtipo

Nos experimentos anteriores a base de dados foi separada em seis categorias, buscando-se assim diferenciar as formas resistentes das formas *naives* do vírus e não apenas o subgrupo. Porém, as amostras resistentes não possuem necessariamente o mesmo padrão, tendo em vista que resistências a remédios diferentes podem estar ligadas às mutações em posições distintas. Conforme descrito no tópico 3.4.6, é crível que seria interessante a separação das amostras de acordo com o subtipo do vírus, não se diferenciando inicialmente entre resistentes e *naives*. Neste mesmo tópico viu-se que esta análise poderia ser feita de duas formas distintas: agrupando as amostras por subtipo antes de realizar o treinamento ou juntando as respostas encontradas de acordo com o subtipo, ignorando-se assim a informação de ser ou não resistente. Em um experimento futuro, a amostra poderia ser submetida a uma nova rede, que por sua vez estará configurada para distinguir os medicamentos para os quais aquela amostra possui resistência.

Nas tabelas 4.7 e 4.8, pode-se ver que os resultados obtidos ao utilizar memórias de 8 bits, e treinamento com os seis grupos, variaram de 86,6% a 92,9%. Já para os resultados agrupando os subtipos antes do treinamento a variação da taxa de acerto foi de 82,6% a 91,1%. No caso da WiSARD com memórias de 16 bits, treinando os seis grupos separadamente o grau de acerto variou de 89,5% a 94% e ao treinar após ter agrupado por subtipo o acerto ficou entre 86,5% e 93,5%. Em nenhum desses casos o desvio padrão foi superior a 5%.

Tabela 4.7 Resultados dos experimentos por subtipo com memórias de 8 bits.

		SEM AGRUPAR		AGRUPANDO	
CÓDIGO	N AMOSTRAS	MÉDIA	DESVIO	MÉDIA	DESVIO
BIN20	4560	91,2%	1,7%	89,1%	2,4%
HIDROBIN	4560	91,4%	2,5%	88,8%	2,8%
HIDROGRAY	4560	91,1%	2,2%	89,4%	2,3%
HIDROGRAY23	4560	92,9%	1,6%	91,0%	1,7%
MMBIN	4560	92,5%	1,9%	91,1%	2,2%
MMGRAY	4560	86,6%	4,1%	83,5%	5,4%
MMGRAY23	4560	91,6%	2,5%	89,8%	3,0%
HIDROMMBIN	4560	87,0%	2,9%	82,6%	2,9%
HIDROMMGRAY	4560	87,3%	4,3%	85,4%	3,8%
HIDROMMGRAY23	4560	88,3%	3,3%	84,2%	2,9%
MÉDIA	4560	90,0%	2,7%	87,5%	2,9%

Pelos resultados expostos aqui nota-se que a melhor forma de se reconhecer o subtipo do vírus é treinando os grupos resistentes e *naives* separadamente usando uma rede com memória de 16 bits. Além disso, os bons resultados obtidos, em alguns casos próximos de 94%, evidenciam que é possível usar essa abordagem em um trabalho futuro, onde seria feita a distinção dos medicamentos para os quais o paciente apresentará resistência.

Tabela 4.8 Resultados dos experimentos por subtipo com memórias de 16 bits.

		SEM AGRUPAR		AGRUPANDO	
CÓDIGO	N AMOSTRAS	MÉDIA	DESVIO	MÉDIA	DESVIO
BIN20	4560	91,2%	2,5%	89,7%	1,6%
HIDROBIN	4560	92,9%	2,4%	93,1%	2,6%
HIDROGRAY	4560	92,3%	2,7%	90,7%	3,1%
HIDROGRAY23	4560	93,3%	2,9%	93,5%	2,2%
MMBIN	4560	94,0%	1,3%	93,5%	1,9%
MMGRAY	4560	90,2%	5,0%	87,8%	4,7%
MMGRAY23	4560	93,5%	2,2%	92,8%	2,4%
HIDROMMBIN	4560	89,5%	2,6%	86,5%	4,0%
HIDROMMGRAY	4560	91,1%	3,2%	88,4%	3,7%
HIDROMMGRAY23	4560	90,8%	4,4%	88,6%	4,2%
MÉDIA	4560	91,9%	2,9%	90,5%	3,1%

4.1.6 Experimento B Resistente Versus B *Naive*

Neste último experimento, buscou-se diferenciar as amostras do subtipo B em relação à existência ou não de resistência. Grande parte dos trabalhos mencionados no Capítulo 1 tem como objetivo a diferenciação das amostras resistentes a um dado medicamento quanto às amostras não resistentes. Além disso, por este subtipo ser o mais abundante na Europa e nas Américas, ele costuma ser o mais estudado, sendo inclusive o utilizado em grande parte dos trabalhos mencionados no primeiro Capítulo.

É importante ressaltar que neste experimento não houve separação por medicamento. Isto pode influenciar o resultado levando-o até mesmo a apresentar acurácia inferior à obtida na categorização dos seis grupos. Porém, é possível que uma taxa de acerto não muito elevada, agora, venha a ser útil em um experimento futuro. No próximo Capítulo será descrito como a WiSARD pode ser aperfeiçoada ao incluir em seu funcionamento, conhecimentos biológicos já existentes.

Desta forma, mesmo que deste experimento não se obtenha resultados tão bons, as futuras alterações desta rede neural certamente implicarão em grandes melhoras. Além disso, é possível que tais aperfeiçoamentos ensejem numa melhora para os resultados dos demais subtipos que ainda não possuem estudos biológicos aprofundados, o que viria a contribuir para o conhecimento em relação aos demais subtipos.

Tabela 4.9 Resultados dos experimentos com o subtipo B para memórias de 8 e 16 bits.

		MEMÓRIA DE 8 BITS		MEMÓRIA DE 16 BITS	
CÓDIGO	N AMOSTRAS	MÉDIA	DESVIO	MÉDIA	DESVIO
BIN20	1584	74,9%	6,4%	73,7%	4,4%
HIDROBIN	1584	73,2%	6,3%	75,4%	5,3%
HIDROGRAY	1584	74,2%	7,3%	76,4%	6,1%
HIDROGRAY23	1584	74,1%	6,7%	76,1%	5,4%
MMBIN	1584	73,5%	4,9%	74,2%	5,7%
MMGRAY	1584	73,9%	2,8%	73,9%	3,6%
MMGRAY23	1584	73,2%	5,2%	74,4%	5,9%
HIDROMMBIN	1584	73,7%	5,9%	74,6%	7,0%
HIDROMMGRAY	1584	74,1%	7,2%	73,8%	4,9%
HIDROMMGRAY23	1584	73,8%	6,7%	75,9%	6,1%
MÉDIA	1584	73,9%	6,0%	74,8%	5,4%

Na tabela 4.9 pode-se ver que, ao utilizar memórias de 8 bits, chega-se a uma taxa de acerto que variou entre 73,2% e 74,8% e desvio padrão de 2,8% a 7,2%. Assim como nos demais experimentos, o uso de memórias de 16 bits fez com que os resultados fossem um pouco melhores. Neste caso, o grau de acerto ficou entre 73,6%

e 76,4% e o desvio padrão variou entre 3,5% e 6,9%. Embora estes resultados não tenham sido muito bons, foram melhores que os obtidos ao diferenciar as seis categorias. Juntando-os com os excelentes resultados obtidos nos experimentos que diferenciavam somente os subtipos, observa-se que pode ser realizada a categorização em etapas. Certamente, ao se segregar as amostras de acordo com o medicamento, bons resultados serão obtidos com a WiSARD.

4.1.7 Comparação Entre as Codificações

No terceiro Capítulo há menção de que para os primeiros experimentos as codificações apresentaram uma variação de 5,6%. Comentou-se também que por ser muito difícil prever como esta variação evoluiria, optou-se por continuar realizando todos os experimentos com todas as codificações. Agora se possui os resultados dos 11 experimentos realizados com cada forma de representação dos aminoácidos, e pode-se, portanto, analisar e comparar o desempenho de cada uma delas.

Na tabela 4.10 há os maiores e os menores valores obtidos para cada configuração da WiSARD, bem como, a diferença entre eles. A numeração utilizada, segue a relação presente na tabela 3.7. Observando-se estes dados percebe-se que as diferenças entre as codificações foram sempre muito pequenas, tendo uma variação média de 5,3%. A coluna “SUBTIPO”, que apresentou média de 6%, se refere às análises nas quais, independentemente das amostras resistentes e *naives* terem sido treinadas separadamente ou não, as respostas eram agrupadas pelo subtipo do vírus.

No ANEXO I pode-se observar que essas escalas geraram números binários muito diferentes. Porém, pelos resultados apresentados ao longo deste Capítulo, tornou-se previsível que para uma mesma configuração de WiSARD as codificações apresentariam resultados semelhantes. Pelo exposto na tabela 4.10, nota-se que apenas para a oitava configuração da WiSARD a diferença foi de 8,6%, o que ainda pode ser considerado um valor baixo. A décima e a décima primeira configurações não possuem valor em relação ao subtipo, pois se referem aos experimentos onde se utilizou apenas amostras do subtipo B.

Tabela 4.10 Diferenças máximas entre codificações por configuração da WiSARD.

WISARD	6 GRUPOS			SUBTIPO		
	MAIOR	MENOR	DIFERENÇA	MAIOR	MENOR	DIFERENÇA
1	68,9%	64,6%	4,2%	78,2%	72,5%	5,6%
2	82,0%	77,8%	4,1%	91,5%	85,5%	6,1%
3	85,3%	79,8%	5,6%	94,7%	87,8%	6,9%
4	90,9%	84,4%	6,5%	99,6%	96,5%	3,0%
5	80,8%	73,7%	7,0%	98,2%	92,2%	6,0%
6	69,0%	62,0%	7,1%	92,9%	86,6%	6,3%
7	69,1%	65,5%	3,7%	94,0%	89,5%	4,4%
8	91,1%	82,6%	8,6%	91,1%	82,6%	8,6%
9	93,5%	86,5%	7,0%	93,5%	86,5%	7,0%
10	74,9%	73,2%	1,7%	-	-	-
11	76,4%	73,7%	2,8%	-	-	-
MÉDIA			5,3%			6,0%

Outro ponto importante é que nos experimentos iniciais não existia uma codificação predominantemente melhor que as demais. Sendo assim, as diferenças apresentadas na tabela 4.10 poderiam em um momento ser consequência de um melhor desempenho da codificação BIN20, e em outro, ter sido decorrente de um resultado muito pior para esta mesma codificação.

Por esta possível característica da relação entre as codificações, calculou-se a média de cada uma. Desta forma, uma codificação que fosse sempre muito eficiente teria naturalmente uma média muito acima das demais. O mesmo vale caso houvesse uma que possuísse sempre um resultado inferior.

Na tabela 4.11, vê-se que não houve nenhuma codificação com média muito acima, ou muito abaixo das demais. Além disso, o desvio padrão referente a estas médias é mais um indicador que comprova a pequena variação existente. Nota-se que para as análises que consideram os seis grupos, a média da taxa de acerto varia entre 76% e 79,4%. Ao se considerar apenas os subtipos, esta taxa varia entre 89,9% e 93,1%. Isso resulta em uma variação máxima para a eficiência das codificações menor que 3,5%.

Complementando estes dados, vê-se que as codificações que usam somente dois terços da escala do código Gray obtiveram resultados melhores na maioria dos experimentos. Porém, a diferença entre essas codificações e as que usaram toda a escala, tanto em binário quanto em código Gray, foi muito pequena. Em relação à informação biológica representada pelas escalas, a hidrofobicidade foi a que gerou melhores resultados. Em seguida, foi obtida a massa molecular, e contrariando-se as expectativas, a junção dessas duas informações resultou em grau de acerto inferior.

Tabela 4.11 Médias da acurácia das codificações.

CÓDIGO	6 GRUPOS		SUBTIPO	
	MÉDIA	DESVIO	MÉDIA	DESVIO
BIN20	76,8%	3,6%	91,1%	1,9%
HIDROBIN	78,6%	3,9%	92,8%	2,0%
HIDROGRAY	78,3%	4,0%	92,2%	2,0%
HIDROGRAY23	79,4%	4,0%	93,1%	1,8%
MMBIN	78,7%	3,9%	93,1%	1,8%
MMGRAY	76,0%	4,5%	90,1%	3,1%
MMGRAY23	78,8%	4,3%	92,8%	1,9%
HIDROMMBIN	76,3%	4,6%	89,9%	2,5%
HIDROMMGRAY	76,4%	4,3%	90,4%	2,7%
HIDROMMGRAY23	77,5%	4,5%	90,9%	2,8%
DESVIO PADRÃO	1,2%		1,3%	

Tais fatos podem ter ocorrido por não ter sido feito um controle rígido em relação à codificação e a distância de Hamming entre os aminoácidos. Sendo assim, dependendo do número atribuído para aquele aminoácido, a combinação das duas informações pode ter gerado valores binários menos adequados.

A hidrofobicidade espalhada em dois terços do código Gray foi a que apresentou, em média, melhores resultados. Porém, conclui-se que a realização dos experimentos com todas as 10 formas de representação binária foi importante. Embora existam algumas diferenças, todas elas possuem praticamente o mesmo grau de eficiência ao representar numericamente os aminoácidos. Em futuros trabalhos, pode ser interessante não só analisar novas informações químicas, como também dar mais ênfase à distância de Hamming entre as representações binárias. Desta forma, a codificação gerada pode vir a ser aperfeiçoada.

4.1.8 Comparação Entre Memórias de 8 bits e 16 bits

Originalmente, os experimentos usando memórias de 16 bits foram realizados somente para verificar se a WiSARD seria capaz de trabalhar com entradas maiores. Entretanto, da obtenção dos primeiros resultados verificou-se que esta configuração apresentava melhora na acurácia. Em decorrência deste fato, passou-se a realizar os experimentos seguintes também com redes de 16 bits.

Tabela 4.12 Comparação entre experimentos com memórias de 8 bits e de 16 bits.

			Memórias de 8 bits		Memórias de 16 bits		Diferença	
	CÓDIGO	N AMOSTRAS	MÉDIA	DESVIO	MÉDIA	DESVIO	MÉDIA	DESVIO
RECONHECIMENTO 6 GRUPOS	BIN20	4560	64,6%	4,6%	65,5%	3,4%	-0,009	0,012
	HIDROBIN	4560	65,7%	4,9%	68,6%	5,1%	-0,029	-0,002
	HIDROGRAY	4560	66,9%	4,0%	68,5%	5,7%	-0,016	-0,016
	HIDROGRAY23	4560	69,0%	4,5%	68,7%	7,0%	0,004	-0,025
	MMBIN	4560	66,9%	5,1%	69,1%	5,9%	-0,022	-0,007
	MMGRAY	4560	62,0%	7,2%	66,4%	8,0%	-0,044	-0,008
	MMGRAY23	4560	67,7%	5,4%	68,8%	7,1%	-0,011	-0,017
	HIDROMMBIN	4560	62,8%	5,2%	66,2%	5,3%	-0,034	-0,001
	HIDROMMGRAY	4560	63,5%	5,0%	66,4%	6,9%	-0,028	-0,018
	HIDROMMGRAY23	4560	64,3%	4,7%	67,1%	5,8%	-0,029	-0,011
RECONHECIMENTO SUBTIPO	BIN20	4560	89,1%	2,4%	89,7%	1,6%	-0,006	0,008
	HIDROBIN	4560	88,8%	2,8%	93,1%	2,6%	-0,043	0,002
	HIDROGRAY	4560	89,4%	2,3%	90,7%	3,1%	-0,013	-0,008
	HIDROGRAY23	4560	91,0%	1,7%	93,5%	2,2%	-0,025	-0,005
	MMBIN	4560	91,1%	2,2%	93,5%	1,9%	-0,024	0,002
	MMGRAY	4560	83,5%	5,4%	87,8%	4,7%	-0,043	0,006
	MMGRAY23	4560	89,8%	3,0%	92,8%	2,4%	-0,029	0,006
	HIDROMMBIN	4560	82,6%	2,9%	86,5%	4,0%	-0,039	-0,011
	HIDROMMGRAY	4560	85,4%	3,8%	88,4%	3,7%	-0,031	0,001
	HIDROMMGRAY23	4560	84,2%	2,9%	88,6%	4,2%	-0,044	-0,013
RECONHECIMENTO B RESISTENTE E B *NAIVE*	BIN20	1584	74,9%	6,4%	73,7%	4,4%	0,012	0,021
	HIDROBIN	1584	73,2%	6,3%	75,4%	5,3%	-0,022	0,011
	HIDROGRAY	1584	74,2%	7,3%	76,4%	6,1%	-0,022	0,012
	HIDROGRAY23	1584	74,1%	6,7%	76,1%	5,4%	-0,020	0,013
	MMBIN	1584	73,5%	4,9%	74,2%	5,7%	-0,007	-0,008
	MMGRAY	1584	73,9%	2,8%	73,9%	3,6%	0,000	-0,008
	MMGRAY23	1584	73,2%	5,2%	74,4%	5,9%	-0,012	-0,007
	HIDROMMBIN	1584	73,7%	5,9%	74,6%	7,0%	-0,010	-0,011
	HIDROMMGRAY	1584	74,1%	7,2%	73,8%	4,9%	0,003	0,023
	HIDROMMGRAY23	1584	73,8%	6,7%	75,9%	6,1%	-0,021	0,006
RECONHECIMENTO SUBTIPO SEM AGRUPAR ANTES DO TREINAMENTO	BIN20	4560	91,2%	1,7%	91,2%	2,5%	0,000	-0,008
	HIDROBIN	4560	91,4%	2,5%	92,9%	2,4%	-0,015	0,002
	HIDROGRAY	4560	91,1%	2,2%	92,3%	2,7%	-0,012	-0,005
	HIDROGRAY23	4560	92,9%	1,6%	93,3%	2,9%	-0,004	-0,012
	MMBIN	4560	92,5%	1,9%	94,0%	1,3%	-0,015	0,006
	MMGRAY	4560	86,6%	4,1%	90,2%	5,0%	-0,036	-0,009
	MMGRAY23	4560	91,6%	2,5%	93,5%	2,2%	-0,019	0,004
	HIDROMMBIN	4560	87,0%	2,9%	89,5%	2,6%	-0,025	0,003
	HIDROMMGRAY	4560	87,3%	4,3%	91,1%	3,2%	-0,037	0,011
	HIDROMMGRAY23	4560	88,3%	3,3%	90,8%	4,4%	-0,025	-0,010
	MÉDIA	-	79,2%	4,2%	81,2%	4,4%	-0,020	-0,002

Em um total de quarenta casos que utilizaram tanto memórias de 8 quanto de 16 bits, houve somente quatro experimentos nos quais o uso de memórias de 8 bits foi melhor. Dentre estes quatro casos, um obteve diferença de 1,2% e para os demais a diferença não chegou a 0,4%, o que é muito baixo. Na tabela 4.12 pode-se observar

que o uso de memórias de 16 bits apresentou uma melhora média de 2,3%. Embora seja uma diferença média pequena, houve muitos casos em que essa diferença passou de 4%. Pelo fato disto estar relacionado a uma acurácia que passa de 84,2% para 88,6%, esta melhora pode ser considerada muito significativa.

É importante ressaltar que o tamanho da memória tem impacto direto no número de memórias utilizadas e consequentemente na capacidade de generalização da rede. Caso o tamanho da memória fosse demasiadamente aumentado, a rede poderia perder eficiência e por isso não foram feitos experimentos com reder utilizando memórias maiores.

4.2 Análise em Relação à Técnica do *Bleaching*

Inicialmente, esta técnica foi utilizada de modo a parar o reconhecimento assim que ocorresse o desempate. Desde os primeiros resultados a maioria possuía valor de *Bleaching* acima de zero. Logo, já era possível se perceber a sua importância. Este fato também chamou a atenção de que certamente haveria uma grande semelhança entre as categorias.

Tal semelhança afetaria em muito a confiança da WiSARD, que ao ser analisada confirmou-se que os valores obtidos eram sempre muito baixos. Isto ocorria devido aos grupos estarem sendo escolhidos com uma pequena diferença entre os discriminadores. Muitas vezes, esta diferença era de apenas 1 ou 2 pontos, o que resultava numa confiança menor que 2%.

Em decorrência desta baixa taxa de confiança, resolveu-se passar à verificação do que acontecia com os resultados caso o *Bleaching* continuasse aumentando até que nenhum grupo pontuasse, mesmo que o desempate já houvesse ocorrido. Para ser possível comparar a categoria e a confiança encontradas inicialmente com os valores obtidos ao se deixar o Bleaching continuar crescendo após o desempate, a resposta passou a ser composta por:

- Valores obtidos assim que ocorresse o desempate;
- Grupos obtidos após o desempate, e maior confiança possível entre eles;
- Valores e grupos obtidos para a maior confiança possível.

Com essas três informações é possível realizar as seguintes análises: número de amostras que resultariam em empate sem o uso do *Bleaching* e consequentemente

em escolha aleatória, taxa de utilização do *Bleaching*, taxa de acerto com *Bleaching* e número de amostras onde o uso do *Bleaching* levou a um aumento na confiança.

4.2.1 Escolhas Aleatórias Sem *Bleaching*

Na primeira parte da resposta, havia então o valor mínimo do *Bleaching* que foi necessário para o desempate dos discriminadores. Quando este valor era zero, significava que não ocorreu empate. Portanto, este resultado seria encontrado mesmo sem o uso desta técnica. Em contrapartida, um valor igual a nove, por exemplo, significava que para todos os valores de *Bleaching* de zero a oito, existiram dois ou mais discriminadores com o mesmo número de pontos. A tabela 4.13 exibe os percentuais das amostras de cada experimento que obtiveram zero como valor do *Bleaching* de desempate.

Tabela 4.13 Percentuais de amostras onde não ocorreram empates.

WISARD	TOTAL DE AMOSTRAS	BIN20	HIDROBIN	HIDRO GRAY	HIDRO GRAY23	MMBIN	MMGRAY	MMGRAY23	HIDRO MMBIN	HIDRO MMGRAY	HIDROMM GRAY23	MÉDIA
1	1205	15,7%	41,4%	43,6%	44,4%	34,1%	35,4%	36,9%	37,2%	36,2%	35,8%	36,1%
2	1205	40,6%	78,0%	74,2%	35,4%	74,4%	73,5%	74,9%	62,5%	67,1%	62,3%	64,3%
3	1205	35,4%	77,8%	35,4%	81,6%	73,7%	74,5%	77,4%	69,1%	70,0%	70,0%	66,5%
4	4482	7,7%	41,2%	58,8%	36,0%	34,0%	29,8%	35,6%	29,5%	29,8%	32,0%	33,4%
5	114	78,0%	94,0%	91,1%	93,9%	86,1%	89,5%	88,6%	89,5%	90,5%	90,5%	89,2%
6	4560	71,1%	87,6%	86,9%	89,6%	84,3%	85,8%	87,5%	83,2%	85,0%	83,3%	84,4%
7	4560	70,3%	91,0%	91,6%	91,6%	87,7%	88,5%	89,9%	87,5%	88,8%	88,2%	87,5%
8	4560	86,2%	97,4%	97,5%	98,3%	97,8%	96,0%	98,1%	93,9%	95,6%	94,6%	95,5%
9	4560	88,3%	98,8%	97,9%	98,6%	98,9%	97,2%	98,6%	96,4%	96,9%	96,8%	96,8%
10	1584	47,4%	73,3%	76,1%	74,8%	71,3%	71,3%	72,2%	69,1%	71,7%	72,2%	69,9%
11	1584	48,3%	79,7%	81,4%	82,0%	74,2%	78,9%	79,5%	73,1%	76,4%	80,1%	75,3%
MÉDIA		53,5%	78,2%	75,9%	75,1%	74,2%	74,6%	76,3%	71,9%	73,5%	73,3%	72,6%

Observando-se o valor referente aos experimentos que usaram codificação Bin20 e balanceamento pelo subtipo B resistente, ou seja, WiSARD de número 4, pode-se ver que este experimento só chegaria a um resultado determinístico em apenas 7,7% das amostras. Embora pareça ser um caso isolado, é na verdade a combinação das duas situações que apresentaram menores médias.

Em grande parte dos experimentos essa média fica acima dos 80%, porém, comparando-se as codificações, é notório que a BIN20 possui o pior desempenho. Sua média é 20% inferior a praticamente todas as demais. Já da análise das redes, tem-se que a quarta e a primeira configurações usadas tiveram os menores desempenhos. Para todos os experimentos com estas configurações, apenas um terço chegaria a um resultado determinístico.

Os resultados melhoram um pouco para as configurações referentes às redes de números dois, três, dez e onze. Porém, para estes experimentos, uma a cada três amostras teria resultado aleatório se não fosse o uso do *Bleaching*. Apenas nos casos onde a WiSARD foi treinada com lotes pequenos é que a média manteve-se acima de 80%, tendo em alguns poucos casos, chegado próximo de 100%.

Ocorre que, dependendo da codificação usada, a aleatoriedade da resposta pode ter uma variação grande. Mesmo com a configuração de maior média, o percentual de amostras não determinísticas pode chegar a mais de 10%. Isto significa que uma em cada dez amostras teria seu desempate realizado por sorteio.

Nas redes de um a quatro, o valor máximo do *Bleaching* poderia chegar até o número de treinamentos realizados com a categoria B, no caso, 647. Para as de número dez e onze, este valor poderia chegar até 99, que era o tamanho do lote definido para estes treinamentos. Nos demais casos o lote tinha apenas 19 amostras de cada grupo, ou 38 quando as amostras foram agrupadas pelo subtipo. Portanto, o valor máximo do *Bleaching* era muito menor que nos casos anteriores. Na comparação entre o tamanho do lote e o valor máximo que o *Bleaching* poderia atingir com os valores da tabela 4.13, vê-se que há uma relação direta entre eles, de onde se conclui que quanto maior o lote, mais difícil foi para a WiSARD chegar a um resultado sem o uso do *Bleaching*.

4.2.2 Taxa de Acerto Com o Uso do *Bleaching*

Em complemento à análise anterior, observa-se a taxa de acerto da rede em relação à utilização do *Bleaching*. Na tabela 4.14 tem-se as taxas de acerto sem o *Bleaching* e na tabela 4.15, a diferença entre elas e os resultados apresentados ao longo do tópico 4.1, que se encontram resumidos na tabela 2 do ANEXO II.

Embora existam certas variações, é claramente visível que os dados da tabela 4.14 são muito semelhantes aos da 4.13. Isto ocorre porque esta medida é uma especificidade da medida anterior. Enquanto na tabela 4.13 se mostrou o percentual

de amostras que possuíam valor de *Bleaching* inicial igual a zero, nesta, o foco foi dado nos valores que possuíam essa característica e também obtiveram resposta correta. Este levantamento é importante, pois serve de base para um melhor entendimento dos dados da tabela 4.15, que exibe a variação da porcentagem dos resultados corretos ao usar o *Bleaching* e ao não utilizá-lo.

Tabela 4.14 Taxa de acerto determinístico sem o *Bleaching*.

WISARD	TOTAL DE AMOSTRAS	BIN20	HIDROBIN	HIDRO GRAY	HIDRO GRAY23	MMBIN	MMGRAY	MMGRAY23	HIDRO MMBIN	HIDRO MMGRAY	HIDROMM GRAY23	MÉDIA
1	1205	13,6%	34,6%	36,0%	36,7%	28,8%	29,9%	31,1%	31,9%	30,5%	30,5%	30,4%
2	1205	32,6%	64,5%	61,8%	29,9%	62,6%	60,9%	63,1%	50,1%	55,5%	50,3%	53,1%
3	1205	29,9%	67,8%	29,9%	71,1%	65,4%	62,6%	66,2%	60,2%	60,2%	61,2%	57,4%
4	4482	7,7%	41,1%	58,7%	36,0%	33,9%	29,7%	35,5%	29,4%	29,8%	32,0%	33,4%
5	114	63,9%	72,2%	71,1%	74,7%	66,8%	68,5%	72,9%	70,4%	66,7%	71,1%	69,8%
6	4560	50,4%	58,6%	59,3%	63,4%	58,3%	54,1%	60,4%	54,0%	55,8%	55,2%	56,9%
7	4560	49,4%	63,8%	64,5%	64,2%	61,7%	59,4%	62,7%	59,3%	60,1%	60,8%	60,6%
8	4560	77,3%	86,9%	87,4%	89,6%	89,4%	81,4%	88,5%	77,9%	82,6%	80,0%	84,1%
9	4560	80,1%	92,6%	89,3%	92,5%	92,8%	86,2%	91,8%	83,9%	86,3%	86,1%	88,2%
10	1584	40,2%	58,3%	61,5%	61,4%	57,5%	56,7%	58,7%	56,2%	58,5%	58,6%	56,8%
11	1584	40,9%	62,9%	65,8%	65,9%	59,5%	61,1%	62,9%	59,4%	60,6%	65,5%	60,5%
MÉDIA		44,2%	63,9%	62,3%	62,3%	61,5%	59,1%	63,1%	57,5%	58,8%	59,2%	59,2%

Saber que 7,7% ou 98,8% dos resultados foram encontrados de modo determinísticos gera uma ideia próxima do desempenho da rede. Porém, nada impede que a rede de menor porcentagem tenha obtido uma acurácia maior. O mesmo não pode ser dito dos dados da tabela 4.15, que permite uma verdadeira noção de como o uso do *Bleaching* garante a melhora os resultados.

Comparando-se a taxa de acerto de rede sem, e com o uso do *Bleaching*, tem-se que, embora existam casos em que a taxa varia menos de 1%, também existem casos onde tal variação atinge 76,6 pontos percentuais. Neste caso, a utilização desta técnica foi responsável por fazer o percentual de resultados determinísticos e corretos passar de 7,7% para 84,4%, o que significa um incrível aumento de quase 1000%.

Tabela 4.15 Diferença entre a taxa de acerto ao utilizar o *Bleaching* e entre não utilizá-lo.

WISARD	TOTAL DE AMOSTRAS	BIN20	HIDROBIN	HIDRO GRAY	HIDRO GRAY23	MMBIN	MIMGRAY	MMGRAY23	HIDRO MMBIN	HIDRO MIMGRAY	HIDROMM GRAY23	MÉDIA
1	1205	51,1%	33,0%	32,9%	30,5%	36,0%	34,9%	33,9%	35,3%	34,2%	37,0%	35,9%
2	1205	45,6%	15,9%	19,8%	49,9%	19,4%	17,6%	17,4%	27,7%	22,7%	28,8%	26,5%
3	1205	49,9%	17,3%	55,4%	13,6%	19,7%	17,6%	16,8%	22,4%	22,6%	22,2%	25,8%
4	4482	76,7%	48,1%	31,5%	54,9%	55,5%	60,7%	54,9%	59,7%	59,8%	57,3%	55,9%
5	114	15,8%	5,2%	3,6%	3,6%	8,8%	6,1%	7,9%	6,2%	7,0%	7,9%	7,2%
6	4560	14,1%	7,1%	7,6%	5,6%	8,6%	7,9%	7,3%	8,9%	7,8%	9,1%	8,4%
7	4560	16,1%	4,8%	3,9%	4,5%	7,4%	7,0%	6,1%	6,9%	6,3%	6,4%	6,9%
8	4560	11,8%	1,9%	2,0%	1,4%	1,7%	2,1%	1,3%	4,7%	2,8%	4,2%	3,4%
9	4560	9,6%	0,5%	1,5%	1,0%	0,7%	1,6%	1,0%	2,5%	2,1%	2,6%	2,3%
10	1584	34,7%	14,9%	12,7%	12,7%	16,1%	17,3%	14,5%	17,5%	15,6%	15,2%	17,1%
11	1584	32,7%	12,5%	10,7%	10,1%	14,7%	12,8%	11,5%	15,2%	13,2%	10,4%	14,4%
MÉDIA		32,6%	14,7%	16,5%	17,1%	17,1%	16,9%	15,7%	18,8%	17,6%	18,3%	18,5%

Calculando a média geral, a utilização do *Bleaching* resulta numa melhora de 18,5 pontos percentuais. Considerando-se somente as amostras onde houve balanceamento, essa taxa de melhora cai para em média 8,7 pontos percentuais. Porém, ainda há um caso onde a melhora chega a 34 pontos, o que representa um aumento de eficiência da WiSARD de 85%, pois a acurácia passa de 40% sem o uso do *Bleaching*, para 74% com o uso da técnica.

4.2.3 Taxa de Uso do *Bleacnhig*

Esta análise representa o percentual de amostras onde, mesmo não sendo preciso o uso do *Bleaching*, isto foi feito de modo a buscar uma melhora na confiança do resultado apresentado. Além de garantir uma grande melhora na taxa de acerto, conforme visto no tópico anterior, ao usar tal técnica fixando as categorias encontradas logo após o desempate, esta pode colaborar em muito para o crescimento da confiança dos resultados.

Tabela 4.16 Taxa de utilização do *Bleaching*.

WISARD	TOTAL DE AMOSTRAS	BIN20	HIDROBIN	HIDRO GRAY	HIDRO GRAY23	MMBIN	MMGRAY	MMGRAY23	HIDRO MMBIN	HIDRO MMGRAY	HIDROMM GRAY23	MÉDIA
1	1205	98,8%	96,1%	96,8%	94,5%	97,7%	98,3%	96,0%	95,4%	96,3%	98,1%	96,8%
2	1205	94,2%	89,9%	91,2%	95,0%	89,1%	89,0%	90,7%	92,0%	91,9%	92,3%	91,5%
3	1205	95,0%	91,5%	95,0%	94,5%	93,2%	94,4%	94,0%	91,9%	93,3%	93,3%	93,6%
4	4482	98,3%	93,0%	91,5%	96,7%	95,2%	96,8%	96,4%	97,2%	95,7%	95,1%	95,6%
5	114	80,7%	66,7%	66,7%	65,1%	76,4%	66,7%	70,3%	66,7%	68,6%	71,1%	69,9%
6	4560	71,7%	73,1%	72,6%	72,8%	72,9%	65,5%	72,3%	68,3%	67,6%	68,0%	70,5%
7	4560	75,6%	70,6%	67,7%	69,4%	69,8%	66,7%	70,9%	63,2%	69,4%	69,1%	69,2%
8	4560	91,8%	91,4%	92,3%	93,7%	93,4%	88,1%	92,1%	87,8%	89,5%	89,2%	90,9%
9	4560	92,8%	94,5%	93,8%	94,0%	95,2%	90,0%	94,9%	90,6%	90,4%	91,8%	92,8%
10	1584	96,7%	96,3%	96,4%	96,4%	98,1%	98,2%	96,9%	96,3%	97,4%	97,4%	97,0%
11	1584	96,7%	92,3%	93,1%	94,9%	95,9%	92,7%	92,5%	92,6%	94,5%	94,4%	94,0%
MÉDIA		90,2%	86,9%	87,0%	87,9%	88,8%	86,0%	87,9%	85,6%	86,8%	87,2%	87,4%

Na tabela 4.16 observa-se que com o uso desta técnica nessa forma, 87,4% das amostras obtêm valor de *Bleaching* maior que 0. Somente pouco mais de 10% dispensou o uso desta técnica, pois obteriam a melhor confiança sem utilizá-la. Embora existam experiências onde 37% das amostras dispensaram o uso do *Bleaching*, há também muitas em que esta utilização foi superior a 90%, inclusive chegando a 98%.

Diferentemente da análise feita no tópico 4.2.1, não se vê claramente uma relação entre o tamanho do lote e a taxa de utilização do *Bleaching*. Porém, as redes cinco, seis e sete, que são as de menor lote, possuem médias muito semelhantes entre si e muito inferiores as demais.

4.2.4 Aumento da Confiança

Semelhante ao que foi feito com a taxa de acerto, a análise da taxa de utilização do *Bleaching* será complementada observando-se a porcentagem das amostras onde realmente a confiança do resultado teve aumento em decorrência do uso desta técnica. Na tabela 4.17 pode-se ver que em média 84% do total das amostras conseguiram melhorar a confiança. Para esta análise, caso forem levados em conta apenas os experimentos balanceados por lote, essa média praticamente não apresentaria variação. Por estes dados determina-se que dos 87,4% das amostras

totais que utilizaram o *Bleaching* buscando uma melhora da confiança, somente 3,4% não registrou tal aumento.

Tabela 4.17 Percentual das amostras que obtiveram aumento da confiança.

WISARD	TOTAL DE AMOSTRAS	BIN20	HIDROBIN	HIDRO GRAY	HIDRO GRAY23	MMBIN	MMGRAY	MMGRAY23	HIDRO MMBIN	HIDRO MMGRAY	HIDROMM GRAY23	MÉDIA
1	1205	85,1%	89,4%	93,8%	89,2%	92,4%	89,0%	93,2%	88,2%	91,6%	90,0%	90,2%
2	1205	84,4%	87,0%	88,8%	86,1%	88,2%	85,5%	88,9%	88,1%	86,6%	88,5%	87,2%
3	1205	86,1%	90,2%	86,1%	93,4%	92,9%	93,1%	93,0%	89,6%	90,4%	91,2%	90,6%
4	4482	84,1%	89,2%	88,9%	90,0%	90,2%	91,3%	90,9%	89,9%	89,3%	89,0%	89,3%
5	114	74,5%	66,7%	65,8%	65,1%	76,4%	64,9%	69,4%	64,9%	66,1%	70,2%	68,4%
6	4560	58,9%	70,5%	69,9%	70,6%	69,8%	62,3%	70,3%	62,8%	63,9%	63,4%	66,2%
7	4560	63,7%	68,0%	65,2%	67,6%	67,6%	64,1%	69,0%	59,6%	66,4%	65,9%	65,7%
8	4560	90,2%	91,0%	92,1%	93,6%	93,3%	86,8%	91,8%	86,9%	88,7%	88,5%	90,3%
9	4560	91,5%	94,2%	93,5%	93,8%	95,2%	89,4%	94,9%	90,2%	89,7%	91,4%	92,4%
10	1584	91,1%	94,8%	95,6%	95,2%	96,8%	97,9%	96,2%	93,1%	96,2%	95,8%	95,3%
11	1584	91,5%	91,6%	92,1%	94,5%	94,6%	91,8%	91,4%	89,8%	92,5%	93,5%	92,3%
MÉDIA		81,9%	84,8%	84,7%	85,4%	87,0%	83,3%	86,3%	82,1%	83,8%	84,3%	84,3%

4.2.5 Análise do Último Valor de *Bleaching*

Nas análises anteriores, compararam-se os resultados obtidos com e sem o uso do *Bleaching*. Porém, é interessante verificar também o que ocorre quando se deixa o valor do *Bleaching* aumentar até que nenhum dos discriminadores pontue.

Em relação à acurácia foi observado que tal modo de usar o *Bleaching* tinha um desempenho muito inferior aos anteriores. Em alguns casos, este fato pode ser facilmente entendido. Para todos os experimentos sem balanceamento a resposta encontrada foi B resistente, e a segunda opção C resistente. Estes eram os grupos com o maior número de amostras, portanto, foram os únicos que continuaram pontuando na medida em que o *Bleaching* aumentava. Percebe-se assim, que caso as amostras não estejam balanceadas, o resultado obtido sempre será o grupo com maior número de treinamentos realizados, o que inclusive condiz com o que deve ser esperado.

Os experimentos que estavam balanceados pelo grupo B obtiveram resultados semelhantes, tendo também uma taxa de acerto muito baixa. Possivelmente isso ocorreu em decorrência do grande número de repetições das amostras do grupo F

naive. Esta foi a escolhida em quase todos os resultados. O mesmo se aplica ao experimento onde se buscou diferenciar B *naive* de B resistente, que teve como resultado quase absoluto o grupo B *naive*.

Os demais experimentos não apresentaram um único grupo como resposta. Porém, de maneira geral sempre existia um que aparecia como resposta para mais de 50% das amostras. Observou-se também que, independente da codificação utilizada, este grupo predominante era quase sempre o mesmo para aquela rede.

4.2.6 Resumo da Análise do *Bleaching*

Pelas análises feitas acima, conclui-se que a melhor maneira de utilização desta técnica é a fixação das categorias selecionadas no momento do desempate, bem como, continuar aumentando o *Bleaching* para descobrir a maior confiança. Sendo utilizado dessa forma esta técnica é muito eficaz e serve como um excelente complemento para a WiSARD.

Pode-se ver que tal técnica não só colabora dando uma maior exatidão à rede, tornando determinísticos reconhecimentos que antes seriam desempatados por sorteio, como também contribui para um aumento da confiança dos resultados. Sendo assim, embora torne o processo de reconhecimento mais lento, o *Bleaching* resolve o problema do baixo valor para a confiança, mencionado no tópico 4.1.1.4. Além disso, é importante perceber que houve uma importante melhora nos resultados tanto nos casos onde as amostras estavam desbalanceadas, quanto naqueles em que o mesmo número de amostras por grupo era usado nos treinamentos.

4.3 Melhora de Tempo de Reconhecimento

Conforme visto anteriormente, a utilização de todas as 99 posições da entrada deveria ser feita para os experimentos que se seguiam, mesmo que tal utilização impactasse em uma confiança menor e em um tempo de execução maior. No tópico anterior foi possível observar como a utilização do *Bleaching* foi fundamental no aumento da confiança nos resultados. Porém, deixá-lo aumentar até que apenas um dos discriminadores pontuasse tornou o reconhecimento mais demorado. Desta forma

passou a ser necessário buscar estratégias que tornassem o reconhecimento mais rápido.

Para melhorar o desempenho da rede, foram feitos três aperfeiçoamentos, que embora sejam muito simples, diminuíram em muito o tempo de resposta da WiSARD:

- Uso de VG-RAM nas memórias da WiSARD;
- Uso de vetor para armazenar os pontos de cada padrão durante reconhecimento;
- Cálculo do *Bleaching* somente nos valores que acarretariam mudança na pontuação.

4.3.1 Uso de VG-RAM nas Memórias da WiSARD

No Capítulo 2, foram descritas duas formas de rede neural sem peso, a WiSARD e a VG-RAM. Os critérios para treinamento e reconhecimento utilizados seguem exatamente a lógica utilizada pela WiSARD. Por isto, até o momento, sempre que este trabalho se referia à rede neural, isto era feito focando na utilização desta rede.

Conforme exposto anteriormente, uma WiSARD possui memórias que armazenam 0's e 1's no endereço correspondente ao pedaço da entrada que está sendo lido. Sendo assim, uma memória de, por exemplo, 8 bits possuirá 256 (2^8), espaços para armazenar 256 possíveis valores da entrada. Vale lembrar também que desses endereços, o esperado é que grande parte dos bits fiquem zerados. Pelo fato de ter se utilizado o *Bleaching*, trabalhar com vetores de 256 bits não era suficiente. Não seria guardada apenas a informação do padrão ter sido apresentado ou não, e sim, o número de vezes que ele foi apresentado. Em decorrência da categoria mais abundante na base de dados disponível possuir 747 amostras, era preciso separar três bits para guardar o valor do *Bleaching*. Desta forma, vetores de 768 bits (256 x 3) seriam suficientes.

Tomando como exemplo os casos em que se utilizaram todas as 99 posições da protease, seria preciso 258 memórias para cada um dos 6 discriminadores e cada uma dessas memórias teria 768 bits, o que leva a um total de 1.188.864 (6 x 258 x 768) bits ou 145,125 Kbytes. Ao invés de manter todos esses bits, nos quais a grande maioria certamente possuirá valor 0, pode-se guardar somente o valor treinado para aquela memória e o número de vezes que este padrão foi apresentado. Ficando

assim, apenas com blocos de 8 bits associados a um número inteiro relativo ao número de vezes que aquele padrão foi treinado.

Cada memória da WiSARD passa então a funcionar como uma VG-RAM. Esta rede interna a memória da WiSARD, ao invés de armazenar uma categoria para um determinado padrão apresentado, associa ele a um valor de *Bleaching*. Neste caso, treinar esta VG-RAM significa aumentar em uma unidade o valor associado ao padrão.

Em um primeiro momento isso pode parecer mais lento, pois a memória não será acessada diretamente. Porém, ocorrerá uma melhora no desempenho devido ao fato de menos informações estarem sendo carregadas para a memória do programa, o que o torna mais leve. Além disso, o esperado é que grande parte das memórias possuam poucos padrões com elevado valor de *Bleaching* associados a eles, o que de fato foi observado.

4.3.2 Uso de Vetor Para Armazenar os Pontos de Cada Padrão Durante o Reconhecimento

Das três melhorias realizadas, essa certamente é a mais simples de ser feita, provavelmente a mais intuitiva, e a que causou diminuição mais significativa no tempo do reconhecimento, onde a rede apresentava maior lentidão.

Uma das características mais interessantes de WiSARD é que ela realiza o reconhecimento apresentando o padrão lido a cada um dos discriminadores uma única vez e compara a soma dos pontos obtidos por cada categoria para saber qual delas é a resposta a ser apresentada. Desta forma, o tempo de resposta é constante e sempre muito baixo. Porém, isso deixa de ser verdade ao utilizar o *Bleaching*. Com esta técnica o tempo de resposta passa a ser multiplicado pelo valor que foi necessário utilizar para chegar ao resultado determinístico. Caso seja utilizada conforme recomendado no tópico 4.2, o tempo aumenta ainda mais, pois passa a variar de acordo com o número de amostras treinadas. Para os experimentos onde não houve nenhuma forma de balanceamento, por exemplo, o tempo de resposta poderia ser até 673 vezes maior.

Para reduzir esse tempo de reconhecimento passou-se a criar em tempo de execução um vetor para cada discriminante da WiSARD. Tal vetor tinha capacidade para armazenar os valores gravados nas memórias relativas aquela entrada. Assim, as memórias voltariam a ser acessadas somente uma única vez, quando o *Bleaching* era igual a zero (valor inicial). Para os demais valores, bastaria contar quais posições

desses vetores eram maiores que este valor. Considerando os experimentos com todas as 99 posições da entrada, por exemplo, cada categoria precisaria de 1 vetor com capacidade para guardar 258 inteiros, ao invés das 258 memórias de 2^8 bits (256). Sendo assim, após o uso do primeiro valor de *Bleaching*, trabalhar-se-ia com uma estrutura de dados 256 vezes menor.

É importante lembrar que neste trabalho utiliza-se VG-RAM dentro das memórias da WiSARD. Embora não se possua valores tão bem definidos, sabe-se que no melhor dos casos, trabalhar-se-ia com 258 memórias cada uma contendo 8 bits e mais um número inteiro. Considerando a base de dados utilizada, este número pode precisar de até 9 bits. No melhor dos casos, ter-se-ia então uma estrutura de 18,14 Kbits. Calculando o tamanho desse vetor, percebe-se que ele pode chegar a 2,25 Kbits (256 x 9), o que acarreta em uma estrutura no mínimo cerca de 8 vezes menor. Além disso, ao guardar os valores encontrados em um vetor, só seria necessário procurar nas memórias da VG-RAM o padrão apresentado uma única vez e somente isto já diminui bastante o esforço computacional envolvido no reconhecimento desta rede.

4.3.3 Cálculo do *Bleaching* Somente nos Valores que Acarretariam Mudança na Pontuação.

Conforme explicado no segundo Capítulo, durante o processo de reconhecimento, o valor do *Bleaching* aumenta sempre de uma unidade. Para cada novo valor, os pontos de todas as memórias são recalculados. Caso o valor armazenado na memória não seja maior que o *Bleaching*, aquela memória para de pontuar.

Com base no exemplo que se segue será demonstrado como o uso do *Bleaching* pode ser otimizado. Supondo uma rede com três memórias para cada discriminador, tem-se que para uma determinada entrada o discriminador “A” apresenta valores 2, 1 e 8 e o discriminador “B” 1, 0 e 9. Quando o *Bleaching* assumir os valores 0, 1 e 2, o discriminador “A” fará respectivamente 3, 2 e 1 pontos, enquanto o “B” somará 2, 1 e 1. Em seguida, para os valores entre 3 e 7, ambos os discriminadores não mudarão seus pontos. Eles só voltarão a ter seus valores alterados quando o *Bleaching* for igual a 8 e a 9.

Neste exemplo, é fácil perceber que somente para os valores: 0; 1; 2; 8 e 9 as memórias precisariam ser verificadas. Para os demais, os pontos dos discriminadores

não serão alterados. Ao observar este fato, percebe-se que o reconhecimento poderia ser otimizado se somente fossem efetuados o recálculo dos pontos das memórias nos valores realmente necessários.

No exemplo anterior calcular-se-ia os pontos para os seguintes valores de *Bleaching*: 0, 1, 2, 8 e 9. Seria obtida assim uma melhora de 50% no desempenho, pois os pontos das memórias seriam verificados apenas cinco vezes ao invés de 10 conforme feito anteriormente. Mesmo que o desempate ocorra para um valor pequeno, pelo *Bleaching* estar sendo usado até que somente um dos discriminadores esteja pontuando, essa implementação é significativa. Embora não seja possível assegurar a melhora obtida com essa abordagem percebeu-se que nos experimentos realizados obtive-se uma queda de mais de 60% no tempo de resposta sem haver mudança nos reconhecimentos das amostras.

4.4 “Mea Culpa” – O Que Eu Não Fiz e Que Ficou Faltando

Embora se tenha conseguido resultados muito bons, existem ainda algumas técnicas que poderiam ter sido utilizados de modo a contribuir para a obtenção de resultados ainda melhores. Além disso, é possível também aumentar o grau de reconhecimento da rede neural. A WiSARD pode ser capaz de reconhecer não somente os subtipos dos vírus, mas também de especificar se aquele vírus é ou não resistente a um medicamento especifico. Assim, tal rede aperfeiçoaria o tratamento atribuído a um determinado paciente.

4.4.1 Reconhecimento Cognitivo

A principal técnica para o aumento do grau de acerto da rede que pode ser incluída a seguir é o “reconhecimento cognitivo”. Tal abordagem visa agregar conhecimentos biológicos já obtidos cientificamente, com um reconhecedor automático de padrão, no caso a rede neural sem peso. Esse procedimento consiste em separar o reconhecimento em etapas, de modo a cada pedaço da entrada seja analisada em um momento diferente. Tais partes devem ser escolhidas usando conhecimento prévio acerca do problema, de modo a combiná-lo com o reconhecimento da WiSARD (GREGÓRIO 1996).

Inicialmente, uma parte dela é observada e são formuladas hipóteses sobre quais categorias podem ser escolhidas. Em seguida, é feito uma previsão sobre outra parcela da entrada para somente depois ser feita a observação dela. Essa segunda observação resulta num refinamento das hipóteses e em uma nova previsão. Depois dessa etapa, é verificada a ultima parte da entrada a qual deve confirmar a hipótese levantada. Caso nesse momento não haja confirmação, as hipóteses geradas anteriormente são refeitas, até que alguma delas seja confirmada (GREGÓRIO 1996).

4.4.2 Reconhecimento de Resistência a Medicamentos

Conforme mencionado no segundo Capítulo, uma das maiores dificuldades no tratamento do HIV se deve à falha terapêutica dos medicamentos antiretrovirais. Portanto, tem se buscado detectar os medicamentos para os quais aquele vírus apresentará resistência de modo que o tratamento possa ser mais eficiente. Sendo assim, a próxima etapa a ser galgada em relação ao poder de reconhecimento da rede neural é a especificação do remédio para o qual o paciente possui resistência.

Foi mostrado no Capítulo 1 que diversos estudos conseguem diferenciar pacientes em falha terapêuticas para medicamentos específicos. SILVA (2009), por exemplo, obtém bons resultados e também determina as posições de mutação com maior frequência para estas mutações. Porém, pelos resultados apresentados neste trabalho fica claro que certamente a utilização de redes neurais sem peso conseguirá o mesmo nível de acurácia, ou até mesmo resultados melhores. Esta rede conseguiu trabalhar com toda a estrutura da protease e deve ter bom desempenho ao lidar com a transcriptase reversa. Além disso, por obter um desvio padrão menor, os resultados seriam também mais estáveis. Deve-se lembrar ainda que o uso do reconhecimento cognitivo pode melhor não só experimentos já realizados com a WiSARD, como também ajudar na obtenção de bons resultados na identificação de resistências a medicamentos.

4.4.3 Usar a Integrase ou Mesmo a Transcriptase Reversa do Vírus

Outra possível forma de se conseguir melhorar a classificação dos subtipos do vírus e das resistências a um dado medicamento é utilizando outras informações além

da protease. Conforme explicado no Capítulo 2, o vírus possui três enzimas que são alvo dos medicamentos atuais: a transcriptase reversa, a protease e a integrase. Como a protease é que possui menor número de aminoácidos, ela foi escolhida para ser analisada em um primeiro momento. Além disso, ela já tinha apresentado bons resultados em diversos trabalhos anteriores.

Em relação às demais enzimas, os experimentos realizados usando memórias de 16 bits mostraram que certamente a WiSARD será capaz de trabalhar de maneira eficiente com estruturas maiores. Portanto, certamente será possível utilizar a integrase, que possui 212 aminoácidos, ou até mesmo a transcriptase reversa (RT) que é constituída de uma cadeia de 560 aminoácidos.

Estas enzimas, sobretudo a RT são estruturas mais completas e com mais informações sobre o vírus. Portanto, podem apresentar mais pontos de semelhanças entre amostras de mesmo comportamento e também possuir diferenças mais fáceis de serem detectadas. Além disso, isto permitirá a identificação das resistências aos demais medicamentos antiretrovirais e não apenas aos inibidores de protease. Pode-se inclusive utilizar todas as três enzimas em conjunto na rede neural, buscando-se uma relação entre elas.

5 Conclusão

5.1 Resumo

Nesse trabalho foi criada uma WiSARD, rede neural sem peso, capaz de reconhecer diferentes tipos de mutação do HIV-1. Foram feitos diversos experimentos, que visavam avaliar o desempenho da rede em relação ao grau de reconhecimento. Nestes experimentos também foi possível estudar de que forma a técnica de *Bleaching* deveria ser utilizada de modo a melhorar a acurácia.

A base de dados utilizada neste trabalho consistia em 1205 amostras da protease do vírus de pacientes infectados com o HIV-1. Estas amostras foram disponibilizadas pelo instituto de virologia da UFRJ. Elas são formadas pelo sequenciamento genético dos 99 aminoácidos que compõe a protease juntamente com o subtipo do vírus e sua característica quanto à resistência a algum medicamento inibidor da protease. Tais dados foram codificados em sequências binárias de 20 bits com base em características químicas dos aminoácidos. Nos experimentos realizados neste trabalho foram usadas a massa molecular e a hidrofobicidade isoladamente e também uma combinação dessas duas informações.

Para mapear os 20 diferentes aminoácidos presentes na protease os valores da hidrofobicidade e a massa molecular foram espalhados em quatro escalas: de 1 a 2^{20}, de 1 a $2/3*2^{20}$, de 1 a 2^{10} e de 1 a $2/3 *2^{10}$. Ao converter esses valores para binário, de forma simples, ou usando Gray code, geraram-se números de 10 e de 20 bits. Os de 10 bits foram combinados de modo a formar novos números de 20 bits na medida em que se concatenavam os 10 bits da massa molecular com os 10 bits da hidrofobicidade de um mesmo aminoácido. Além disso, também foi utilizada a matriz Bin20 para a codificação, totalizando 10 formas diferentes de representar os aminoácidos em sequências de 20 bits. Com essas codificações foram realizados um total de 110 experimentos.

Inicialmente variaram-se as posições da protease usadas. Foram testados três intervalos distintos. O primeiro deles era composto por posições de resistência ou assinatura de subtipo do vírus já conhecidas na literatura, totalizando 22 posições. O segundo continha todos os 99 aminoácidos da protease. Por fim, as 27 posições de maior relevância na mutação foram escolhidas por uma prévia análise dos dados.

Tais experimentos obtiveram bons resultados. Porém, como o grau de acerto

entre os grupos possuía grande variação, realizou-se então o balanceamento dos dados. Estes novos experimentos foram feitos somente com o intervalo que compreendia toda a entrada, pois tal intervalo havia apresentado bom resultado e não dependia da base de dados utilizada.

Para balancear os grupos utilizaram-se três métodos: repetição das amostras dos grupos menos populosos, utilização de apenas fração dos grupos mais populosos e da totalidade do grupo com menos amostras, separação dos grupos em lotes com mesmo número de amostras. Dentre estas técnicas de balanceamento, a que melhor se adequou ao problema foi a que realizou o loteamento dos dados. Com essa forma de balanceamento realizou-se experimentos variando o tamanho da memória da rede. Experimentou-se WiSARD de 8 bits e de 16 bits, obtendo melhores resultados e melhor velocidade na rede de 16 bits.

Em seguida, foram realizados experimentos mudando o foco do reconhecimento, procurando detectar apenas o subtipo da amostra, não diferenciando mais entre resistente e *naive*. Por fim, foi feito um último grupo de experimentos focando em diferenciar apenas as amostras do subtipo B entre resistentes e *naives*. Nestes últimos casos, realizaram-se experimentos efetuando o balanceamento por lote e em redes com memórias de 8 e 16 bits.

Para os experimentos realizados focando o reconhecimento dos grupos: B resistente, B *naive*, C resistente, C *naive*, F resistente e F *naive*, devem-se considerar apenas os que possuem balanceamento por lote. Nestes, o maior grau de acerto obtido foi de 69%, com desvio padrão de 5% apenas. No caso da categorização do subtipo, podem-se considerar tanto os experimentos em que as amostras *naives* foram treinadas em um grupo diferente das resistentes, como também aqueles em que estas amostras eram previamente agrupadas de acordo com o seu subtipo. Em muitos desses experimentos a acurácia ficou acima de 90% chegando a atingir no melhor dos casos grau de acerto de 93,9% com um desvio padrão de apenas 1%. No último experimento realizado, onde se buscava diferenciar dentre as amostras de subtipo B as que eram resistentes das que eram *naives*, obtive-se grau de acerto de 76,4% com um desvio padrão de 6%.

Conforme dito anteriormente, também foi avaliada a técnica de *Bleaching*, que atuaria nos casos em que houvesse empate, de modo a tornar o desempate determinístico e não mais aleatório. Para tal análise foram guardadas três informações para cada amostra testada. A primeira refere-se ao momento em que houve o desempate. A segunda era relativa à resposta de maior confiança, mantendo os mesmos grupos selecionados no caso anterior. Por fim, guardou-se também o resultado obtido para o último valor de *Bleaching* testado.

Analisando estas três informações pode-se verificar que sem o uso desta técnica apenas 72,6% dos experimentos chegariam a um resultado determinístico. Isso significa que, em média, aproximadamente um em cada quatro reconhecimentos seria randômico. Além disso, viu-se que o uso do *Bleaching* melhorou o grau de acerto do resultado em 18,5 pontos percentuais e aumentou a confiança em 84% dos experimentos. Apenas 13% dos experimentos não apresentaram diferença entre o primeiro e o segundo resultado armazenado. Portanto, somente para uma parte muito pequena dos resultados a melhor confiança foi encontrada no primeiro valor do *Bleaching*.

Ao analisar os resultados obtidos com o último valor do *Bleaching*, foi possível confirmar a necessidade de se trabalhar com grupos de amostras balanceados. Nos experimentos em que não se balanceou os dados, o grupo mais abundante sempre era encontrado como resposta. Portanto, conclui-se que é imprescindível o uso desta técnica. Percebe-se também que a melhor forma de usá-la é fixar os grupos obtidos no momento em que ocorre o desempate, mas deixar que o *Bleaching* continue atuando de modo a mostrar qual a confiança que aquela resposta pode alcançar. Esta foi a forma que apresentou uma melhor relação entre grau de acerto e confiança.

5.2 Trabalhos futuros

Embora este trabalho tenha apresentado resultados muito bons, ainda há muitas melhorias e abordagens que podem ser realizadas tendo-o como base. No Capítulo anterior foram descritas alguns desses aperfeiçoamentos que poderiam fazer parte deste trabalho, mas que acabaram não sendo incorporados. Nesta Seção serão descritas outras possíveis abordagens interessantes a serem realizadas. Elas são referentes tanto ao refinamento das técnicas utilizadas, quanto ao foco dos experimentos com a WiSARD.

5.2.1 Inclusão do Raio Molecular ou Outras Informações Químicas Para a Criação de Novas Codificações

Neste trabalho foram selecionadas apenas duas características químicas acerca dos aminoácidos para criar as codificações binárias. Tais informações de fato

são muito importantes e apresentaram resultados muito bons. Porém, é possível que outras apresentem desempenho ainda melhor. Elas podem ainda ser combinadas de modo que os aminoácidos fiquem mais bem representados.

Uma dessas características, por exemplo, é o raio molecular. Por ser um valor numérico, seria tão simples de codificá-lo quanto no caso das informações anteriores. Pode-se também usar informações acerca da estrutura da molécula, no que tange ao número de átomos de carbono, hidrogênio e demais elementos químicos por exemplo.

Até mesmo em relação às características já utilizadas, é possível realizar abordagens diferentes. Para a hidrofobicidade, podem ser separados alguns bits da codificação para diferenciar entre hidrofóbico, neutro e hidrofílico. Toda e qualquer informação química que afete o comportamento biológico do aminoácido e facilite ou atrapalhe a existência de mutação pode ser codificada numericamente e verificada.

5.2.2 Rede Neural que Possa Ser Treinada e Destreinada de Modo Automático, com Retroalimentação (Feedfoward)

No Capítulo 2 viu-se que é muito simples realizar o treinamento de uma rede neural sem peso, principalmente no caso da WiSARD. De maneira simplificada, percebe-se que tanto a WiSARD, quando a VG-RAM são treinadas adicionando a informação de que o padrão foi apresentado à rede. De maneira análoga, pode-se criar um procedimento que apague ou simplesmente altere essa informação, fazendo a rede entender que aquele padrão não foi aprendido, ou simplesmente foi esquecido.

Na forma tradicional como a WiSARD e programada, ao tentar apagar um padrão treinado anteriormente corre-se o sério risco de interferir em outros padrões gravados. Porém, o que permite pensar em fazer isso sem afetar os outros treinamentos e sem correr o risco de desconfigurar a WiSARD é a técnica do *Bleaching* descrita no Capítulo 2. Foi visto que ao usar esta técnica passa-se a marcar uma memória com valores numéricos que representam a quantidade de treinamentos que ela recebeu e não mais um valor booleano indicando se ela foi ou não treinada. Sendo assim, remover um aprendizado significaria diminuir o valor que está gravado na memória em uma unidade. Isto faria diferença na hora de executar o *Bleaching*, mas não indicaria que aquela memória não foi usada para treinar outros padrões daquela mesma categoria.

É importante fazer a ressalva que durante esse processo somente deve-se remover uma unidade do valor das memórias correspondentes ao padrão

apresentado, caso todas estejam com valor maior que zero. Pois, caso contrário, pode-se apagar um padrão que não foi aprendido para essa categoria.

Semelhante ao que provavelmente ocorre com o aprendizado humano, quando uma criança observa um novo padrão, ela tenta reconhecê-lo com base nos padrões que já aprendeu. A cada resposta dada ao padrão apresentado, ele pode confirmar, ou mesmo corrigir um aprendizado anterior. Na medida em que os padrões vão sendo apresentados, a taxa de aprendizado naturalmente diminui e a criança consegue ser capaz de reconhecer com eficiência novos padrões apresentados.

Com base nesta analogia, após a etapa de remoção de um padrão treinado ser testada, pode-se passar a fazer isso de modo automático. A WiSARD começaria vazia e apenas precisaria reconhecer os padrões apresentados. O primeiro padrão naturalmente implicaria em um treinamento. Para os padrões seguintes, a rede tentaria reconhecer com base no que já foi treinado. Em caso de erro a WiSARD treinaria esse novo padrão na categoria correta. Seria importante que depois de feito esse treinamento a rede tentasse reconhecer novamente este padrão, para se certificar que ele foi aprendido corretamente. Caso a resposta não seja positiva a este teste, pode ser necessário apagar esse “conhecimento” dos lugares onde ele aparece armazenado erroneamente.

Após diversos padrões terem sido apresentados, o esperado é que a rede já esteja com um grau de acerto extremamente elevado. A taxa de aprendizado deve estar também praticamente zerada. Estes reconhecimentos podem ser realizados até que a rede atinja uma taxa de acerto satisfatória, pré-determinada. Embora construir uma rede com alta taxa de acerto seja muito interessante, este procedimento pode apresentar um custo muito alto. Se esta rede tiver como objetivo, por exemplo, obter 100% de acerto, poderá ser preciso um tempo muito grande.

É importante perceber que este experimento dá margem a outro muito mais importante. O fato de se ter construído um reconhecedor de padrões com alto desempenho usando uma WiSARD permite ver com facilidade o que ela aprendeu e o que a faz ter resultados tão bons. No caso dos experimentos com as mutações do HIV isso pode revelar características importantes das mutações. Poderia, por exemplo, mostrar quais as combinações de mutações ocorridas e quais posições foram responsáveis para que determinada amostra do vírus se tornasse resistente a um medicamento. Desta forma, será possível descobrir os pontos mais interessantes que os geneticistas e virologistas devem observar no momento de estudar tais mutações. É fundamental ressalvar que para a realização deste último estudo, deve-se tomar muito cuidado com o embaralhamento dos bits feito pela WiSARD. Tal processo pode até mesmo impossibilitar que a etapa de aprendizado da rede se estabilize.

5.2.3 *Bleaching* Percentual

Nas análises feitas no Capítulo 4 mostrou-se como foi fundamental o uso da técnica do *Bleaching*. Porém ao longo dos experimentos realizados neste trabalho, observou-se que quando uma categoria possui muito mais treinamentos que as demais, esta técnica tende a privilegiar a categoria mais treinada. Neste trabalho, foi usada uma abordagem convencional para resolver este problema, realizando um loteamento da base de dados.

Porém, é possível experimentar balancear os treinamentos de outra forma. Ao invés de equilibrar o número de amostras na base de dados, pode-se equilibrar o peso de cada treinamento no momento em que o novo padrão é apresentado à rede. Assim, o balanceamento da rede não seria dependente das amostras apresentadas e as categorias ficariam sempre equilibradas. Para isso, basta que além de usar o *Bleaching*, seja guardado também o número de treinamentos realizados em cada categoria. Portanto, a informação do percentual que aquele valor de *Bleaching* representa para aquela memória será mantida.

Da forma como esta técnica foi usada neste trabalho, não há como saber se uma memória com valor 10 armazenado é uma memória que foi muito treinada, ou se representa uma exceção para aquela categoria. Porém, ao saber quantos treinamentos aquele discriminador possui, pode-se determinar o grau de representatividade e importância daquele padrão. Além disso, embora uma categoria com poucos treinamentos apresente valores de *Bleaching* baixos, percentualmente estes valores poderão ser tão elevados quanto os valores das categorias que foram treinadas para muitos padrões. Isto também criará um equilíbrio constante entre os discriminadores.

Tabela 5.1 Padrões treinados.

Categoria:	Padrão Treinado	Memória 1	Memória 2
A	1010	0100	0100
A	1011	0200	1100
A	0101	0210	1110
B	1001	0100	0010

O exemplo a seguir torna mais fácil o entendimento deste conceito. Nele, serão realizados treinamentos de uma entrada com 4 bits em uma WiSARD com memórias

com 2 bits e com o uso *Bleaching*. Na tabela 5.1 pode-se ver o que acontece com as memórias da rede, quando ela é treinada com os padrões 1010, 1011, 0101, para a categoria A e com o padrão 1001 para a categoria B. Em seguida, observa-se na tabela 5.2 o que ocorre ao realizar o reconhecimento do padrão 1001.

Tabela 5.2 Pontos por categoria do reconhecimento realizado.

Valor do Bleaching	Pontos de A	Pontos de B	Resultado Apresentado
0	1 + 1 = 2	1 + 1 = 2	Empate
1	1 + 0 = 1	0 + 0 = 0	Categoria A

Embora o padrão que se busca reconhecer tenha sido treinado na categoria B, o fato da categoria A ter realizado outros treinamentos semelhantes, impediu a WiSARD de reconhecê-lo corretamente. Caso o *Bleaching* não estivesse sendo usado, o reconhecimento resultaria em empate, e seria decidido de maneira aleatória entre essas duas categorias. Portanto, sem esta técnica a WiSARD teria 50% de chance de acertar. Embora não seja um resultado determinístico, já seria um resultado melhor que o obtido com o *Bleaching*.

Tabela 5.3 Padrões treinados.

Categoria:	Padrão Treinado	Memória 1	Memória 2	Número de Treinamentos
A	1010	0100	0100	1
A	1011	0200	1100	2
A	0101	0210	1110	3
B	1001	0100	0010	1

Nas tabelas 5.3 e 5.4 observa-se o que aconteceria usando o *Bleaching* percentual para a mesma situação anterior. É visível que, não só a WiSARD seria capaz de reconhecer corretamente o padrão apresentado como pertencente à categoria B, como também esta seria a única categoria a pontuar caso o *Bleaching* continuasse sendo utilizado. Além disso, por se tratar de um padrão treinado, tal categoria pontuaria em todas as memórias, tendo o resultado uma confiança de 100%, conforme deveria ser esperado para um padrão utilizado no processo de treinamento.

Tabela 5.4 Resultado do reconhecimento com nova forma de uso do *Bleaching*.

Valor do Bleaching	Pontos de A			Pontos de B			Resultado Apresentado
	M1	M2	Total	M1	M2	Total	
0/3 = 0	2/3 > 0 ? SIM	1/3 > 0 ? SIM	1+1 =2	1/1 > 0 ? SIM	1/1 > 0 ? SIM	1+1 =2	Empate
1/3 = 0,33	2/3 > 1/3 ? SIM	1/3 > 1/3 ? NÃO	1+0 =1	1/1 > 1/3 ? SIM	1/1 > 1/3 ? SIM	1+1 =2	Categoria B
2/3 = 0,66	2/3 > 2/3 ? NÃO	1/3 >2/3 ? NÃO	0+0 =0	1/1 > 2/3 ? SIM	1/1 > 2/3 ? SIM	1+1 =2	Categoria B

6 Referências Bibliográficas

ADESOKAN, A.A., ROBERTS, V.A., LEE, K.W., *et al*., 2004. *Prediction of hiv-1 integrase/viral DNA interactions in the catalytic domain by fast molecular docking*. J. MED. CHEM., V.47, PP. 821-828.ALEKSANDER, I., 1990. “An Introduction to Neural Computing”, Chapman and Hall, London.

ALEKSANDER, I. AND T.J. STONEHAM, 1979. *A Guide To Pattern Recognition Using Random Access Memories*, IEEE Journal Computers And Digital Techniques, Vol. 2(L), Pp. 29-40.

ALEKSANDER, I. 1998. *From Wisard to Magnus: A Family of Weightless Virtual Neural Machines*”. In: Ram-Based Neural Networks. World Scientific pp.18–30

AUSTIN, J., 1998. *Ram-Based Neural Networks*.

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

BEERENWINKEL, N. *et al.*, 2003. *Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes.* Nucleic Acids Res. 31, 3850–3855

BEERENWINKEL, N. *et al.,* 2002. *Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype.* Proc. Natl. Acad. Sci. U. S. A. 99, 8271–8276

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

BRODINE SK, MASCOLA JR, WEISS PJ, ITO SI, PORTER KR, ARTENSTEIN AW, *et al*. 1995. *Detection of diverse HIV-1 genetic subtypes in the USA*. Lancet. ; 346:1198-9.

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

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

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

CIARELLI, PATRICK MARQUES ; OLIVEIRA, ELIAS ; BADUE, CLAUDINE, 2009. *Multi-Label Text Categorization Using a Probabilistic Neural Network.* International Journal of Computer Information Systems and Industrial Management Applications, v. 1, p. 133-144.

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

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

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

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

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

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

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

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

FRANÇA, H. L. ; SILVA, J. C. P. ; LENGERKE, O. ; FRANÇA, F. M. G. ; DUTRA, M. S. 2009. "*Um sistema de visão artificial para o controle de perseguição de movimento por uma plataforma Stewart*". In: 2o Congreso Internacional De Ingeniería Mecatrónica Unab, 2009, Bucaramanga. Anais Do 2o Congreso Internacional De Ingeniería Mecatrónica Unab, 2009.

FREIMAN, JOSÉ PAULO E PAMPLONA, EDSON DE O., 2005. "*Redes neurais artificiais na previsão do valor de commodity do agronegócio*". V Encuentro Internacional De Finanzas. Santiago, Chile, 19 A 21 De Janeiro De 2005

GAO, F., VIDAL, N., LI, Y., *et al*., 2001. "*Evidence of two distinct subtypes with is the HIV-1 subtype a radiation*", AIDS Research and Human Retroviruses, v. 17(8), pp. 675-688.

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

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

GREGORIO M., 1996. "*Is that portal gothic? A hybrid system for Recognizing architectural portal shapes*", MVA '96: IAPR Workshop on Machine Vision Applications, Tokyo, Japan, pp. 389-392.

GRIECO, B. P. A.; LIMA, P. M. V.; DE GREGORIO, M.; FRANÇA, F. M. G, 2009. "*Extracting Fuzzy Rules From "Mental" Images Generated By Modified Wisard Perceptrons*". In: 17th European Symposium on Artificial Neural Networks, 2009, Bruges. Proc. Of Esann 2009. Evere, Belgium : D-Side, 2009. P. 313-318.

GRIECO, B. P. A.; LIMA, P. M. V.; GREGORIO, M.; FRANÇA, F. M. G., 2010. "*Producing Pattern Examples from Mental Images?*" Neurocomputing (Amsterdam) , V. 73, P. 1057-1064.

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

HIRSCH MS, BRUN-VEZINET F, CLOTET B, *et al*.,2003. "*Antiretroviral drug resistance testing in adults infected with human immunodeficiency virus type 1*". Recommendations of an International AIDS Society-USA Panel. Clin Infect Dis 2003; 37:113-128.

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

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

LABORATÓRIO VIRCO: (http://www.vircolab.com) acessado em 05/2011

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

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

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

MINISTÉRIO DA SAÚDE DST/AIDS (http://www.aids.gov.br), acessado em 05/2011.

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

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

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

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

PATTICHIS C.S., SCHIZAS C.N., SERGIOU A., SCHNORRENBERG F., 1994. "*A Hybrid Neural Network Electromyographic System: Incorporating the WiSARD Net*", IEEE World Congress on Computational Intelligence, Proceedings of the 1994 IEEE International Conference on Neural Networks, June 28-July 2, Orlando, Florida, Vol. VI, pp. 3478-3483.

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

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

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

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

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

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

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

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

SHAFER RW, WINTERS MA, PALMER S, MERIGAN TC 1998. "*Multiple concurrent reverse transcriptase and protease mutations multidrug resistance of HIV-1 isolates from heavily treated patients*". Ann Intern Med 128: 906-911.

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

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*". Doutorado, Instituto Alberto Luiz Coimbra De Pós-Graduação E Pesquisa De Engenharia (COPPE) Da Universidade Federal Do Rio De Janeiro (UFRJ). Rio de janeiro.

SILVA, R. M., ARRUDA, M. B. ALVES, M. R., TANURI A., BRINDEIRO, R. M. ; NOBRE, F. F.,2008. "*Identificação De Mutações Do HIV-1 Em Pacientes Com Falha Terapêutica Ao Nelfinavir Usando Modelo Computacional Híbrido*.". In: 21º. Congresso Brasileiro De Engenharia Biomédica, 2008, Salvador, BA. Anais Do 21º. Congresso Brasileiro De Engenharia Biomédica, 2008. P. 1764-1767.

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

SOUZA, A. F.; MELOTTI, B. Z.; BADUE, C., 2009. "*Multi-Label Text Categorization with a Data Correlated VG-RAM Weightless Neural Network*". International

Journal of Computer Information Systems and Industrial Management Applications, v. 1, p. 155-169.

SOUZA, A. F., PEDRONI, F, OLIVEIRA, E., CIARELLI, P. M., HENRIQUE, W. F., VERONESE, L. P., BADUE, C., 2009. "*Automated Multi-label Text Categorization with VG-RAM Weightless Neural Networks. Neurocomputing*" (Amsterdam), v. 72, p. 2209-2217.

SOUZA, A. F.; PEDRONI, F.; OLIVEIRA, E. ; CIARELLI, P. M. ; HENRIQUE, W. F. ; VERONESSE, L., 2007. "*Automated Free Text Classification of Economic Activities using VG-RAM Weightless Neural Networks*". In: International Conference on Intelligent Systems Design and Applications (ISDA), 2007, Rio de Janeiro. Proceedings of ISDA.

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

SPIRA, S., M. A. WAINBERG, H. LOEMBA, D. TURNER, AND B. G. BRENNER. 2003. "*Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance*". J. Antimicrob. Chemother. 51:229-240.

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

THOMÉ, A. C. G. 2011. "*Redes Neurais: Uma ferramenta para KDD E Data Mining*. Disponível em: http://www.labic.nce.ufrj.br. Acesso em Maio 2011.

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

VELLA S 2002. "*Antiretroviral drug resistance and HIV/AIDS in the developing world*". J HIV Ther 7: 53-55.

VELLA S, PALMISANO L 2000. "*Antiviral therapy: state of the HAART*". Antiviral Res 45: 1-7.

WANG, D., LARDER, B., 2003. *Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks.* J. Infect. Dis. 188, 653–660

WANG, D., LARDER, B., REVELL, A., MONTANER, J., HARRIGAN, R., WOLF, F., LANGE, J., SCOTT WEGNER, S., RUIZ, L., PEREZ-ELIAS, M. J., EMERY, S., GATELL, J., MONFORTE, A. A., TORTI, C., ZAZZI, M., LANE C. 2009. *"A comparison of three computational modeling methods for the prediction of virological response to combination HIV therapy.*" Artificial Intelligence in Medicine 47: 63—74.

WLODAWER, A., GUSTCHINA, A., 2000. "*Structural and biochemical studies of retroviral proteases*", Biochemical et Biophysical Act, v. 1477, pp. 16-34.

ANEXO I

Tabela ANEXO I.1 Codificações usando escala de hidrofobicidade.

Aminoácido	Binário	Gray Code	Dois Terço de Gray Code
I	11111111111111111111	10000000000000000000	11111111111111111111
V	11101101101101000001	10011011011011100001	11010001010001000001
L	11101110111001011010	10011001100101110111	11010000111000100010
F	11101000001110000101	10011100001001000111	11010111101110000110
C	11001010011110001110	10101111010001001001	11000101100001101110
M	11000110110110101010	10100101101101111111	11000110110110010010
A	10101111111011101010	11111000000110011111	01001111111011010010
G	10001100011101010110	11001010010011111101	01110011011100100111
T	10000010100001001001	11000011110001101101	01111100100000101000
S	10000010011000110010	11000011010100101011	01111101100110101010
W	01101111110011111010	01011000001010000111	01101111110011110100
Y	01100110001010110011	01010101001111101010	01100110000100101010
P	01101111101111001001	01011000011000101101	01101111010000110110
H	00110101100010111101	00101111010011100011	00110010011010111101
Q	00110010000110001001	00101011000101001101	00110001110101110110
N	00110010101111111010	00101011111000000111	00110001001111111011
E	00111011001111000001	00100110101000100001	00110100110000111101
D	00111011111000110011	00100110000100101010	00110100000110101011
K	00100111110110110101	00110100001101101111	00010111110110110010
R	00000000010011111011	00000000001110000110	00000000000100000011

Tabela ANEXO I.2 Codificações usando Massa Molecular.

Aminoácido	Binário	Gray Code	Dois Terço de Gray Code
I	01110000001111110000	01001000001000001000	01101111101111101110
V	01010100101011010100	01111110111110111110	00100100010010100100
L	01110000001111110000	01001000001000001000	01101111101111101110
F	10110011001100110011	11101010101010101010	01001100110011001100
C	01011100100011011100	01110010110010110010	00100011011010100011
M	10010011101100010011	11011010011010011010	01010011010011010011
A	00011101100010011101	00010011010011010011	00011010011010011110
G	00000001111110000010	00000001000001000011	00000001111110000001
T	01011000100111011000	01110100110100110100	00100110100110100110
S	00111101000010111101	00100011100011100011	00111100111010111101
W	11111111111111111111	10000000000000000000	11111111111111111111
Y	11010010101101010010	10111011111011111011	11001010010001001010
P	01010000101111010000	01111000111000111000	00101111001110101101
H	10011111100000011111	11010000010000010000	01011111011111011111
Q	10001101110010001101	11001011001011001011	01110001110001110001
N	01110010001101110010	01001011001011001011	01101010001101101010
E	10001111110000001111	11001000001000001000	01110000001111110000
D	01110100001011110100	01001110001110001110	01101011110011101011
K	10001101110010001101	11001011001011001011	01110001110001110001
R	11000100111011000100	10100110100110100110	11000010111011000010

Tabela ANEXO I.3 Codificações usando 10 bits com a hidrofobicidade e 10 bits com a massa molecular.

Aminoácido	Binário	Gray Code	Dois Terço de Gray Code
I	1111111111011000001	1000000000100100001	1111111110110111110
V	1111110011010101010	1000001010011111011	1111110011001010011
L	1111101000111000001	1000001110010010001	1111110010011011110
F	1111110000101100110	1000001000111010101	1111110000010011001
C	1111011100010111001	1000110010011100101	1111011010001000110
M	1111011010100100111	1000110111110110100	1111011001010100110
A	1111001011000111011	1000101110000010011	1111000010000011010
G	1110110100000000100	1001101110000000110	1101000100000000011
T	1110101101010110001	1001111011011101001	1101001110001001101
S	1110101101001111010	1001111011001000111	1101001010001111001
W	1110100001111111111	1001110001100000000	1101011110111111111
Y	1110011011101001010	1001010110101110111	1101010100110010100
P	1110100001010100001	1001110001011110001	1101011110001011110
H	1101111011100111110	1011000110110100001	1101111011010111110
Q	1101111001100011110	1011000101110010000	1101111001011100000
N	1101111001100011011	1011000101110010110	1101111001011100011
E	1101111111011010000	1011000000010011000	1101111111011010110
D	1101111111011001000	1011000000010010110	1101111111011010100
K	1101110010100011011	1011001011110010110	1101101010011100011
R	1101011000110001001	1011110100101001101	1100100110110000101 1

ANEXO II

Tabela ANEXO II.1 Resultados dos experimentos iniciais detalhados por categoria.

			GERAL		B RESISTENTE		B *NAIVE*		C RESISTENTE		C *NAIVE*		F RESISTENTE		F *NAIVE*	
	CÓDIGO	N AMOSTRAS	MÉDIA	DESVIO	MÉDIA	DESVIO	MÉDIA	DESVIO	MÉDIA	DESVIO	MÉDIA	DESVIO	MÉDIA	DESVIO	MÉDIA	DESVIO
CLASSICAS 8 SEM BALANCEAMENTO	BIN20	1205	64,7%	1,8%	99,3%	0,7%	0,0%	0,0%	12,1%	7,4%	4,0%	8,4%	8,9%	9,3%	0,0%	0,0%
	HIDROBIN	1205	67,6%	2,0%	97,5%	1,9%	0,0%	0,0%	24,5%	11,5%	0,0%	0,0%	32,4%	10,0%	0,0%	0,0%
	HIDROGRAY	1205	68,9%	2,0%	98,8%	1,0%	0,0%	0,0%	34,8%	9,7%	4,0%	8,4%	22,7%	10,3%	0,0%	0,0%
	HIDROGRAY23	1205	67,2%	1,9%	98,3%	1,7%	0,0%	0,0%	18,2%	8,4%	4,0%	8,4%	30,2%	12,6%	0,0%	0,0%
	MMBIN	1205	64,8%	0,7%	97,9%	1,8%	1,0%	3,2%	2,5%	4,4%	4,0%	8,4%	27,9%	10,8%	0,0%	0,0%
	MMGRAY	1205	64,8%	1,3%	99,2%	1,1%	0,0%	0,0%	1,9%	4,2%	4,0%	8,4%	22,3%	12,0%	0,0%	0,0%
	MMGRAY23	1205	65,1%	1,5%	98,8%	1,5%	0,0%	0,0%	5,7%	6,2%	2,0%	6,3%	22,8%	5,2%	0,0%	0,0%
	HIDROMMBIN	1205	67,2%	2,2%	97,6%	2,0%	0,0%	0,0%	33,6%	9,9%	4,0%	8,4%	16,1%	7,4%	0,0%	0,0%
	HIDROMMGRAY	1205	64,6%	1,8%	98,5%	2,0%	0,0%	0,0%	13,0%	7,1%	4,0%	8,4%	11,6%	6,0%	0,0%	0,0%
	HIDROMMGRAY23	1205	67,6%	2,4%	98,5%	1,3%	1,0%	3,2%	27,8%	13,1%	4,0%	8,4%	19,9%	9,2%	0,0%	0,0%
TODA 8 SEM BALANCEAR	BIN20	1205	78,3%	2,9%	98,8%	1,5%	35,3%	9,6%	62,3%	20,7%	14,0%	13,5%	43,5%	16,7%	10,0%	21,1%
	HIDROBIN	1205	80,4%	3,3%	99,5%	0,9%	34,4%	10,1%	68,9%	13,3%	10,0%	14,1%	52,2%	12,8%	20,0%	25,8%
	HIDROGRAY	1205	81,6%	2,5%	99,1%	1,1%	36,4%	11,0%	78,0%	11,1%	18,5%	20,0%	50,5%	14,4%	15,0%	24,2%
	HIDROGRAY23	1205	79,8%	3,1%	99,1%	1,4%	36,4%	11,0%	69,6%	17,7%	24,5%	22,7%	42,9%	15,0%	10,0%	21,1%
	MMBIN	1205	82,0%	2,4%	99,3%	1,0%	25,2%	8,4%	81,2%	15,5%	10,0%	14,1%	61,6%	15,0%	5,0%	15,8%
	MMGRAY	1205	78,5%	3,9%	98,9%	2,1%	33,3%	11,5%	67,5%	16,1%	0,0%	0,0%	44,6%	18,1%	15,0%	24,2%
	MMGRAY23	1205	80,5%	3,1%	99,5%	0,7%	31,4%	10,5%	77,3%	13,7%	4,0%	12,6%	48,4%	13,9%	15,0%	24,2%
	HIDROMMBIN	1205	77,8%	2,4%	98,8%	1,2%	34,2%	9,3%	57,3%	12,1%	10,0%	10,5%	47,1%	14,3%	15,0%	24,2%
	HIDROMMGRAY	1205	78,3%	2,1%	99,2%	0,9%	35,2%	10,5%	56,1%	11,1%	12,0%	14,0%	48,5%	10,1%	15,0%	24,2%
	HIDROMMGRAY23	1205	79,1%	3,3%	98,9%	1,1%	38,4%	9,4%	67,5%	14,5%	14,0%	13,5%	40,9%	17,4%	15,0%	24,2%
27M 8 SEM BALANCEAR	BIN20	1205	79,8%	3,1%	99,1%	1,4%	36,4%	11,0%	69,6%	17,7%	24,5%	22,7%	42,9%	15,0%	10,0%	21,1%
	HIDROBIN	1205	85,2%	2,7%	98,9%	1,4%	34,3%	11,7%	87,0%	6,3%	28,5%	19,2%	70,6%	13,9%	15,0%	24,2%
	HIDROGRAY	1205	85,3%	2,4%	99,1%	0,6%	34,3%	10,7%	94,8%	5,1%	30,5%	25,2%	61,0%	15,0%	20,0%	25,8%
	HIDROGRAY23	1205	84,7%	3,2%	98,9%	1,4%	33,3%	10,5%	89,6%	9,6%	26,5%	23,1%	64,8%	16,3%	20,0%	25,8%
	MMBIN	1205	85,1%	2,7%	98,4%	1,8%	28,3%	12,4%	95,4%	5,5%	6,0%	9,7%	75,9%	15,2%	15,0%	24,2%
	MMGRAY	1205	80,2%	3,3%	99,5%	0,9%	34,4%	13,0%	65,9%	11,5%	2,0%	6,3%	56,9%	23,0%	20,0%	25,8%
	MMGRAY23	1205	83,1%	2,8%	98,5%	1,6%	30,4%	12,0%	86,3%	13,8%	6,0%	9,7%	66,3%	18,3%	15,0%	24,2%
	HIDROMMBIN	1205	82,6%	2,6%	99,1%	0,9%	35,4%	13,0%	82,6%	8,5%	20,5%	18,9%	54,4%	13,7%	15,0%	24,2%
	HIDROMMGRAY	1205	82,7%	3,1%	98,7%	0,9%	39,4%	10,1%	77,4%	14,2%	24,5%	18,3%	59,6%	15,8%	15,0%	24,2%
	HIDROMMGRAY23	1205	83,4%	3,8%	98,7%	1,4%	35,4%	13,0%	78,8%	14,8%	26,5%	21,1%	65,5%	16,4%	20,0%	25,8%

Tabela ANEXO II.2 Resultados dos experimentos balanceados pelo grupo B e pelo grupo F *naive* detalhados por categoria.

			GERAL		B RESISTENTE		B *NAIVE*		C RESISTENTE		C *NAIVE*		F RESISTENTE		F *NAIVE*	
	CÓDIGO	N AMOSTRAS	MEDIA	DESVIO	MEDIA	DESVIO	MEDIA	DESVIO	MEDIA	DESVIO	MEDIA	DESVIO	MEDIA	DESVIO	MEDIA	DESVIO
BALANCEAMENTO B	BIN20	4482	84,4%	1,5%	81,6%	5,4%	78,0%	2,2%	80,0%	6,3%	85,0%	1,3%	87,0%	2,6%	94,8%	0,4%
	HIDROBIN	4482	89,3%	1,2%	91,6%	2,5%	79,0%	2,1%	85,0%	7,8%	89,6%	2,1%	95,8%	2,8%	94,8%	0,4%
	HIDROGRAY	4482	90,2%	0,7%	91,0%	3,0%	81,0%	1,2%	87,8%	6,9%	91,0%	1,9%	95,6%	1,7%	94,8%	0,4%
	HIDROGRAY23	4482	90,9%	1,5%	91,3%	3,6%	87,5%	3,0%	81,8%	6,9%	88,7%	2,1%	96,1%	1,6%	100,0%	0,0%
	MMBIN	4482	89,4%	1,1%	90,1%	3,4%	84,9%	1,6%	88,2%	7,5%	83,0%	2,4%	95,3%	2,4%	94,8%	0,4%
	MMGRAY	4482	90,4%	1,3%	91,3%	2,8%	80,6%	1,9%	88,2%	5,8%	85,3%	2,3%	97,1%	1,2%	100,0%	0,0%
	MMGRAY23	4482	90,4%	1,3%	89,8%	2,9%	83,5%	2,3%	88,2%	3,8%	86,7%	2,2%	94,4%	4,9%	100,0%	0,0%
	HIDROMMBIN	4482	89,1%	1,7%	89,1%	2,2%	81,9%	1,7%	84,7%	7,2%	83,7%	2,4%	95,0%	3,9%	100,0%	0,0%
	HIDROMMGRAY	4482	89,6%	1,5%	91,7%	2,5%	83,7%	2,4%	82,7%	7,0%	83,9%	2,1%	97,2%	1,6%	98,4%	2,6%
	HIDROMMGRAY23	4482	89,3%	1,6%	87,7%	4,4%	83,3%	2,3%	85,5%	6,5%	85,8%	3,0%	93,4%	3,5%	100,0%	0,0%
BALANCEAMENTO F *NAIVE*	BIN20	114	79,8%	7,3%	100,0%	0,0%	70,0%	35,0%	85,0%	24,2%	65,0%	33,7%	80,0%	35,0%	80,0%	25,8%
	HIDROBIN	114	77,3%	6,8%	95,0%	15,8%	65,0%	33,7%	80,0%	25,8%	60,0%	39,4%	85,0%	24,2%	80,0%	25,8%
	HIDROGRAY	114	74,7%	6,9%	90,0%	21,1%	70,0%	42,2%	70,0%	25,8%	60,0%	31,6%	75,0%	35,4%	85,0%	24,2%
	HIDROGRAY23	114	78,3%	6,7%	90,0%	21,1%	60,0%	39,4%	75,0%	26,4%	70,0%	25,8%	85,0%	24,2%	90,0%	21,1%
	MMBIN	114	75,6%	10,2%	85,0%	24,2%	80,0%	35,0%	75,0%	26,4%	55,0%	36,9%	85,0%	24,2%	80,0%	25,8%
	MMGRAY	114	74,6%	8,3%	90,0%	21,1%	60,0%	45,9%	75,0%	26,4%	55,0%	36,9%	90,0%	21,1%	70,0%	35,0%
	MMGRAY23	114	80,8%	9,1%	95,0%	15,8%	85,0%	33,7%	85,0%	24,2%	65,0%	33,7%	85,0%	24,2%	70,0%	25,8%
	HIDROMMBIN	114	76,6%	11,7%	85,0%	33,7%	80,0%	35,0%	85,0%	24,2%	50,0%	40,8%	75,0%	26,4%	85,0%	24,2%
	HIDROMMGRAY	114	73,7%	6,7%	80,0%	35,0%	60,0%	39,4%	70,0%	35,0%	55,0%	36,9%	85,0%	24,2%	85,0%	24,2%
	HIDROMMGRAY23	114	78,9%	7,5%	85,0%	33,7%	70,0%	35,0%	85,0%	24,2%	65,0%	33,7%	80,0%	25,8%	85,0%	24,2%