ESCOLA POLITÉCNICA

DE PERNAMBUCO

d.s.c. UPE
Departamento de Sistemas e Computação

APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS NA MODELAGEM DE UM CLASSIFICADOR DE FORMAS CLÍNICAS DE DENGUE UTILIZANDO DADOS GENÔMICOS

Trabalho de Conclusão de Curso

Engenharia da Computação

Thiego Wanderley Fontes de Oliveira

Orientador: Prof. Fernando Buarque Lima Neto

Co-Orientador: Bartolomeu Acioli-Santos

Recife, Dezembro 2009

UNIVERSIDADE
DE PERNAMBUCO

THIEGO WANDERLEY FONTES DE OLIVEIRA

APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS NA MODELAGEM DE UM CLASSIFICADOR DE FORMAS CLÍNICAS DE DENGUE UTILIZANDO DADOS GENÔMICOS

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia da Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Recife, dezembro de 2009.

Dedico este trabalho aos meus pais
Wanderley e Cláudia, minha irmã Thiale,
e a minha noiva Kelly, que são as
pessoas quem mais amo nesta vida.

Agradecimentos

Agradeço à Deus por me dar discernimento, guiar, e cuidar durante a caminhada até aqui.

Aos meus pais e minha irmã que são minha fortaleza, e que sempre me deram total apoio e orientação nas decisões que tive que tomar. Agradeço muito à Kelly, minha namorada, que carinhosamente me acompanha.

À minha família que esteve sempre unida, e foi compreensiva com minhas ausências.

Aos amigos da Koinonia e do RadaR que me alegraram em todas as programações e saídas juntos.

À todos os companheiros do LaViTE que tiveram paciência em responder os questionamentos e curiosidades sobre as rotinas do laboratório, e que foram pacientes quando eu não pude realizar minha atividades. Agradeço também ao meu ex co-orientador Dr. Carlos Calzavara que foi compreensivo com a mudança do projeto.

Ao meu co-orientador Dr. Bartolomeu Acioli por me dar total base de conhecimento na área biológica, que era uma desconhecida pra mim, e me apoiar durante todo o desenvolvimento desta monografia.

Ao meu orientador Prof. Dr. Fernando Buarque por ter me dado inspiração, e vontade de buscar os desafios na ciência, sendo também professor de Metodologia Científica e Inteligência Artificial, disciplinas que foram fundamentais para escolha do tema deste trabalho.

Agradeço ao Prof. Dr. Mêuser Valença por tirar várias dúvidas de corredor, e por lecionar as disciplinas Redes Neurais e Mineração de Dados, fundamentais para este trabalho.

Aos colegas de faculdade Marcel Caraciolo, Tiago Bockholt, Luiz Soares, por ter disponibilidade para me ajudar, e a Gustado Oliveira que liderou projetos de outras disciplinas durante este TCC.

E à todos que de uma forma ou outra me ajudaram e não foram citados.

Resumo

Das primeiras manifestações clínicas da dengue até seu diagnóstico, muitos pacientes desenvolvem quadros graves que podem evoluir a óbito, dada a impossibilidade de uma intervenção mais intensa e apropriada para aqueles pacientes que vão desenvolver as formas graves de dengue, em especial, a dengue hemorrágica. Considera-se três formas clínicas para a dengue: clássica, clássica complicada, e febre hemorrágica, estas duas ultimas sendo as mais letais da doença. A diferenciação das formas de dengue, segundo recomendação da Organização Mundial de Saúde (OMS), é feita baseada em aspectos clínicos e exames laboratoriais, que em certos casos podem levar a diagnósticos imprecisos. Este trabalho teve como objetivo o desenvolvimento de um método prognóstico baseado em dados genômicos (polimorfismos genéticos), tendo como suporte inteligência computacional. A base de dados incluiu 105 pacientes da coorte de dengue do LaviTE, 26 FHD, 49 DCC e 30 DC obtidos de aplicação de técnicas de genotipagem em massa (*Illumina)*. Foi desenvolvido um sistema de gerenciamento de banco de dados para pré-processar dados genotípicos (322 loci polimórficos), extraindo resultados estatísticos como: O.R, equilíbrio de H-W e o Teste de Fisher. A abordagem univariada ressaltou a participação de 6 loci gênicos relacionados a imunidade inata como importantes na definição do fenótipo severo ou homorrágico em dengue. Na avaliação multivariada foi implementado um modelo utilizando Redes Neurais Artificiais (Multilayer Perceptron) capaz de classificar com uma sensibilidade de 85% casos de dengue severa (DH+DCC). Os dados indicam uma potencial utilização desta ferramenta como método auxiliar de prognóstico da dengue grave nos serviços de saúde, auxiliando no tratamento do paciente, controle e combate à doença.

Abstract

The delay between initial symptoms and reaching a diagnosis is sometimes the cause of death for many patients, given the impossibility of a more detailed intervention in the patients that will develop the severe form of dengue. There are three levels of severity of clinical forms of Dengue, namely, Classic Dengue Fever, Complicated Dengue Fever, and Hemorrhagic Dengue Fever. The two latter are considered the lethal ones. The differentiation between those clinical forms follows the recommendation of the World Health Organization (WHO), and is based on the observation of clinical aspects and some laboratorial exams, but these two means could be inaccurate. Because of that, this research looks for to minimize this problem, by developing a tool that uses artificial intelligence to better help correct diagnosis, and based in genomic data (genetic polymorphism). The database used a cohort of 105 patients from LaViTE, 26 DHF, 49 CDF an 30 DF obtained from the application of a technique of mass genotyping (Illumina). The developed, a System of management of database, pre-processed genotypic data (322 polymorphic *loci)*, drawing results of some statistical calculations like Odds Ratio, H-W Equilibrium, Fisher Test. The univariable approach showed the participation of 6 *loci* related to the innate immunity as important on the definition of the phenotype severe or hemorrhagic form of dengue. The Artificial Neural Networks model used tackled genotypic data, and produced correct classification with a sensibility of 85% the cases of severe dengue from classical dengue fever. After further calibrations, this tool is deemed to be used in real diagnosis prediction and could also facilitate the implementation of public programs for Prevention, Control and Treatment of Dengue.

Sumário

Índice de Figuras

Índice de Tabelas

Tabela de Símbolos e Siglas

CPqAM - Centro de Pesquisas Aggeu Magalhães

DC – Dengue Clássica

DCC – Dengue Clássica Complicada

DEN – Vírus da Dengue

DG – Dengue Grave

DNA – Ácido Desoxirribonucléico

DSC – Departamento de Sistemas e Computação

Eq.H-W – Equilíbrio de Hardy Weinberg

FHD – Febre Hemorrágica da Dengue

FIOCRUZ – Fundação Oswaldo Cruz

Hip – Hipótese

IA – Inteligência Artificial

LaViTE - Departamento de Virologia e Terapia Experimental

MLP – Multilayer Perceptron (Perceptron Multicamada)

mRNA - Ácido Ribonucléico mensageiro

MS – Ministério da Saúde

NR – Não houve reação

O.R. – Odds-Ratio

OMS – Organização Mundial da Saúde

PCR – Reação em Cadeia de Polimerase

RNA – Redes Neurais Artificiais

RT-PCR – Transcrição Reversa por Reação em Cadeia de Polimerase

SCD – Síndrome do Choque da Dengue

SGBD – Sistema de gerenciamento de banco de dados

SNP – Polimorfismo Simples de Nucleotídeo

SVN – Support Vector Machine (Máquina de Vetor de Suporte)

XX – Forma genotípica homozigota mais prevalente

XY – Forma genotípica heterozigota

YY – Forma genotípica homozigota menos prevalente

Capítulo 1

Introdução

Após relevantes descobertas da biologia molecular nas décadas de 50, foi determinada a estrutura do DNA e o modo como a informação genética é codificada por quatro nucleotídeos (Adenina – A, Guanina – G, Citocina – C, e Timina – T) [1] [2]. Desde então teve início grande número de pesquisas nessa área, posteriormente batizada de biologia molecular. O advento do método de seqüenciamento automático do DNA [3] e a conseqüente massificação dos estudos genômicos, com especial atenção para a conclusão do Projeto do genoma humano [4] , foram determinantes no desenvolvimento de uma nova área de estudo da biologia: a Bioinformática.

Nos anos recentes, a Bioinformática, definida como a computação direcionada à informação genética, tem se tornado uma das áreas de maior visibilidade na ciência moderna [5]. A cada dia a bioinformática vem se caracterizando não somente como uma disciplina voltada para desenvolvimento de ferramentas moleculares ou de análise de seqüências mas, sim, como uma verdadeira área de pesquisas, com epistemologia e metodologia próprias as quais são de interesse para muitos pesquisadores (e este trabalho igualmente) [6].

A aplicação da Bioinformática no estudo de doenças é bem variada. Ferramentas computacionais têm sido empregadas no estudo de padrões de transcrição gênica (o transcriptoma), padrões de tradução (o proteoma) bem como de em análises de diferenças de composição do DNA de indivíduos e populações (o genoma), onde tem trazido bons resultados para o entendimento dos mecanismos genéticos que suportam o desenvolvimento do quadro clínico (fenótipo). Tudo isso trazendo melhorias em terapias e diagnósticos das mesmas.

A dengue é uma doença viral endêmica no Brasil. É transmitida pelo vetor *Aedes aegypt,* e uma vez infectado o paciente pode evoluir do quadro benigno (dengue clássica) ao quadro hemorrágico. Dentre as variadas vertentes de estudo da dengue, clínicas, epidemiológicas, a bioinformática aplicada a estudos de

genética de populações e genômica funcional têm ganho destaque dada as grandes potencialidades advindas da sua utilização, na busca de marcadores moleculares para o desenvolvimento das formas mais graves da doença.

1.1 Motivação

A saúde pública brasileira vem sendo atingida por ciclos de epidemias de dengue de 1986 até os dias atuais. De acordo com o Ministério da Saúde (MS), mais de quatro milhões de casos de dengue foram notificados no país neste período. Muitos desses casos têm levado os pacientes ao óbito, causado, em sua maioria, pela demora no diagnóstico da doença [7].

A dengue é uma doença febril aguda, de origem viral que pode ser de curso benigno ou grave. Seu diagnóstico é feito através de exames que tem resultados precisos, embora sejam dispendiosos. Quando seu diagnóstico é rápido, o tratamento pode ser feito antes do agravamento do quadro clínico, prevenindo os óbitos [8]. O diagnóstico das formas clínicas da dengue é realizado seguindo-se a recomendação da OMS, sendo este baseado apenas em aspectos clínicos. Isso implica em certo grau de imprecisão devido a similaridade dos sintomas iniciais da infecção pelo vírus dengue (DENV) com outras doenças, como gripe, leptospirose, infecção urinária, dentre outras [9]. Outro importante aspecto no diagnóstico envolvendo dengue é a incapacidade dos testes laboratoriais atuais para identificar pacientes com propensão a desenvolver formas severas da doença.

Este estudo aborda o aspecto genético da dengue. Utilizando dados genômicos de pacientes (SNPs) infectados pelo DENV como subsídio para a ferramenta de inteligência computacional visando análises. Os resultados deste trabalho poderão culminar em métodos alternativos de prognóstico das formas graves da dengue.

Este trabalho capitaliza em pesquisas biológicas realizadas no Departamento de Virologia e Terapia Experimental (LaViTE), do Centro de Pesquisas Aggeu Magalhães (CPqAM), da FIOCRUZ, onde foram identificados praticamente 2000 genes diferencialmente expressos em pacientes com a forma branda da dengue e pessoas que manifestaram os sintomas severos da dengue hemorrágica; sendo este

derivado de um projeto principal intitulado *Identificação de marcadores biológicos preditivos de respostas clínicas ao dengue através de análise proteômica* cujas considerações éticas foram aprovadas pelo CONEP, sob o no. 4909. Após diversas analises estatísticas [10], concluiu-se que cerca de 40 genes deveriam ser para um estudo mais aprofundado de expressão gênica, baseado na técnica de Reação em Cadeia da Polimerase (do inglês, *Polymerase chain reaction* – PCR) quantitativa em tempo real, e observou-se que tais genes tinham expressão diferenciadas para dengue hemorrágica, especialmente na fase convalescente, um dado relevante que sugere a presença e polimorfismos genéticos nos genes [10]. Assim, este problema, por sinal, é uma das principais motivações pela qual coortes de pacientes dengue são estabelecidas visando o amplo estudo da doença. [11].

1.2 Objetivos

1.2.1 Objetivo Principal

- Este trabalho tem como objetivo desenvolver uma ferramenta computacional que utiliza técnicas de inteligência computacional e seja capaz de definir mais precisamente as formas graves da dengue utilizando dados genômicos (SNPs) e ferramentas de inteligência computacional (redes neurais) (seção 3.2)

1.2.2 Objetivos Secundários

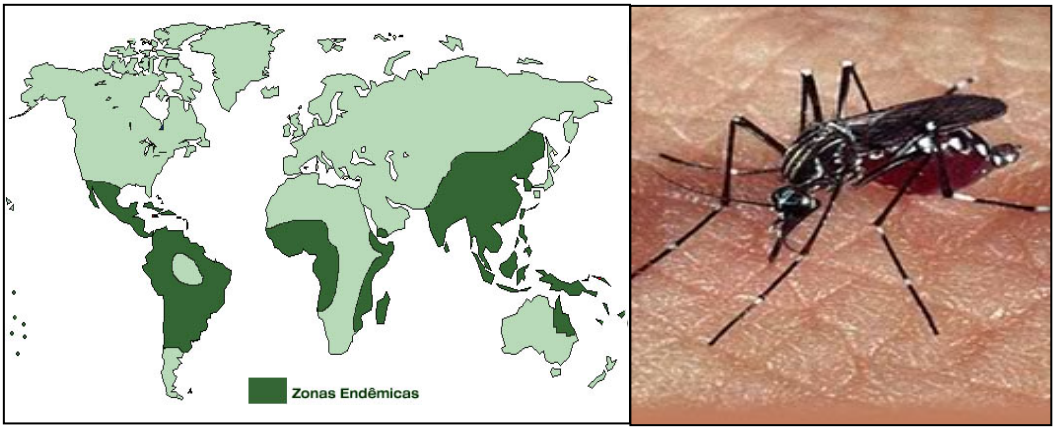
- Realizar pré-processamento dos dados obtidos a partir da técnica de genotipagem em massa de pacientes com diferentes manifestações de dengue;
- Desenvolver funcionalidade capaz de realizar cálculos estatísticos univariados de *Odds-Ratio*, Teste de Fisher e equilíbrio de Hard Weinberg sobre os dados genômicos dos pacientes (seção 4.4);
- Desenvolver um modelo computacional utilizando Redes Neurais Artificiais (*Multilayer Perceptron*), para estudo genômico multivaridado (multigênico) como alternativa de predição das formas graves da dengue.

Capítulo 2

Aspectos gerais da Dengue

2.1 Epidemiologia

A dengue é uma doença viral causada por um flavivírus (arbovírus) e transmitida por mosquitos. Segundo a Organização Mundial de Saúde a dengue atinge cerca de 80 milhões de pessoas anualmente em mais de cem países (figura 1). Destas, cerca de 550 mil necessitam de hospitalização e pelo menos 20 mil chegam a óbito, constituindo-se em um grave problema de saúde pública [12]. Os vetores transmissores do vírus da dengue são os artrópodes hematófagos *Aedes albopictus* e o *Aedes aegypti* (figura 1. B). O primeiro vive principalmente em zonas rurais, enquanto os mosquitos da espécie *A. aegypti* são especializados em sobreviver nas zonas urbanas. Essa especialização e a sua maior habilidade em transmitir o vírus para os humanos são características que os garante como principais transmissores da dengue no mundo [13].



A B

Figura 1. A - Distribuição da dengue no mundo. (Fonte: http://www.escola24horas.com.br/imagens/mapa_dengue.jpg). B – Mosquito *Aedes aegypti* (Fonte: http://www.faunabrasil.com.br).

Existem quatro sorotipos antigênicos distintos do vírus da dengue, que vão de vírus da dengue 1 a 4 (DEN-1 a 4) [14]. O DEN foi primeiramente isolado no Japão e classificado como DEN-1. O DEN-2 foi isolado na Nova Guiné, e o DEN-3 e 4 a partir de pacientes com quadros hemorrágicos, em 1956, nas Filipinas [13].

O ciclo de transmissão do vírus da dengue começa quando o mosquito pica uma pessoa infectada, então, o vírus multiplica-se no intestino médio do vetor, chegando finalmente às glândulas salivares do inseto, de onde sairá, durante o repasto, para a corrente sangüínea de outra pessoa que será infectada. Assim que penetra na corrente sangüínea humana o vírus passa a se multiplicar em órgãos específicos, como o baço, o fígado e os tecidos linfáticos. Esse período é conhecido como incubação e dura de quatro a sete dias. Em seguida, o vírus é novamente liberado na corrente sangüínea, quando os pacientes desenvolvem os primeiros sintomas. O DEN também se replica nas células sangüíneas e atinge a medula óssea, comprometendo a produção de plaquetas. Durante sua multiplicação, formam-se substâncias que agridem as paredes dos vasos sangüíneos, provocando uma perda de plasma para o interstício, podendo culminar em sérios distúrbios no sistema circulatório como hemorragias e queda da pressão arterial [15].

2.2 Histórico

Os primeiros relatos sugerindo a ocorrência da dengue foram documentados na enciclopédia chinesa publicada durante a dinastia Chin (265 a 420 d.C.) [16]. A partir do século XVIII começaram a surgir episódios importantes de infecção pela dengue como em Jacarta, em 1779, e Taiwan, em 1916 [13]. A primeira epidemia da dengue tipo hemorrágico na Ásia ocorreu em Manila, Filipinas, em 1953-1954 [17]. A descrição mais segura de um surto de dengue nas Américas aconteceu na Filadélfia em 1780 [16].

Já no Brasil, o primeiro caso confirmado de dengue foi reportado em Boa Vista, estado de Roraima. Isto aconteceu em 1982, onde foram isolados os sorotipos 1 e 4 do vírus da dengue [18]. Em 1986, houve uma epidemia no estado do Rio de Janeiro causada pelo sorotipo 1, cidade que também apresentou duas outras grandes epidemias em 1990 e 2000, devido à introdução dos sorotipos 2 e 3,

respectivamente [19] [20]. Depois da introdução do sorotipo 3, em 2000, foram reportadas epidemias em todo o país, inclusive no estado de Pernambuco [11].

Em Pernambuco, a primeira epidemia descrita ocorreu em 1987 pelo sorotipo 1, com 2.118 casos reportados. Em 1995, depois de sete anos sem registros de casos autóctones, uma nova epidemia ocorreu com a introdução do sorotipo 2 [11]. Em 2002 houve o registro de outra epidemia devido à entrada do sorotipo 3, na qual a Região Nordeste apresentou a maior taxa de incidência regional do Brasil, tendo o estado de Pernambuco o maior índice da região [13]. Desde então, os três sorotipos (1,2 e 3) estão circulando no estado pernambucano.

2.3 Formas Clínicas

A grande maioria dos pacientes infectados pelo vírus da dengue apresenta um quadro auto-limitado benigno, conhecido como dengue clássica (DC). Esta forma é caracterizada por febre alta (39ºC a 40ºC), associada à cefaléia, prostração, mialgia, artralgia, dor retroorbitária, exantema maculopapular acompanhado ou não de prurido. No final do período febril, podem surgir manifestações hemorrágicas e em casos mais raros, podem existir sangramentos maiores, que não são exclusivos da forma hemorrágica, chamados de vasculopatia [21].

Em uma porção menor, a doença apresenta evolução para a febre hemorrágica da dengue (FHD). Uma forma muito mais severa, que apresenta além das características inicias da DC, plaquetopenia (plaquetas < 100.000/mm3), aumento da permeabilidade capilar, extravasamento de plasma e anormalidades homeostáticas, sobretudo hemoconcentração. Este quadro pode ainda evoluir para insuficiência circulatória e síndrome do choque por dengue (SCD), a forma mais comum de morte causada por essa patologia [22]. Casos complicados cujos critérios não completam suficientemente as exigências (OMS) para FHD, especialmente a hemoconcentração, são classificados como dengue clássica complicada (DCC) [9]. Nessa situação, a presença de um dos itens a seguir caracteriza o quadro: alterações neurológicas; disfunção cardiorrespiratória; insuficiência hepática; plaquetopenia igual ou inferior a 50.000/mm3; hemorragia digestiva; derrames cavitários ou leucometria global igual ou inferior a 1.000/ mm3 [21]. Esta última forma clínica passou a ser aceita pelo MS, entretanto para a OMS, permanecem apenas as

formas DC e FHD. Por isso neste trabalho as duas possibilidades foram levadas em conta, segundo o MS, as formas DC, DCC, e FHD, e segundo a OMS as formas DC e FHD, desta forma os casos de DCC foram tratados como casos de DC. E uma última forma de tratar as formas clínicas, em que surge uma classificação de risco, a forma grave, ou Dengue Grave (DG) que agrupa os casos que apresentam riscos para os pacientes por ela acometidos, que engloba os casos de DCC e de FHD.

Um indivíduo não pode ser infectado duas vezes pelo mesmo sorotipo do vírus da dengue, já que uma vez infectado por um sorotipo, o paciente adquire resistência a este, entretanto, devido à existência dos quatro diferentes sorotipos do vírus, um paciente pode ser infectado uma ou mais vezes pela dengue, mas por sorotipos diferentes. Denomina-se assim, que um paciente infectado pela primeira vez pelo dengue, é uma infecção do tipo primária, e se já tiver sido infectado anteriormente pelo vírus, é do tipo secundária. Existe um modelo que classifica o tipo de infecção baseado no teste Elisa de IgG [23].

2.4 Polimorfismos Genéticos e a dengue.

Os polimorfismos simples de nucleotídeos (SNPs) podem ser uma alternativa de estudo para compreensão das diferentes formas de resposta do homem frente ao vírus dengue. Isso é possível porque as SNPs podem alterar a seqüência protéica do gene, modificar a estabilidade e estrutura do mRNA (RNA mensageiro), bem como alterar a quantidade de proteína traduzida pelo gene. Por essas características, as SNPs já vêm sendo utilizadas como marcadores moleculares para diversas doenças. Para dengue existem estudos que mostram que polimorfismos de certos *loci* estão associados ao risco ou proteção à forma mais grave da doença, como é o caso dos genes do MBL2 [24], e do Fator H [25], TNF-α [26] entre outros, por isso se faz necessário o estudo com uma abordagem ampla da relação entre polimorfismos no desenvolvimento do fenótipo severo em dengue.

2.5 Modelos de classificação das Formas Clínicas

A classificação das formas clínicas de forma automática é objeto de estudo por vários grupos de pesquisa. Estudos utilizando aspectos clínicos como parâmetros são os mais comuns, em 2008 Lee desenvolveu um modelo utilizando o Algoritmo do Vizinho mais Próximo [27], e em 2009 outro estudo com a mesma coorte utilizando desta vez o Algoritmo de Árvore de Decisão [28], entretanto, esta abordagem clínica é questionada pela relatividade dos sintomas, e o aspecto subjetivo relacionado ao que o paciente está sentindo, ou apresentando.

Uma abordagem com suporte genético seria uma alternativa à utilização de dados clínicos. Mesmo requerendo estudos muito mais custosos o aspecto genético produz resultados precisos e com mais segurança científica. Alguns trabalhos já apontam para este caminho. A utilização de dados genômicos [24], [26], [29] ou de expressão gênica [30] revelarem alguns possíveis marcadores moleculares baseados em testes estatísticos. Também métodos de inteligência computacional (*Support Vector Machine*) têm sido empregados na tentativa de diferenciação das formas clínicas da dengue [31].

Todas estas técnicas se diferenciam pelo tipo de dados utilizados como parâmetros (clínicos) ou pela técnica utilizada para análise, algumas técnicas estatísticas, e outra de computação inteligente. É difícil fazer uma comparação entre os resultados destas, uma vez que os dados utilizados são todos diferentes.

Capítulo 3

Inteligência Computacional para Problemas de Classificação em Bioinformática

A Inteligência Artificial (IA) é uma área da Ciência da Computação que estuda o desenvolvimento de técnicas computacionais inspiradas na capacidade humana de resolver problemas e, sobretudo, aprender com eles. Entre os estudiosos da área existe um debate que divide a área de pesquisa entre IA fraca e IA forte. O argumento desta divisão reside no fato de ser possível ou não construir uma máquina consciente.

Neste capítulo, serão apresentadas variações de Redes Neurais Artificiais, uma importante técnica de inteligência computacional, com potencial de aplicação na busca de assinaturas ou métodos genéticos preditivos. Nosso objetivo também é o de fornecer uma visão geral à cerca da aplicação dessas técnicas no domínio da bioinformática.

3.1 Técnicas de Inteligência Computacional

As técnicas de computação inteligente tiveram seu desenvolvimento logo após a segunda guerra mundial com a publicação do artigo de Alan Turing intitulado "*Computing Machinery and Inteligence*" [32]. Inspirados na capacidade de aprender do homem foram desenvolvidos alguns modelos computacionais. Porém só com o surgimento do computador moderno estas técnicas se tornaram aplicáveis na solução de problemas reais. Atualmente a computação inteligente é utilizada para uma gama de aplicações que vão desde sistemas utilização em jogos de estratégia como xadrez (em 1997 um programa de computador que derrotou o campeão

mundial de xadrez Garry Kasparov) até aplicações de suporte a decisão que utilizam dados históricos para melhorar o desempenho no futuro [33].

Como as técnicas de computação inteligente apresentam excelentes resultados na tarefa de classificação, são candidatos naturais para auxiliar nos desafios encontrados na busca por assinaturas ou métodos preditivos genéticos em doenças. São muitas as técnicas consagradas de inteligência computacional tais como Árvores de Decisão, Raciocínio baseado em casos, Redes Bayesianas e Otimização baseado em enxame de partículas. Entretanto, as redes neurais artificiais, por ser conhecidamente uma ferramenta poderosa em classificação de padrões, foram selecionadas e utilizadas para o estudo genético (cap. 4) e serão melhor discutidas nesta revisão.

3.2 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNA) são modelos matemáticos que se assemelham às estruturas neuronais biológicas que têm capacidade computacional adquirida por meio de aprendizado e conseqüente, generalização [34]. Inspirados no sistema nervoso as RNAs são constituídas por unidades de processamento simples chamadas neurônios (assim como no modelo biológico). A força da conexão entre os neurônios é conhecida como pesos sinápticos e o local de armazenamento do conhecimento adquirido [34]. A principal virtude das redes neurais artificiais é a sua capacidade de generalização, ou seja, produzir respostas adequadas a dados de entrada não conhecidos previamente, de forma a alcançar um objetivo desejado. Essa capacidade, derivada da combinação de neurônios e sinapses em grandes redes, torna possível resolver problemas computacionais complexos.

São algumas das principais características das redes neurais artificiais que as fazem ter desempenho superior ao de modelos convencionais [35]:

1 A representação do conhecimento é interna à própria rede, não dependendo de estruturas adicionais;

2 A não-linearidade, que possibilita sua aplicação em sistemas complexos;

3 Adaptabilidade: uma vez construída uma rede eficiente para determinada aplicação, pode ser utilizada em tempo real, sem a necessidade de ter sua arquitetura alterada a cada atualização de dados;

4 Capacidade de auto-aprendizado. As redes neurais artificiais, depois de treinadas, não necessitam de conhecimentos de especialistas para tomar decisões, baseiam-se apenas nos exemplos históricos que lhes são fornecidos;

5 Tolerância a falhas: como os elementos de processamento da rede operam em paralelo, a destruição ou defeito em algum deles não torna a rede inoperante, podendo até mesmo não causar grandes problemas no funcionamento geral do sistema;

6 Imunidade a ruídos: dados reais sempre contêm ruído; as redes conseguem separar o ruído da informação relevante.

Historicamente, as pesquisas sobre redes neurais artificiais tiveram início com base no clássico artigo de Warren McCulloch e William Pitts, de 1943 *A logical calculus of the ideas immanent in nervous activity*, no qual descrevem um modelo matemático para os neurônios [36]. Esse modelo é o utilizado para a implementação dos neurônios artificiais.

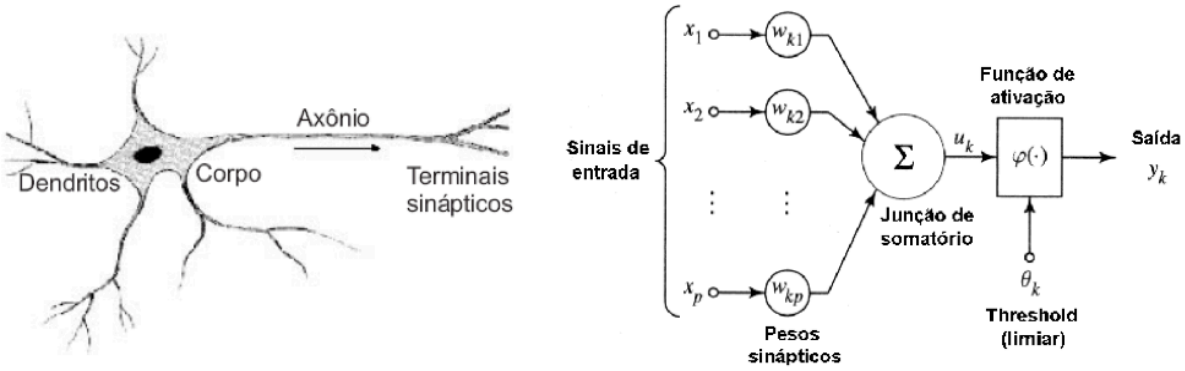


Figura 2. Modelo de neurônio biológio e proposto por McCulloch e Pitts, mostrando a similaridade da forma de entrada de dados, e a saída por outra extremidade após a inferência.

O neurônio biológico pode ter seu funcionamento simplificado para facilitar a compreensão. Os dendritos recebem os “sinais” de entrada do neurônio, estes sinais são decorrentes de estímulos enviados por outros neurônios. Os “sinais“ influenciam de forma diferenciada cada neurônio. O grau de influências de um neurônio sobre

outro se modifica com o tempo em decorrência principalmente da aprendizagem. Os estímulos recebidos pelo corpo do neurônio são “somados“ e caso superem um limiar de ativação este neurônio “dispara” propagando estímulos para outros neurônios ligados a este. O neurônio artificial proposto por McCulloch e Pitts tem um funcionamento muito semelhante ao modelo biológico sendo composto por sinais de entrada, pesos Sinápticos, uma função de somatório, uma função de ativação (ou ‘threshold’) e a saída que se comportam como terminais sinápticos, dendritos, corpo e axônio.

Os sinais de entrada são as variáveis que serão utilizadas na RNA. Os pesos sinápticos são valores que são ajustados durante a aprendizagem da rede. Estes valores são multiplicados pelos valores de entrada e determinam a influência do sinal de entrada no neurônio. Os pesos podem assumir valores positivos e negativos. Os valores positivos podem ser associados a sinapses excitatórias e os de valores negativos, à sinapse inibitória dos neurônios biológicos. A função de junção somatório tem a finalidade de somar a contribuição de cada sinal de entrada multiplicada pelo seu peso gerando o valor que será passado como parâmetro para função de ativação. A função de ativação consiste numa função que define um limiar de ativação, ou seja, define se o neurônio será ativado. Várias funções podem ser utilizadas para este fim tais como: rampa, degrau, sigmóide ou logística (utilizada neste estudo), e hiperbólica como mostrado na figura 3.

O modelo desenvolvido na década 40 foi bastante disseminado e em 1958, a partir da publicação do livro “*Principles of neurodynamics* de Rossenblatt” onde sistematizou várias idéias sobre os *Perceptrons*, que são modelos de neurônios baseados no modelo de McCulloch e Pitts [37]. Após este período os algoritmos de Perceptron foram desenvolvidos e aperfeiçoados. No final da década de 60, entretanto, um estudo de Minsky e Papert (1969) [38] demonstraram que os modelos apresentados até então resolviam apenas problemas de associação de padrões linearmente separáveis. Com esta constatação as RNAs passaram as duas décadas seguintes relegadas a um plano secundário.

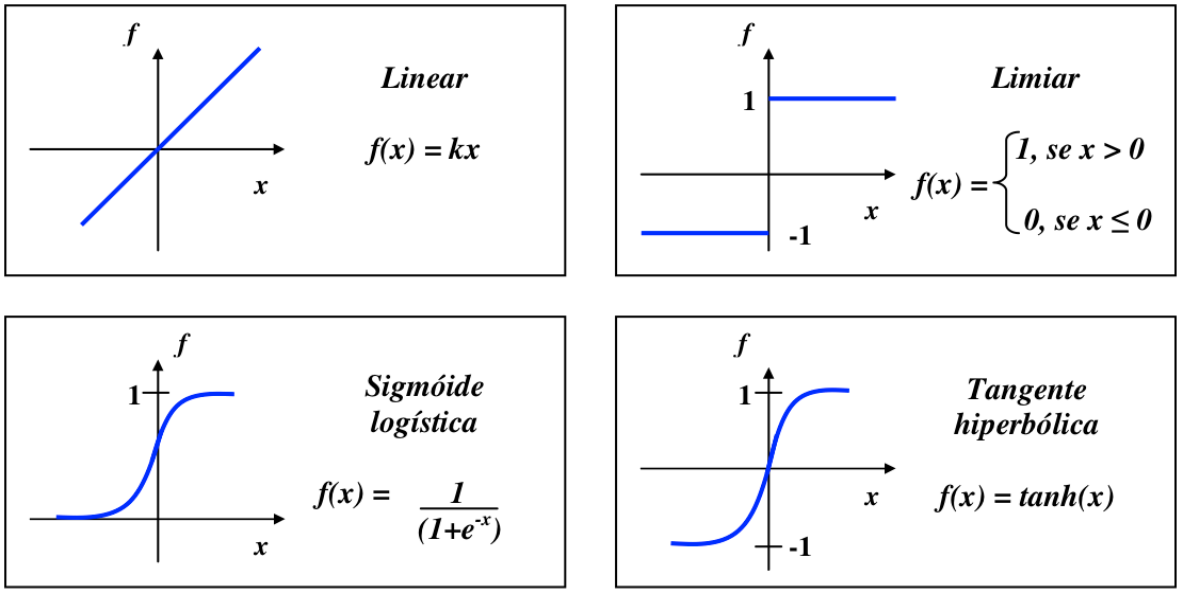


Figura 3. Gráficos das principais funções de ativação que podem ser utilizadas em uma RNA. As funções limiar e tangente hiperbólica variam de -1 a 1, e a sigmóide logística de 0 a 1.

Em anos mais recentes a contribuição de Rumelhart, Hinton e Willians [39] que resolveram o problema de associação a padrões não-linear com a criação da Regra Delta Generalizada, mais conhecida como Algoritmo de correção de erros de retropropagação para Redes Perceptron de várias camadas (Multilayer Perceptron) foi de grande relevância no aperfeiçoamento das RNA. Associada ao desenvolvimento da informática esta abordagem impulsionou o desenvolvimento das RNAs que vem sendo pesquisada e aplicada em diversas áreas desde então [40].

As pesquisas sobre as RNAs demonstraram várias formas de organizar os neurônios, criando assim o conceito de topologia das Redes Neurais Artificiais. A topologia consiste na determinação das características da rede como: a quantidade de neurônios, a quantidade de camadas (*Multilayer Perceptron*) ou de forma cíclica (Redes de Hopfield). A topologia ou arquitetura das RNAs apresenta um gama de alternativas que possuem características que favorecem a aplicação em domínio específico. Ou seja, a escolha da RNA mais apropriada depende do problema a ser resolvido.

Além da topologia da RNA é importante definir a forma de aprendizagem da rede. Basicamente existem duas possibilidades: o aprendizado supervisionado e o não supervisionado. No método supervisionado um conjunto de entradas e as saídas

esperadas são fornecidos para que a rede ajuste os seus pesos (sinápticos). Já no método não supervisionado não existe o supervisor para fornecer os resultados esperados para um conjunto de entradas. Assim a rede deve buscar associações relevantes a partir da extração das propriedades estatísticas exclusivamente com os dados de entrada, criando classes e grupos representativos.

Na seção seguinte será realizada uma apresentação mais detalhada do *Perceptron* de Camada Única, e das Redes *Multilayer Perceptron* (MLP). As MLP são as RNAs mais utilizadas, por obter bons resultados para a maior parte das aplicações.

3.2.1 *Perceptron* de Camada Única

As redes que utilizam neurônios do tipo *perceptron* com uma única camada são úteis na solução de problemas que admitem separação de classes por hiperplanos, ou seja, padrões linearmente separáveis. Rosenblatt [41] demonstrou que sempre que um *perceptron* (figura 4) for treinado com padrões retirados de duas classes linearmente separáveis, será capaz de convergir para uma superfície de decisão formada por um hiperplano entre essas duas classes. Quando o *perceptron* é construído com um único neurônio, esse somente será capaz de classificar padrões apenas entre duas classes. Porem, com o aumento do numero de neurônios, o numero de classes separáveis aumentara proporcionalmente.

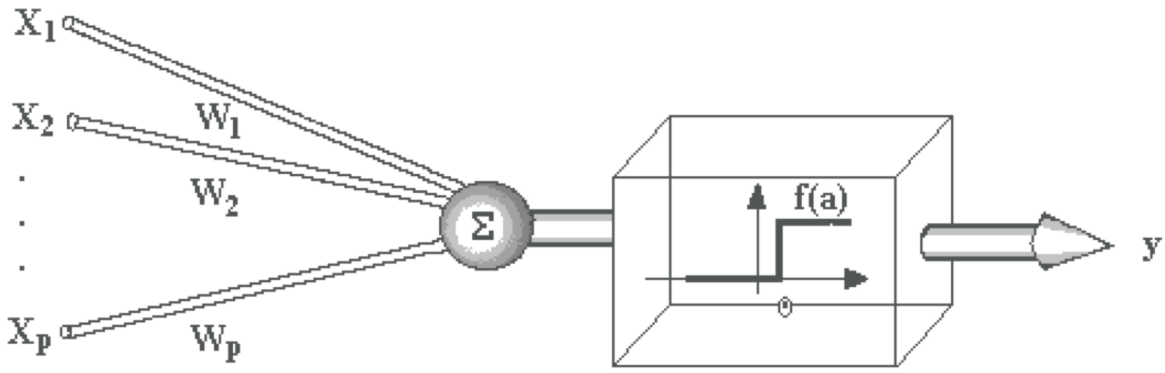


Figura 4. Representação de um *perceptron* de Rosenblatt com apenas um neurônio, e utilizando a função discriminante a função linear.

3.2.2 *Multilayer Perceptron* (MLP)

As Redes MLP são constituídas de no mínimo três camadas de neurônios, sendo uma camada de entrada e uma de saída e no mínimo uma camada intermediária (ou escondida). Os sinais de entrada são propagados de uma camada para a seguinte até atingir a saída da rede. É importante observar que cada neurônio de uma camada esta ligado a todos os neurônios da camada seguinte (fig. 5).

A rede MLP tem sido utilizada com grande sucesso na solução de problemas com alto grau de não-linearidade. No treinamento supervisionado das MLP o algoritmo de retro-propagação do erro para ajustar os pesos das conexões entre os neurônios é utilizado. Basicamente este processo é constituído por duas etapas. Na primeira o sinal é propagado da entrada para a saída (*feedforward*) e posteriormente a retro-propagação do erro (*backpropagation*) [40].

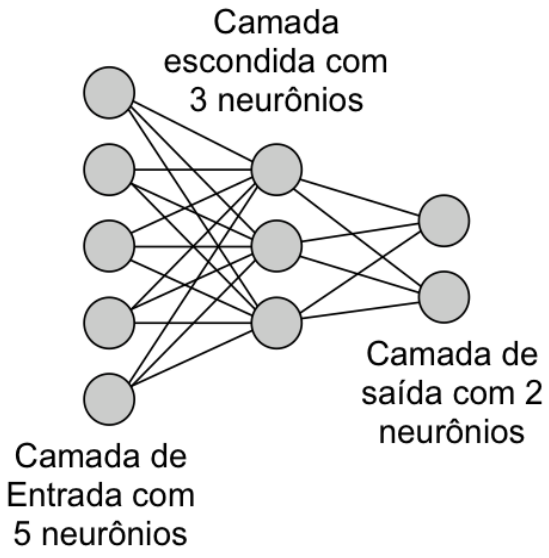


Figura 5. Arquitetura de uma RNA Multilayer Perceptron com apenas uma camada escondida com 3 neurônios. Esta arquitetura é largamente utilizada em problemas de clasificação.

Na fase do *feedforward* o vetor de entrada de dados é apresentado e processados pela camada de entrada, sendo, então, propagados para a camada seguinte. Este processo se repete a cada camada até atingir a camada de saída. Os sinais nesta etapa (*feedforward*) são conhecidos como sinal funcional e são mantidos como pesos estáticos. Na fase do *backpropagation* uma regra de correção de erro é aplicada. Uma das regras utilizadas é a do gradiente descendente, que

busca minimizar o erro médio quadrático. Após o cálculo do valor a ser ajustado nos pesos, o sinal percorre o caminho inverso da rede ajustando o peso dos neurônios. Nesta etapa o sinal é então chamado de sinal de erro.

Inicialmente, o processo de desenvolvimento de rede MLP consiste na coleta de dados. Como as RNAs trabalham apenas com valores numéricos é necessário codificar os dados categóricos (i.e pré-processamento). Finalmente, os dados devem ser divididos para utilizar uma parte no treinamento da rede e outras duas partes para teste e validação.

No treinamento de uma rede MLP dois parâmetros são importantes para a qualidade do resultado: a taxa de aprendizagem e o momento. A partir desses parâmetros o treinamento é realizado até atingir os critérios para encerramento do treinamento. A parada no treinamento pode ser realizada por vários artifícios, por exemplo: a partir da validação cruzada que verifica se o treinamento está produzindo melhorias no resultados. Porém, além da validação cruzada, o critério do número de épocas é utilizado como limite superior do treinamento. Ou seja, no caso da validação cruzada não parar o treinamento este será finalizado ao atingir o número de épocas estabelecido. É importante observar que a parada do treinamento pelo número de épocas é um indício de que os ajustes não estão produzindo bons resultados ou por subajustamento ou superajustamento.

Firmados os conhecimentos da técnica que vamos empregar, os dados de entrada precisam estar ajustados para sua execução, e essa etapa será abordada no capítulo a seguir.

Capítulo 4

Bioinformática e a Dengue: Modelos Univariado e Multivariado para definição de risco de formas severas da dengue

Como afirmado anteriormente no capítulo 2, a classificação das formas clínicas da dengue utilizando como parâmetros os aspectos clínicos pode ter certo grau de imprecisão devido a similaridade dos sintomas iniciais da infecção pelo vírus dengue com outras doenças, e como alternativa a esta similaridade, estudos tem demonstrado a participação de fatores genéticos no desenvolvimento do fenótipo severo em dengue. Assim, o objetivo deste trabalho foi pré-processar dados de genotipagem em massa de pacientes com dengue, e criar um modelo classificador das formas clínicas da dengue, como alternativa aos métodos de classificação tradicionais (com critérios clínicos).

4.1 Coorte de Pacientes

O Laboratório de Virologia e Terapia Experimental (LaViTE), do Centro de Pesquisas Aggeu Magalhães/FIOCRUZ, realiza atualmente um estudo de coorte composta por pacientes voluntários com dengue clássica (DC), dengue clássica complicada (DCC) e febre hemorrágica da dengue (FHD) da cidade do Recife-PE. Os voluntários da coorte foram cadastrados na pesquisa quando diagnosticados com dengue em três hospitais participantes do estudo (Hospital Santa Joana, Hospital Esperança e Instituto Materno-Infantil de Pernambuco – IMIP). Esses voluntários foram examinados do primeiro ao quarto dia de detecção da doença, com avaliações adicionais a cada 24-48h, dependendo do estado clínico dos pacientes. A cada

voluntário é solicitada uma doação de 4 a 5 amostras de sangue, dos dias 1º ao 30º após o início dos sintomas.

A caracterização clínica inclui a história e avaliação clínica, hemograma, contagem de plaquetas, enzimas hepáticas, sorologia para dengue por ELISA-IgM (Bio-Manguinhos/Fiocruz) e ELISA-IgG (PANBIO), isolamento viral em células C6/36, e diagnóstico molecular de dengue por RT-PCR viral. Cada voluntário e respectiva amostra recebe uma numeração, sendo os dados introduzidos em um sistema digital, em tempo real, e validados dentro de 24h. As informações são integradas a um banco de dados que inclui os dados clínicos completos, resultados de pesquisa e o respectivo inventário de amostras de PBMCs congeladas [9].

Este trabalho conta com uma população teste composta por 105 pacientes, dos quais 30 tiveram DC, 49 DCC e 26 FHD. Na tab. 1 mostra como os dados estão distribuídos em relação a todas as características clínicas e de cada paciente de forma equitativa. Os pacientes DCC foram incluídos no grupo dengue clássica, assim como é recomendado pela OMS. Uma segunda coorte de pacientes totalmente independentes da primeira foi utilizada no estudo para efeito de controle genético. Esta foi composta por 99 indivíduos, os quais depois da caracterização clínica mostraram-se isentos de infecção pelo vírus da dengue.

Este projeto foi derivado de um projeto principal intitulado *Identificação de marcadores biológicos preditivos de respostas clínicas ao dengue através de análise proteômica* cujas considerações éticas foram aprovadas pelo Conselho Nacional de Ética em Pesquisa (CONEP), sob o no. 4909.

Tabela 1. Distribuição da quantidade de pacientes dengue quanto ao sexo, tipo de infecção, diagnóstico clínico e idade. Mostra que a quantidade de paciente está distribuído equitativamente por todas as características.

Sexo		Tipo de Infecção		Diagnóstico Clínico			Idade
Homem (n= 54)	Mulher (n = 51)	Primária (n = 44)	Secundária (n = 60)	DC (n=30)	DCC (n=49)	FHD (n=26)	Anos
9	7	9	6	10	5	1	0 - 15
3	4	3	4	-	5	2	16 - 20
6	4	8	2	3	4	3	21 - 25
4	10	6	8	3	6	5	26 - 30
9	4	6	7	2	8	3	31 - 35
5	6	5	6	3	5	3	36 - 40
8	3	3	8	3	5	3	41 - 45
5	3	1	7	3	5	-	46 - 50
5	10	3	12	3	6	6	51 - 84

4.2 Tratamento dos dados da genotipagem em massa

O DNA dos pacientes dengue foi enviado para ser submetido a uma técnica de genotipagem em massa (Illumina®) na *University of Washington*. Os resultados destas análises não foram suficientes para serem analisados com os softwares especializados dado que os softwares comerciais exigem um elevado número de pacientes e de dados genômicos como *input*. Assim, se fez necessária a construção de uma ferramenta de apoio às analises dos dados genéticos destes pacientes.

Os dados da genotipagem foram constituídos por 322 *loci* gênicos de 105 pacientes infectados pelo dengue e 99 pacientes não-dengue. Informações biológicas como sexo e idade; tipo da infecção (primária ou secundária) e o diagnóstico clínico foram também considerados. Com este conjunto de dados foi construído um banco de dados relacional, separando em tabelas diferentes os dados

genéticos, dos demais dados biológicos e fenotípicos dos pacientes, aliado a um sistema simples de gerenciamento. Para tal foi utilizado o *FileMaker Pro,* ferramenta proprietária de fácil manipulação, já utilizada em outros projetos do LaViTE. A fig. 6 mostra as três tabelas do banco de dados; na tabela paciente estão os dados de cada paciente, na tabela *alleles* estão os dados genotípicos, e na tabela *locus*, é onde estão resultados de cálculos estatísticos.

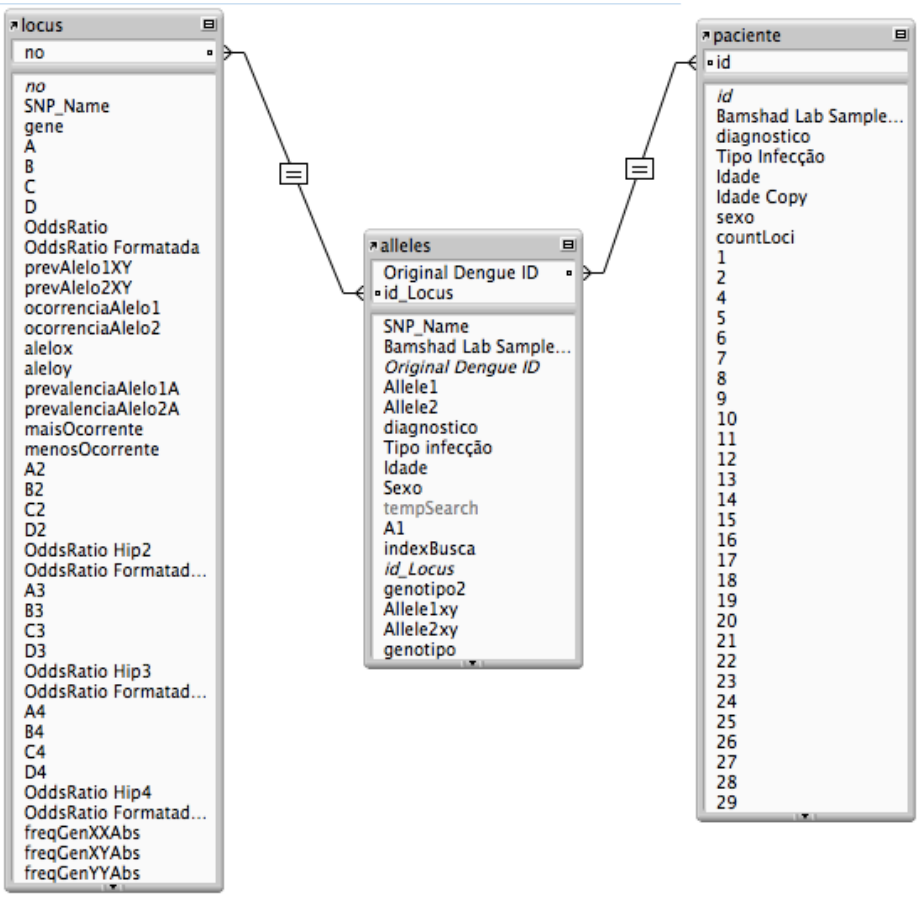


Figura 6. Tabelas relacionadas do banco de dados utilizado para segmentação dos dados de genotipagem em massa de pacientes dengue, desenvolvido utilizando a ferramenta *FileMaker Pro* v.10

4.3 O problema da padronização dos dados genéticos

Genes são segmentos de DNA que carregam informações genéticas. Em cada gene existem vários *loci* gênicos, formados por pares de base de nucleotídeos. No DNA

estas bases podem ser adenina (A), guanina (G), citosina (C) e timina (T). É neste "alfabeto" que os dados da genotipagem são representados, em que o *locus* que tiver dois alelos iguais é chamado homozigoto e o *locus* que tiver dois pares de base diferentes é heterozigoto.

Baseado nessas informações, foi realizada uma contagem de alelos dos pacientes por *locus*, elegendo-se então o alelo prevalente e o não prevalente na população. A partir daí fez-se a seguinte transformação na base de dados: em vez de um conjunto de pares de letras (representando alelos), as informações gênicas foram representadas por *locus* homozigoto prevalente (XX), *loci* heterozigoto (XY), *loci* homozigoto não prevalente (YY), e *loci* não reativo (indefinido ou NR). Esta estratégia permitiu a padronização das informações de todos os *loci*, independentemente dos alelos (ou das bases nitrogenadas A,T, C ou G; ou indefinido, se não reagente) que definiam cada variação alélica, uma importante adaptação para realização de cálculos simples, como os de freqüências alélicas e genotípicas para cada *locus* (importantes dados genéticos).

4.3.1 As quatro hipóteses genéticas testadas para cada locus gênico

O problema em questão é encontrar uma forma de classificar/predizer as formas clínicas da dengue, baseado nas informações genéticas dos pacientes. Uma possível solução seria utilizar um único *locus* como elemento de "predição" clínica. Para isso, foi utilizada uma interpretação genética onde duas características genotípicas (definição de dois "grupos" genotípicos, a partir dos três genótipos de cada *locus*; Ex. AA x AT +TT) seriam comparadas a duas características fenotípicas (definição de dois "grupos" clínicos, a partir das três formas de manifestação da dengue; Ex. DC x DCC + DH) aplicados em tabelas de contingência 2x2 [42], variando-se a forma de agregação das diferentes categorias em cada uma das hipóteses, tal como apresentado na tab 2. Nas hipóteses 1 (Hip1) e 2 (Hip2) foram agrupadas as formas genotípicas homozigota menos ocorrente (YY) e a forma heterozigota (XY), mantendo a forma homozigota mais ocorrente (XX) isolada; já nas hipóteses 3 (Hip3) e 4 (Hip4), agrupou-se a forma XX com a XY, mantendo a forma YY isolada. Mantendo-se o mesmo raciocínio, as Hip1 e Hip3, tiveram as formas clínicas FHD e DCC agrupadas, criando a forma DG, seguindo a norma do MS. Nas Hip2 e Hip4, a forma clínica FHD ficou isolada das formas clínicas DCC e DC, que

segundo as recomendações da OMS, formam a forma clínica DC. Este tipo de avaliação genética em agrupamentos permite a definição acerca dos efeitos de cada alelo, se dominante (expressa seu fenótipo tanto em homozigose como em heterozigose) ou recessivo (só se expressa em homozigose); ou se o efeito do alelo de risco/proteção se dá sobre as formas severas de dengue (DH+DCC) ou apenas sobre a dengue hemorragia (FHD).

Então de forma prática, o objetivo do agrupamento do problema em hipóteses, é de separar o problema em grupos, a fim de estudar o efeito do polimorfismo isolando as formas clínicas, e também os genótipos, medindo se o efeito se apresenta em alguma das hipóteses.

Tabela 2. As quatro hipóteses genéticas. As hipóteses 1 (Hip1) e 2 (Hip2) foram agrupadas as formas genotípicas homozigota menos ocorrente (YY) e a forma heterozigota (XY), mantendo a forma homozigota mais ocorrente (XX) isolada, e a permutação dos outros atributos nas outras tabelas

Hip1	XX	XY+YY		**Hip2**	XX	XY+YY
FHD+DCC	A	B		FHD	A2	B2
DC	C	D		DCC+DC	C2	D2
Hip3	XX+XY	YY		**Hip4**	XX+XY	YY
FHD+DCC	A3	B3		FHD	A4	B4
DC	C3	D3		DCC+DC	C4	D4

4.4 Seleção das Variáveis estatísticas

Em face das hipóteses criadas, pode-se utilizar técnicas estatísticas para seleção dos *loci,* que neste caso são os atributos, na busca de um marcador uni variado para classificação dos dados (definição de risco isolado). Então foi desenvolvido no sistema de gerenciamento do banco de dados (SGBD) o cálculo para estabelecer o *Odds-Ratio (indicador de risco),* o Equilíbrio de Hardy Weinberg (indicador genético do estado de equilíbrio da população), e o Teste de Fisher (método estatístico para determinação do nível de significância do OR) para validação de cada *locus (i.e. dos efeitos dos alelos variantes em diferentes combinações genotípicas).*

O *Odds-Ratio* é a medida do efeito causado, demonstrando a força da associação de duas características diferentes entre si [43]. Com o calculo do *Odds-Ratio* é Possível predizer o quanto aquele *locus* (na verdade genótipos que apresentam o alelo variante) pode ser importante para a definição da forma clínica da dengue, isto é, quanto maior o *Odds-Ratio* maior será a freqüência dos genótipos com alelos variantes num tipo de dengue (ex. DC) em relação à outra forma clínica (FHD) da hipótese em consideração. O calculo do Odds-Ratio se dá conforme a figura abaixo.

	$Y=1$	$Y=0$
$X=1$	p_{11}	p_{10}
$X=0$	p_{01}	p_{00}

$$\frac{p_{11}/(p_{11}+p_{10})}{p_{10}/(p_{11}+p_{10})} \Big/ \frac{p_{01}/(p_{01}+p_{00})}{p_{00}/(p_{01}+p_{00})} = \frac{p_{11}p_{00}}{p_{10}p_{01}}$$

Figura 7. A esquerda uma tabela de contingência 2x2, e a direita a formula do calculo do *Odds-Ratio*, que em resulta a medida da razão entre as grandezas.

No calculo do *Odds-Ratio*, pode acontecer de a razão ser menor que 1, isso quando o divisor é maior que o dividendo, significando que a razão neste caso é inversa, ou seja, ao invés de medir o risco daquela configuração genética de ter a forma clínica mais grave, mede a proteção. Mas para efeito do *ranking* foi criada uma variável chamada de *odds-ratio* formatado, que nos casos da razão ser inversa, a divisão é invertida, resultando em uma taxa direta.

Para determinação do nível de significância do *Odds-Ratio*, foi aplicado o teste exato de Fisher. Este teste mede a significância estatística em análises de tabelas de contingência quando as amostras são pequenas. Na figura 8 é mostrada a formula do teste exato de Fisher.

	men	women	total
dieting	a	b	$a+b$
not dieting	c	d	$c+d$
totals	$a+c$	$b+d$	n

$$p = \binom{a+b}{a}\binom{c+d}{c} \Big/ \binom{n}{a+c} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

Figura 8. Fórmula do teste exato de Fisher com um exemplo da comparação entre homens e mulheres que fazem ou não dieta. A formula é a razão entre o produtório do fatorial da soma das quantidades de homens e mulheres que fazem ou não dieta agrupados dois a dois; e o produtório do fatorial de cada um e o total.

Em 1908 Godfrey Harold Hardy e Wilhelm Weinberg afirmaram que em uma população mendeliana dentro de determinadas condições, as freqüências alélicas permanecerão constantes ao passar das gerações. Independentemente de um alelo ser raro ou freqüente, sua freqüência permanecerá no tempo, desde que essas condições sejam mantidas [44]. Estes autores, em trabalhos independentes, estabeleceram um calculo de medida do estado de equilíbrio das populações com base nas freqüências alélicas e genotípicas observadas e esperadas (teóricas) do grupo em estudo. O calculo é estabelecido de acordo com a figura 9.

		Fêmeas	
		A (*p*)	**a (*q*)**
Machos	**A (*p*)**	AA (p^2)	Aa (pq)
	a (*q*)	aA (qp)	aa (q^2)

- $f(\mathbf{AA}) = p^2$
- $f(\mathbf{Aa}) = 2pq$
- $f(\mathbf{aa}) = q^2$

Figura 9. Calculo do Eq. H-W. Tendo os valores calculados das freqüências genotípicas (XX - f(AA); XY – f(AA); YY – f(AA)) se calcula as freqüências esperadas, aplicando-os ao teste do Qui-Quadrado, os valores maiores iguais a 95% atendem ao Eq. H-W.

Com todos os dados estatísticos (OR, Fisher, e Eq.-H.W.), pode-se fazer algumas análises acerca do grau de importância de cada *locus (locus* contendo alelos variantes) criando um *ranking* para o problema de classificação das formas clínicas. Essa métrica ranqueia o *locus* que melhor pode classificar dentro da

hipótese estabelecida, de acordo com *Odds-Ratio*. Foi definido, então, que para este estudo seriam relevantes apenas os *loci* que estavam em Equilíbrio genético (Eq. H-W), com Odds-Ratio maior ou igual a 2,4 (ou entre zero e 0,41666, caso a predominância seja inversa), e que tinham significância estatística mínima a 95%, segundo teste exato de Fisher.

4.5 Resultados das análises univariadas

Após este ponto de corte, verificamos que 47 *loci* tiveram um alto grau de *Odds-Ratio*, respeitando-se os critérios ditos acima. Seis deles foram considerados de alta relevância (Fisher a 1%) biológica para dengue, todos eles representando genes relacionados à imunidade inata tais como MBL2, e o IFNg (tab.3). Outros *loci*, entretanto, apresentaram elevados ORs de forma "artificial". Nestes casos, um dos elementos da matriz 2x2 era zero, ou muito baixo. Uma das interpretações para isso, é que a amostra de pacientes é muito baixa para encontrar indivíduos variantes (freqüência de alelos variantes é muito baixa ou inexistente) na população.

Tabela 3. Resultados univariados (6 loci), mostrando o nome do gene onde se encontra o *loci*, o posicionamento, a apresentação dos alelos, o efeito variante, a taxa e o tipo de efeito (se proteção ou risco), o efeito molecular e a forma clínica associada

Gene	***locus***	**Tipo**	**Alelos**	**Efeito do alelo variante**	**Risco/Proteção**	**Efeito molecular**	**Associação clínica**
IFNg	rs2069727	3'UTR	A/G	GG/AG	OR 8,3 proteção	Alta expressão	DG
MBL2	rs10824793	Intron	A/G	AA	OR 3,14 proteção	Baixa expressão	DG
OASL	rs3213545	Synon	C/T	TT/TC	OR 3,7 risco	?	DG
STAT1	rs2066804	Intron	C/T	TT	OR 15,4 proteção	?	DG
FAAD	rs1131715	3'UTR	C/T	CC	OR 9,15 risco	?	DG
INDO	rs3739319	Intron	A/G	AG/AA	OR 3,5 proteção	?	FHD

Embora os dados apresentados na tab.3 sejam por si só de grande importância, outras questões sobressaem quando pensamos numa análise genética mais ampla. Tal como outras doenças a dengue é, antes de qualquer coisa, multigênica e multifatorial. Assim, o que pode acontecer quando, por exemplo, um indivíduo tiver um genótipo com alelo que confere risco pra FHD em um dado *locus*, e num outro *locus* este mesmo indivíduo apresenta genótipo com alelos de proteção para a FHD? Os efeitos se anulam? Ou um *locus* se sobressai em relação ao outro?

Uma alternativa pra se tentar interpretar as diversas formas de interação entre *loci* polimórficos (sinergias ou antagonismos) é a construção de modelos matemáticos/bioinformáticos que possibilite este entendimento multivariado (multigênico) da doença. Neste trabalho foi desenvolvido um modelo baseado em redes neurais, apresentado e discutido na próxima sub-seção (4.6).

4.6 Modelo de rede neural para a análise genética multivariada (multigênica) em dengue

O Weka, é um software desenvolvido por pesquisadores da Universidade de Waikato na Nova Zelândia [45], e possui uma coleção de algoritmos para Mineração de Dados. Como se trata de uma ferramenta *open source,* e já está disponibilizada na *web* a muito tempo, é bastante utilizada em projetos e tem atualizações constantes, com melhoras tanto nos algoritmos, quanto na parte gráfica. Por tal, o Weka foi escolhido pra ser a base do projeto desenvolvido em Java, utilizando o código fonte do Weka, e fazendo a chamada das funções e métodos necessários. Para isso se fez necessário um estudo do código fonte dos métodos para classificação utilizando Redes Neurais Artificiais com MLP. Visto isso, passamos para o estudo do formato de dados utilizados como entrada para o Weka, o ARFF.

No desenvolvimento deste projeto, as 4 hipóteses (tab. 2) poderiam combinar entre si com as formas de agrupamento dos genótipos e as formas clínicas, por isso foi criada uma lista com as possíveis organizações dos dados, para serem exportados, transformados em ARFF e utilizados como *datasets* para a RNA. A

tabela 4 mostra a seqüência em que os dados foram exportados do banco de dados, em formato de arquivo .mer, onde os nomes das variáveis ficam na primeira linha, e os atributos entre aspas-duplas, todos separados por ponto e virgula. Este aquivo .mer foi transformado para arquivos de variáveis separadas por vigulas (.csv), e utilizado um método do Weka que converte do padrão .csv para arff. Esta ferramenta de transformação cria uma primeira linha com uma relação precedida de um @ (@*relation*) e o nome do arquivo. Reconhece que na primeira linha estão os *labels* de cada atributo, criando para cada atributo de entrada ou saída uma linha com: @*attribute,* o nome do atributo, e o tipo (numérico, binário, ou nominal). Como neste exemplo: @attribute 6H numeric. Uma linha após os atributos estão os dados (@*data*) todos separados por virgulas, assim como no padrão .csv. Isso feito para todos os *datasets,* ficando assim os dados prontos para ser utilizados na rede neural.

Tabela 4. Tabela que mostra a ordem, e como os dados foram exportados do banco de dados, separando as maneiras que os dados poderiam se configurar de acordo com as hipóteses criadas, a organização dos fenótipos, se seriam todas as formas clínicas, ou agrupadas em DC e FHD, ou DC e DG, e os genótipos, se agrupados, ou todos independentes

Dataset	*Loci*	Hip.	Fenótipo	Genótipos
1	p. n. letais ^ fisher <=0,05	-	DC DCC FHD	XX XY YY NR
2	p. n. letais ^ fisher <=0,05	1 3	DC DG	XX XY YY NR
3	p. n. letais ^ fisher <=0,05	2 4	DC FHD	XX XY YY NR
4	p. n. letais ^ fisher <=0,05	1	DC DG	XX (XY+YY) NR
5	p. n. letais ^ fisher <=0,05	3	DC DG	(XX+XY) YY NR
6	p. n. letais ^ fisher <=0,05	2	DC FHD	XX (XY+YY) NR
7	p. n. letais ^ fisher <=0,05	4	DC FHD	(XX+XY) YY NR
8	Odds-Ratio >=2,4	-	DC DCC FHD	XX XY YY NR
9	Odds-Ratio >=2,4	1 3	DC DG	XX XY YY NR
10	Odds-Ratio >=2,4	2 4	DC FHD	XX XY YY NR
11	Odds-Ratio >=2,4	1	DC DG	XX (XY+YY) NR
12	Odds-Ratio >=2,4	3	DC DG	(XX+XY) YY NR
13	Odds-Ratio >=2,4	2	DC FHD	XX (XY+YY) NR
14	Odds-Ratio >=2,4	4	DC FHD	(XX+XY) YY NR
15	top 6	-	DC DCC FHD	XX XY YY NR
16	top 6	1 3	DC DG	XX XY YY NR
17	top 6	2 4	DC FHD	XX XY YY NR
18	top 6	1	DC DG	XX (XY+YY) NR
19	top 6	3	DC DG	(XX+XY) YY NR
20	top 6	2	DC FHD	XX (XY+YY) NR
21	top 6	4	DC FHD	(XX+XY) YY NR

O desenvolvimento da ferramenta que utiliza a inteligência computacional foi no Eclipse, em Java [46], criando um pacote com os métodos de chamadas das funções *Multilayer Perceptron*, em que uma classe realiza o treinamento da MLP, e

outra faz o teste da rede criada resultando as estatísticas do algorítmo. O objetivo foi encontrar o agrupamento de *loci* que em conjunto (multivariado), considerando as hipóteses apresentadas na tab. 2, que melhor classificasse o fenótipo do paciente.

A rede neural foi executada utilizando apenas uma camada oculta, e a quantidade de neurônios foi empiricamente definida com a metade da soma de neurônios da camada de entrada com a camada de saída. Outros parâmetros da RNA também foram definidos empiricamente e ajustados ao passo que eram executadas, e haviam melhoras nos resultados. Foi estabelecido então que seriam utilizados 500 ciclos máximos para treinamento, juntamente com 25% dos dados para validação cruzada, O treinamento prosseguiu até que o erro na validação (calculado pelo erro médio quadrático) aumentasse continuamente por no máximo 400 ciclos, não ocorrendo isso seguiria até o máximo predeterminado. Essa técnica é utilizada para evitar o *overfitting* (treinamento em excesso) dos dados, e para que o modelo não “decore” os dados apresentados no treinamento, trazendo assim uma RN com uma boa capacidade de generalização [47]. Um parâmetro que mereceu uma maior atenção foi a taxa de aprendizado, pois se utilizando um valor muito baixo o processo de aprendizado pode ser muito lento, alem de ter uma grande chance de ficar preso a mínimos locais. Por outro lado, se for muito alto, pode gerar uma instabilidade numérica durante o treinamento, e para auxiliar neste caso, foi utilizado também o momento, que possibilita aumentar os passos de treino sem aumentar a instabilidade do algoritmo [47]. A taxa de aprendizado foi determinada como 0,3, e o momento 0,2.

Para a aplicação dos dados na MLP, primeiro foi necessária a execução de pré-processamentos dos dados. Visto que o único dado numérico de entrada é a idade, foi realizada a normalização deste. O passo seguinte foi a transformação dos dados genotípicos de cada paciente em unidades unárias. Isso foi feito estabelecendo que cada atributo foi replicado em tantas quantas fossem as representações possíveis deste, e cada atributo deste novo grupo de atributos, seria representado por 1 ou 0. Por exemplo: o *dataset* 1 (tabela 4) poderia ter 4 possíveis representação genotípicas (XX, XY, YY, NR). Então, para cada *locus*, foram criados outros 4 atributos: se esse *locus* for XX, os atributos seriam 1 0 0 0; se fossem XY seriam 0 1 0 0; se YY 0 0 1 0; e se NR 0 0 0 1. Isto foi utilizado também para

representar o sexo e o tipo de infecção do paciente, utilizando apenas dois parâmetros para sua representação. Este artifício foi aplicado para evitar um efeito de ordem na rede neural.

4.7 Resultados das Execuções da MLP Aplicada aos Datasets de Pacientes

Definidos os parâmetros iniciais e os *datasets,* o passo seguinte foi a execução da rede neural segundo a ordem da tabela 4. Todos os testes foram executados em um MacBook com processador Intel Core 2 Due de 2.16 GHz, com 4 GB de memória RAM, utilizando o sistema operacional Mac OS X 10.6.2. Para cada *dataset* utilizado nos testes, alguns parâmetros foram ajustados na busca por melhorias nos resultados de classificação. O primeiro resultado levado em consideração, foi a percentagem de instâncias classificadas corretamente. Entretanto, um aspecto importante a ser considerado, é a sensibilidade do método. Sensibilidade e especificidade são cálculos estatísticos da medida de precisão do teste. No trabalho em questão, a sensibilidade mede percentualmente o quanto o algorítmo consegue classificar corretamente os pacientes com a forma mais severa da doença, e a especificidade mede percentualmente o quanto o algorítmo consegue excluir da classe de pacientes com a forma severa os pacientes acometidos pela a forma branda [48], e os erros são chamados de falso positivo e falso negativo respectivamente. Esse aspecto é imprescindível, uma vez que clinicamente, é muito mais perigoso que um paciente acometido pela forma grave da dengue seja dispensado de cuidados mais detalhados, se comparado ao caso de um paciente com a doença em sua forma branda fique aos cuidados médicos desnecessariamente. Contudo, a busca é por um algoritmo de classificação otimizado para ambos os testes.

Todos os *datasets* (tab. 4) foram utilizados para treinamento e testes do algoritmo, e o *dataset* 11, que utiliza como entrada a idade, sexo, tipo de infecção, os dados genotípicos agrupados conforme a Hip1, e ranqueados pelo *Odds-Ratio* ≥ 2,4, foi o que teve melhores resultados, tanto o total de pacientes classificados

corretamente, como para sensibilidade. Isto feito avaliando os testes com as modificações dos parâmetros de entrada da MLP. Então outros testes foram executados. Para verificar se com esse mesmo conjunto de *loci,* os dados genotípicos e clínicos, agrupados ou não, poder-se-ia obter também bons resultados.

A partir deste momento duas novas avaliações foram realizadas com o *dataset* 11: a primeira delas os dados foram testados desconsiderando os agrupamentos de genótipos e formas clínicas. Na segunda avaliação, apenas a forma de agrupamento das formas clínicas DC/DG, sendo esta a única avaliação estatisticamente relevante (tab. 2). Uma explicação para isso é que a forma clínica DCC tem seu aspecto clínico muito semelhante aos casos de FHD, ficando as diferenças genéticas mais evidenciadas nesta diferenciação de extremos clínicos.

Então a arquitetura da Rede Neural ficou como mostrado na figura 10. Os 14 *loci* (tab. 4), a idade, o sexo e o tipo de infecção foram representados por 62 neurônios de entrada mais o *threshold;* 31 neurônios na camada escondida e dois neurônios como classes.

Tabela 5. Loci utilizados no *dataset* 11, o numero o respectivo *locus* (Polimorfismo), gene, *Odds-Ratio,* o seu formatado, o teste de Fisher, e o valor do teste do Qui-Quadrado para o Eq-HW

no	Polimorfismo	gene	*OddsRatio*	*OddsRatio* Formatado	fisher	Valor p_EqHW
6	OAS3_029396	OAS3	0,3979	2,5129	0,030552011	0,01
29	rs10824793	MBL2	0,3182	3,1429	0,008897528	0,05
41	rs1131715	FADD	0,3333	3,0000	0,010472951	0,025
64	rs12712526	EIF2AK2	0,3657	2,7343	0,020801106	0,05
104	rs1800450	MBL2	2,5226	2,5226	0,035345803	0,01
129	rs2069727	IFNG	8,5263	8,5263	0,000123908	0,01
146	rs2234978	FAS	2,5734	2,5734	0,029806827	0,01
150	rs2240188	OAS3	0,4135	2,4182	0,033590755	0,01
207	rs310199	JAK1	0,3546	2,8199	0,018825692	0,01
213	rs3135932	IL10RA	2,6667	2,6667	0,025979146	0,01
214	rs3136705	CD2	2,5641	2,5641	0,026425493	0,01
216	rs3213545	OASL	0,2658	3,7625	0,003738891	0,01
261	rs4251580	IRAK4	3,1958	3,1958	0,011262437	0,01
295	rs486907	RNASEL	2,8750	2,8750	0,019273295	0,01

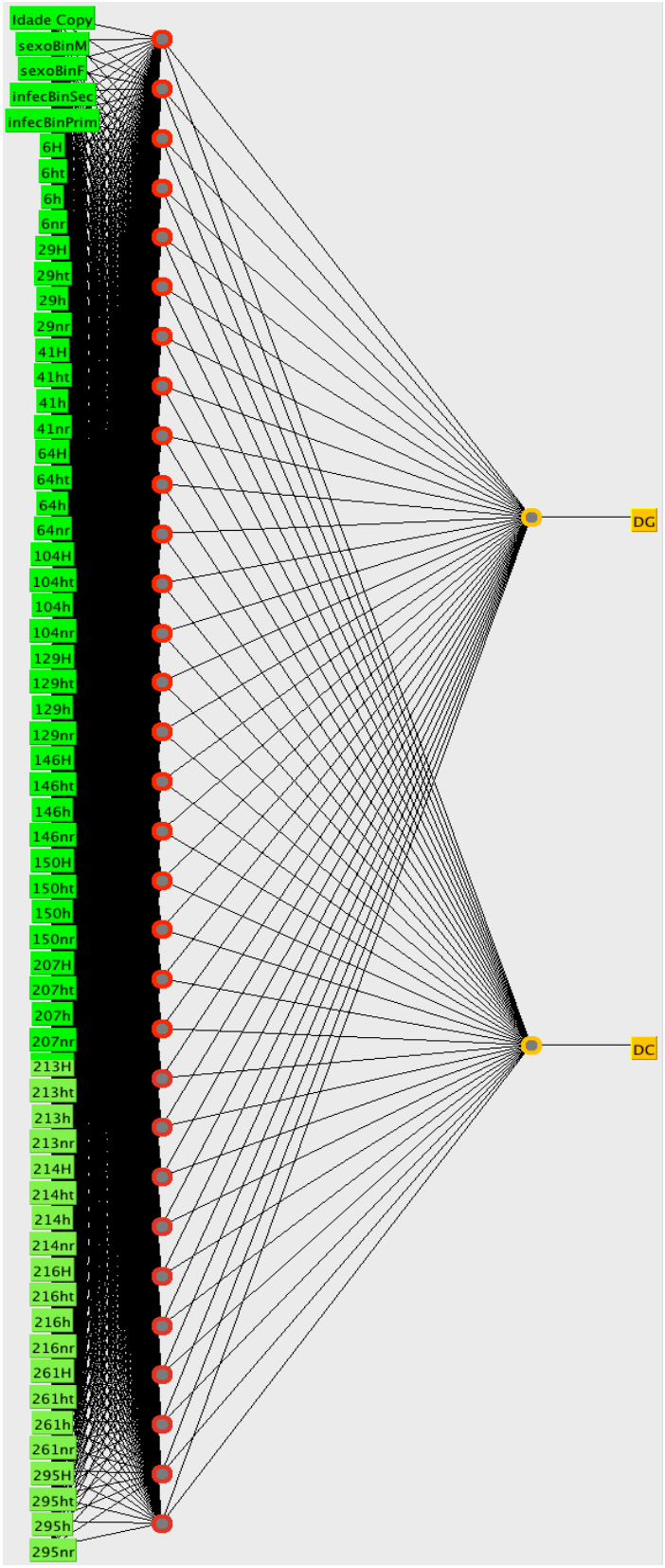


Figura 10. Arquitetura da Rede Neural final utilizada para classificação de formas clínica da dengue mostrando os neurônios de entrada e saída.

4.7.1 Resultados das simulações

Com a arquitetura da rede neural definida, os parâmetros da RNA finais utilizados foram: 600 épocas máximas para treinamento, e 400 épocas para os 25% dos dados utilizados para validação; 0,3 para taxa de aprendizagem, com 0,2 como momento. Dois terços dos dados (69 instâncias) foram utilizados para o treinamento do algorítmo, e o outro um terço (36 instâncias) para os testes.

Ao carregar o arquivo .ARFF na pasta de bases de dados do algorítmo se excuta a classe MLP_Treinar, e o MLP é treinado, após isso a classe MLP_Executar utilizada o modelo de classificação criado durante o treinamento para o teste da parte separada dos dados, retornando na tela as estatísticas do treinamento. Todos os dados foram randomizados antes da execução da MLP. Para estabelecer um resultado confiável, o algoritmo foi executado 50 vezes com dados de treinamento e teste diferentes, modificando a semente geradora da randomização, como mostrado na tabela 5**.** O resultado foi uma media de 75,83% de pacientes classificados corretamente, com uma sensibilidade media de 84,47% com desvio padrão de 8%, e uma especificidade media de 57,93% com desvio padrão de 16%. Esse resultado deixa claro que o algoritmo consegue com mais precisão predizer os pacientes acometidos pelas forma grave da dengue, do que classificar os pacientes que tiveram apenas a forma mais branda. Outras execuções foram feitas modificando a quantidade de dados separados para teste, mas esta foi a que apresentou melhores resultados.

Este é o primeiro relato de utilização de RNA's para classificação de formas clínica da dengue. Os resultados apresentados, embora preliminares, despontam como alternativa promissora para classificação das formas graves de dengue, dados os valores de sensibilidade de 85% e especificidade de 58%. Abordagens anteriores que utilizaram coortes bem estabelecidas obtiveram resultados próximos para dengue. Lee [27] utilizando o métodos de *likelihood-ratio* obtiveram 97% e 60% de sensibilidade e especificidade respectivamente. Em outro trabalho do mesmo grupo [28] através de Árvores de Decisão foram obtidos 100% e 46% dos mesmo valores. Nós acreditamos que com a utilização de PSO [49] para melhoramento da RNA e aumento de numero de casos de dengue estudados (coorte) o método proposto aqui possa apresentar rendimentos semelhantes aos resultados já publicados. Este

modelo pode ajudar a esclarecer as interações entre polimorfismos/genes como fatores biológicos relevantes na evolução de manifestações graves da dengue na nossa população.

Tabela 6. Resultados da execução do algoritmo utilizando o *dataset* 11, mostrando para cada semente de randomização a porcentagem de instancias classificadas corretamente, a sensibilidade e especificidade da classificação, com suas médias e desvio padrão

Semente	% Instancias Corretas	Sensibilidade	Especificidade
1	94,44	1	0,818181818
2	75,00	0,96	0,272727273
3	77,78	0,826086957	0,692307692
4	80,56	0,793103448	0,857142857
5	77,78	0,827586207	0,571428571
6	72,22	0,75862069	0,571428571
7	66,67	0,666666667	0,666666667
8	75,00	0,851851852	0,444444444
9	66,67	0,833333333	0,333333333
10	80,56	0,875	0,666666667
11	72,22	0,777777778	0,555555556
12	80,56	0,84	0,727272727
13	72,22	0,954545455	0,357142857
14	69,44	0,739130435	0,615384615
15	72,22	0,821428571	0,375
16	66,67	0,76	0,454545455
17	66,67	0,642857143	0,75
18	75,00	0,916666667	0,416666667
19	75,00	0,851851852	0,444444444
20	72,22	0,76	0,636363636
21	77,78	0,807692308	0,7
22	75,00	0,782608696	0,692307692
23	83,33	0,958333333	0,583333333
24	75,00	0,782608696	0,692307692
25	80,56	0,88	0,636363636
26	75,00	0,769230769	0,7
27	83,33	0,821428571	0,875
28	69,44	0,863636364	0,428571429
29	72,22	0,72	0,727272727
30	75,00	1	0,357142857
31	72,22	0,777777778	0,555555556
32	75,00	0,913043478	0,461538462
33	80,56	0,88	0,636363636
34	77,78	0,92	0,454545455
35	80,56	0,92	0,545454545
36	75,00	0,833333333	0,583333333
37	86,11	0,851851852	0,888888889
38	77,78	0,909090909	0,571428571
39	80,56	0,925925926	0,444444444

Tabela 6. Continuação

Semente	% Instancias Corretas	Sensibilidade	Especificidade
40	61,11	0,76	0,272727273
41	77,78	0,769230769	0,8
42	77,78	0,814814815	0,666666667
43	77,78	0,916666667	0,5
44	77,78	0,913043478	0,538461538
45	72,22	0,791666667	0,583333333
46	83,33	0,875	0,75
47	86,11	0,862068966	0,857142857
48	83,33	0,956521739	0,615384615
49	72,22	0,807692308	0,5
50	61,11	1	0,125
Média	75,83	0,844795489	0,579385448
D. Padrão	6,31	0,082966559	0,16957461

Capítulo 5

Conclusão e Trabalhos Futuros

Neste trabalho, foi desenvolvida uma ferramenta para auxiliar na classificação de formas clínicas de pacientes infectados pelo vírus da dengue. Esta ferramenta incorporou Redes Neurais Artificiais como mecanismo de inferência, e utilizou de forma original dados provenientes de polimorfismos genéticos.

Desenvolveu-se também, um banco de dados afim de armazenar pré-processamentos (dados provenientes da genotipagem), e uma ferramenta própria para a manipulação deste banco de dados. Estas ferramentas também são inovadoras, por retornar informações estatísticas e biológicas relevantes de um grande conjunto de *loci* (322 neste caso). Entretanto, por ter sido desenvolvido em uma plataforma propietária, o *Filemaker Pro,* este sistema de gerenciamento de banco de dados (SGBD) ainda que autorizado para distribuição, poderia ser desenvolvido em outra linguagem de programação, e assim poderia até mesmo ser comercializado [23].

A aplicação do modelo proposto é eminentemente de pesquisa científica. Entretanto, por se tratar de classificação clínica de pacientes possui interesse potencial elevado para aplicações reais (em unidades de saúde pública). Uma vez que a coorte estudada é pequena, uma expansão deste estudo poderia trazer resultados de maior respaldo estatístico, para posteriormente ser passível de aplicação de auxilio no *front* médico - paciente nas epidemias de dengue.

5.1 Contribuição

Foi criado um SGBD para armazenamento de dados genéticos, que auxilia o estudo preliminar de *loci* que podem ser analisados para estudos univariados ou multivariados de classificadores. Foi implementado um modelo de classificação de formas clínicas da dengue multivariado (multigênico), que utiliza o algorítmo de Redes Neurais Artificiais. Este algorítmo busca minimizar os erros de classificação.

Este trabalho pode ser um ponto de partida para se desenvolver uma nova linha de pesquisa no DSC (Departamento de Sistemas e Computação), ou seja em problemas com aplicações de tecnologia da informação na área médica de genômica associada a epidemias.

5.2 Trabalhos Relacionados

Paralelo a este trabalho, foi desenvolvido um trabalho que utiliza a mesma base de dados, que foi extraída do SGBD desenvolvido durante este projeto, a fim de buscar assinaturas genéticas multivariadas. Nesse estudo, técnicas estatísticas como regressão logística e o critério de informação Akaike são os motores de inferência. Apesar desse trabalho ainda estar em fase de validação dos resultados, será uma possível boa base de comparação de resultados.

5.3 Dificuldades Encontradas

Por ser uma área que requer conhecimentos específicos de biologia, como biologia celular e molecular, microbiologia, genética, virologia, epidemiologia, dentre outras, o entendimento da base teórica foi lento e representou uma dificuldade para este trabalho monográfico. Vários artigos de todas essas áreas foram lidos, não obstante a pesquisa ficou centrada em trabalhos que tivessem sido elaborados utilizando as mesmas técnicas para os mesmo fins, o que não foi encontrado. Por ser então um trabalho multidisciplinar e até certo ponto inovador, foram encontradas todas as dificuldades que um trabalho fora da escopo da engenharia da computação pode encontrar. A exigüidade de tempo foi outro fator que dificultou o desenvolvimento deste trabalho, uma vez que ele foi desenvolvido em paralelo a outras quatro disciplinas, o estágio, e outras atividades extra-pesquisa.

5.4 Trabalhos Futuros

Por ter sido um trabalho desenvolvido em menos de um semestre, muito poderia ser elaborado e desenvolvido para o melhoramento dos resultados desta pesquisa. Mesmo se utilizando de uma seleção de grupos de atributos a partir de definições

estatísticas, poderia ser feita uma seleção de atributos mais automática, para encontrar um conjunto de *loci* mais eficiente. Outro trabalho que pode ser desenvolvido, é a utilização de *Particle Swarm Optimization* (PSO) [49] para estabelecer os melhores parâmetros da MLP, ou ainda, utilizar em substituição da RNA, o próprio algoritmo de PSO.

Uma proposta de melhoramento deste projeto, seria o aumento da Coorte de pacientes, para fazer uma base de treinamento dos algoritmos maior, e ter mais dados para teste, trazendo assim um modelo validado e mais confiável.

Referências Bibliográficas

1. Watson, J.D. and F.H. Crick, *The structure of DNA.* Cold Spring Harb Symp Quant Biol, 1953. **18**: p. 123-31.
2. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-8.
3. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
4. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
5. Ouzounis, C.A. and A. Valencia, *Early bioinformatics: the birth of a discipline – a personal view.* Bioinformatics, 2003. **19**(17): p. 14.
6. Guimarães, A.C.R., *Introdução a Bioinformática.* Laboratório de Genômica Funcional e Bioinformática - Instituto Oswaldo Cruz, 2008.
7. Siqueira, J.B., Jr., et al., *Dengue and dengue hemorrhagic fever, Brazil, 1981-2002.* Emerg Infect Dis, 2005. **11**(1): p. 48-53.
8. Gubler, D.J., *Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century.* Trends Microbiol, 2002. **10**(2): p. 100-3.
9. Cordeiro, M.T., et al., *Dengue and dengue hemorrhagic fever in the State of Pernambuco, 1995-2006.* Rev Soc Bras Med Trop, 2007. **40**(6): p. 605-11.
10. Nascimento, E.J., et al., *Gene Expression Profiling during Early Acute Febrile Stage of Dengue Infection Can Predict the Disease Outcome.* PLoS One, 2009. **4**(11): p. e7892.
11. Cordeiro, M.T., et al., *Characterization of a dengue patient cohort in Recife, Brazil.* Am J Trop Med Hyg, 2007. **77**(6): p. 1128-34.
12. Silva-Júnior, J.B.S.J. and e. al., *Dengue no Brasil - Histórico, Situação Atual e Perspectivas.* Revista Ciência e Ambiente, 2002. **25**: p. 436-444.
13. Souza, J.L.D., *Dengue - diagnóstico, tratamento e prevenção*. 1 ed. 2007, Rio de Janiero: Rubio.
14. Batista, R.S., *Medicina Tropical: Abordagem atual das doenças infecciosas e parasitárias*. Vol. 2. 2001, Rio de Janeiro: Cultura Médica.
15. FIOCRUZ. *Exposição Dengue*. 2008 02/09/2008]; Available from: http://www.invivo.fiocruz.br/dengue/virus_pt.htm.
16. Teixeira, M.G., M.L. Barreto, and Z. Guerra, *Epidemiologia e medidas de prevenção do dengue.*, MS, Editor. 1999, Informe Epidemiológico do SUS. p. 5-33.
17. Fersh, J.W., *Philippine hemorrhagic fever: a clinical, laboratory and necropsy study.* J. Lab. Clín. Méd., 1969. **73**: p. 451-458.
18. Travassos da Rosa, A.P.A., et al., *Surto de dengue em Boa Vista, Território de Roraima.* Boletim Epidemiológico, 1982. **14**: p. 93-100.
19. Nogueira, R.M. and e. al, *Dengue in the State of Rio de Janeiro, Brazil, 1986-1998.* Mem Inst Oswaldo Cruz, 1999. **94**: p. 297-304.
20. Nogueira, R.M. and e. al., *Dengue virus type 3 in Rio de Janeiro, Brazil.* Mem Inst Oswaldo Cruz, 2001. **96**: p. 925-6.
21. Brasil, M.S., *Dengue: diagnóstico e manejo clínico*, M.d. Saúde, Editor. 2005, Editora do Ministério da Saúde: Brasília.
22. Gubler D, K.G., *Dengue and dengue hemorrhagic fever*. CAB International ed. 1997, London.

23. Cordeiro, M.T., et al., *Reliable classifier to differentiate primary and secondary acute dengue infection based on IgG ELISA.* PLoS One, 2009. **4**(4): p. e4945.
24. Acioli-Santos, B., et al., *MBL2 Gene polymorphisms protect against development of thrombocytopenia associated with severe dengue phenotype.* Hum Immunol, 2008. **69**(2): p. 122-8.
25. Nascimento, E.J., et al., *Alternative complement pathway deregulation is correlated with dengue severity.* PLoS One, 2009. **4**(8): p. e6782.
26. Fernandez-Mestre, M.T., et al., *TNF-alpha-308A allele, a possible severity risk factor of hemorrhagic manifestation in dengue fever patients.* Tissue Antigens, 2004. **64**(4): p. 469-72.
27. Lee, V.J., et al., *Predictive value of simple clinical and laboratory variables for dengue hemorrhagic fever in adults.* J Clin Virol, 2008. **42**(1): p. 34-9.
28. Lee, V.J., et al., *Decision tree algorithm in deciding hospitalization for adult patients with dengue haemorrhagic fever in Singapore.* Trop Med Int Health, 2009. **14**(9): p. 1154-9.
29. Sakuntabhai, A., et al., *A variant in the CD209 promoter is associated with severity of dengue disease.* Nat Genet, 2005. **37**(5): p. 507-13.
30. Calzavara-Silva, C.E., et al., *Early molecular markers predictive of dengue hemorrhagic fever.* An Acad Bras Cienc, 2009. **81**(4): p. 671-7.
31. Gomes, A., et al., *Classification of dengue fever patients based on gene expression data using support vector machines*. 2009.
32. Turing, A.M., *Computing machinery and intelligence.* New Series, 1950. **59**(236): p. 433-460.
33. Buarque, F., et al., *InteliMED*. 2009, Finep, CNPq, UPE, DSC, PRIMISE, FCM, UFPE, Nutes, GTIS: Recife.
34. Braga, A.P., A.C.P.L.F. Carvalho, and T.B. Ludemir, *Redes Neurais Artificiais*. 2003. p. 30.
35. Ambrósio, P.E., *Redes neurais artificiais no apoio ao diagnóstico diferencial de lesões intersticiais pulmonares.* 2002, FAculdade de Filosofia, Ciências e Letras de Ribeirão Preto: Ribeirão Preto.
36. McCulloch, W.S. and W. Pitts, *A logical calculus of the ideas immanent in nervous activity.* Bulletin of Mathematical Biophysics, 1943. **5**: p. 115-133.
37. Loesch, C. and S.T. Sari, *Redes Neurais Artificiais: Fundamentos e MOdelos*. 1996, Blumenau: FURB.
38. Minsky, M. and S. Papert, *Perceptrons; an introduction to computational geometry*. 1969: MIT Press.
39. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, *Learning representations by back-propagating errors.* Nature, 1986. **323**: p. 533-536.
40. Milaré, C.R., *Extração de conhecimento de Redes Neurais Artificiais utilizando sistemas de aprendizado simbólico e algoritmos genéticos.* 2003, Universidade de São Paulo: São Carlos.
41. Rosenblatt, F., *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.* Psychological Review, 1958. **65**(6): p. 386-408.
42. Mosteller, F., *Association and Estimation in Contingency Tables.* Journal of the American Statistical Association, 1968. **63**(321): p. 1-28.
43. Edwards, A.W.F., *The measure of association in a 2x2 table.* Journal of the Royal Statistical Society, 1963. **126**(1): p. 109-114.
44. Castle, W.E., *The laws of Galton and Mendel and some laws governing race improvement by selection.* Proc. Amer. Acad. Arts Sci., 1903. **35**: p. 233-242.

45. Hall, M., et al., *The WEKA Data Mining Software: An Update.* SIGKDD Explorations, 2009. **11**.
46. Technology, J. *Java Technology*. 2009 [cited 2009 17/08/2009]; Available from: http://java.sun.com.
47. Valença, M., *Fundamentos das Redes Neurais*. Vol. 1. 2007, Recife: Livro Rápido. 381.
48. Aragon, D.C., *Avaliação de Métodos Estatísticos Aplicados ao Estudo de Testes Diagnósticos na Presença do Viés de Verificação*. 2007, Universidade de São Paulo: São Paulo.
49. Kennedy, J.E., R., *Particle swarm optimization*, in *Neural Networks IEEE International Conference*. 1995: Australia. p. 1942-1948.