Pós-Graduação em Ciência da Computação

Flávia Roberta Barbosa de Araújo

INFERÊNCIA DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO UTILIZANDO ALGORITMOS BASEADOS EM *RELEVANCE LEARNING VECTOR QUANTIZATION*

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE
2017

Flávia Roberta Barbosa de Araújo

INFERÊNCIA DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO UTILIZANDO ALGORITMOS BASEADOS EM *RELEVANCE LEARNING VECTOR QUANTIZATION*

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Orientadora: *Katia Silva Guimarães*

RECIFE
2017

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

A662i Araújo, Flávia Roberta Barbosa de
Inferência de polimorfismos de nucleotídeo único utilizando algoritmos baseados em *Relevance Learning Vector Quantization* / Flávia Roberta Barbosa de Araújo. – 2016.
126 f.: il., fig., tab.

Orientadora: Katia Silva Guimarães.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2016.
Inclui referências e apêndice.

1. Ciência da computação. 2. Interação epistática. I. Guimarães, Katia Silva (orientadora). II. Título.

004 CDD (23. ed.) UFPE- MEI 2017-94

Flávia Roberta Barbosa de Araújo

Inferência de Polimorfismos de Nucleotídeo Único Utilizando Algoritmos Baseados em Relevance Learning Vector Quantization

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutora em Ciência da Computação

Aprovado em: 21/02/2016.

_____

Orientador: Profa. Dra. Katia Silva Guimarães

BANCA EXAMINADORA

_____

Profa. Dra. Renata Maria Cardoso Rodrigues de Souza
Centro de Informática / UFPE

_____

Prof. Dr. Renato Vimieiro
Centro de Informática / UFPE

_____

Prof. Dr. Paulo Salgado Gomes de Mattos Neto
Centro de Informática / UFPE

_____

Profa. Dra. Maria Emília Telles Walter
Instituto de Ciências Exatas / UnB

_____

Prof. Dr. Jones Oliveira de Albuquerque
Departamento de Estatística e Informática / UFRPE

Só imaginação não é o bastante, pois a realidade da natureza é muito mais surpreendente do que qualquer coisa imaginável. Esta aventura só é possível porque gerações de pesquisadores seguiram rigorosamente um conjunto de regras.

-Teste ideias através de experimentos e observações. Desenvolva as ideias que passarem no teste, rejeite as que falharam. Siga as evidências aonde quer que vão e questione tudo.

—NEIL DEGRASSE TYSON IN COSMOS

Resumo

Embora duas pessoas compartilhem mais de 99% do DNA, as variações são extremamente relevantes para determinar as variações fenotípicas. Dentre essas variações, os polimorfismos de nucleotídeo único (SNP) são alterações pontuais mais conhecidas por influenciar no aumento no risco de doenças. Os SNPs podem atuar individualmente ou através de interações com outros SNPs (interações epistáticas). A inferência das interações epistáticas é um problema que vem sendo amplamente estudado, sendo utilizados dados genômicos de estudos de associação ampla do genoma (GWAS) com pacientes casos e controles. Diversas abordagens computacionais foram propostas, utilizando diferentes estratégias para lidar com os desafios de inferir as interações mais relevantes. O primeiro desafio encontrado neste estudo, esta relacionado à grande quantidade de dados (cerca de 500 a 900 mil SNPs). O segundo desafio esta associado ao número de possíveis interações entre SNPs, o que leva a um problema combinatorial. E o terceiro desafio, relaciona-se com o baixo poder estatístico das interações, sendo mais custoso identificá-las. A combinação desses desafios, tornam este um problema difícil de ser tratado. Nesta tese, são utilizadas diferentes metodologias, selecionadas para verificar suas capacidades em lidar com o problema da inferência da interações epistáticas. Dentre estas, são avaliadas técnicas de seleção de características e abordagens computacionais na detecção das interações entre SNPs, assim como algoritmos de aprendizagem de máquina baseados em *Relevance Learning Vector Quantization* (RLVQ). Nos experimentos realizados, os algoritmos baseados em RLVQ apresentaram resultados satisfatórios ao identificar as interações relevantes entre SNPs em dados com até 5 interações, utilizando requisitos computacionais relativamente baixos quando comparados a outras abordagens descritas na literatura. Um estudo mais extenso foi realizado, com o objetivo de identificar um ajuste ideal dos parâmetros e verificar as capacidades e limitações de cada algoritmo. Com os resultados obtidos através desse ajuste de parâmetros, foi possível levantar hipóteses referente a influência da quantidade de interações entre SNPs e da dimensionalidade dos dados em função dos parâmetros utilizados nos algoritmos. Considerando essas análises, foi possível propor uma nova metodologia denominada iGRLVQ-SNPi, baseada em algoritmos de RLVQ, para lidar de forma mais eficiente com o problema da inferência das interações entre os SNPs. Com o iGRLVQ-SNPi, foi possível avaliar interações de ordem n, sem que para isso, fosse necessário informar o número de interações que se deseja avaliar. Nos experimentos realizados, o iGRLVQ-SNPi obteve uma excelente acurácia nos diferentes conjuntos de dados testados, e sendo comparativamente melhor ou tão eficiente quanto outras abordagens de inferência epistáticas avaliadas, utilizando um menor custo computacional.

Palavras-chave: GWAS. SNPs. Interação Epistática. Seleção de Características. Aprendizagem de Quantização Vetorial.

Abstract

Although two people share more than 99% of DNA, variations are extremely relevant for determining phenotypic variations. Among these variations, single nucleotide polymorphisms (SNPs) are punctual changes known to influence the increased risk of disease. SNPs can act individually or through interactions with other SNPs (epistatic interactions). The inference of epistatic interactions is a problem that has been extensively studied, using genomic data from genome wide association studies (GWAS) with cases and controls patients. Several computational approaches were proposed, using different strategies to deal with the challenges of inferring the most relevant interactions. The first challenge found in this study is related to the large amount of data (about 500 to 900 thousand SNPs). The second challenge is the number of possible interactions between SNPs, which leads to a combinatorial problem. And the third challenge is related to the low statistical power of the interactions, being more difficult to identify them. The combination of these challenges makes this a hard problem to address. In this thesis, different methodologies were used, they were selected to verify their abilities in dealing with the problem of inference of the epistatic interactions. Among these, we evaluate techniques of feature selection and computational approaches in the detection of interactions between SNPs, as well as machine learning algorithms based on Relevance Learning Vector Quantization (RLVQ). In the experiments performed, the RLVQ-based algorithms presented satisfactory results by identifying the relevant interactions between SNPs in data with up to 5 interactions, using relatively low computational requirements when compared to other approaches described in the literature. A more extensive study was carried out with the objective of identifying an optimal adjustment of the parameters and verifying the capacities and limitations of each algorithm. With the results obtained through this adjustment of parameters, it was possible to raise hypotheses regarding the influence of the amount of interactions between SNPs and the dimensionality of the data as a function of the parameters used in the algorithms. Considering these analyzes, it was possible to propose a new methodology called iGRLVQ-SNPi, based on RLVQ algorithms, to deal more efficiently with the problem of inference of the interactions between the SNPs. With iGRLVQ-SNPi, it was possible to evaluate n-order interactions, without it being necessary to inform the number of interactions to be evaluated. In the experiments performed, iGRLVQ-SNPi obtained an excellent accuracy in the different datasets tested, and was comparatively better or as efficient as other evaluated epistatic inference approaches, using a lower computational cost.

Keywords: GWAS. SNPs. Epistatic Interaction. Feature Selection. Learning Vector Quantization.

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

AD	Doença ou mal de Alzheimer.
ADNI	Base de dados *Alzheimer's Disease Neuroimaging Initiative*.
APOE	Apoliproteína E, associada como fator de risco da doença de Alzheimer.
APP	Gene da proteína precursora da amilóide, do inglês: *Amyloid Precursor Protein*.
AR	Métrica *Average Rank*.
BCR-ABL	Gene que produz uma proteína híbrida tirosina quinase.
BEAM	Ferramenta *Bayesian Epistasis Association Mapping*.
BOOST	Ferramenta *Boolean Operation based Screening and Testing*.
CINOEDV	*Co-Information Based Method for Detecting and Visualizing n-order Epistatic Interactions*.
DNA	Ácido desoxirribonucléico. Molécula formada pela união de nucleotídeos.
FCBF	Filtro de Seleção de Característica *Fast Correlation-based Feature Selection*.
GLVQ	Algoritmo *Generalized Learning Vector Quantization*.
GNG	Algoritmo *Growing Neural Gas*
GPU	Unidade de Processamento Gráfico, do inglês: *Graphics Processing Unit*.
GRLVQ	Algoritmo *Generalized Relevance Learning Vector Quantization*.
GWAS	Estudos de associação ampla do genoma, do inglês: *Genome-wide association study*.
iGRLVQ	Algoritmo *Incremental Generalized Relevance Learning Vector Quantization*.
LD	Desequilíbrio de Ligação, do inglês: *Linkage disequilibrium*.
LHS	*Latin Hypercube Sampling*.
LINEs	*Long Interspersed Nuclear Elements*.
LMCI	Comprometimento Cognitivo Leve e Tardio, do inglês: *Late Mild Cognitive Impaiment.*
LVQ	Algoritmo *Learning Vector Quantization*.
MCI	Comprometimento Cognitivo Leve, do inglês: *Mild Cognitive Impaiment.*
MAF	Frequência do menor alelo, do inglês: *Minor allele frequence*.
MDR	Ferramenta *Multifactor Dimensionality Reduction*.
NG	Algoritmo *Neural Gas*.

NGS	Sequênciamento de Nova Geração, do inglês: *Next-Generation Sequencing*.
RLVQ	Algoritmo *Relevance Learning Vector Quantization*.
RNA	Ácido ribonucléico. Molécula formada pela união de nucleotídeos.
SINEs	*Short Interpersed Nuclear Elements*.
SNP	Polimorfismo de nucleotídeo único, do inglês: *Single Nucleotide Polymorphism*.
SNPns	Um tipo de SNP, SNP não-sinônimo.
SOM	Algoritmo *Self-Organizing Maps*.
SGNG	Algoritmo *Supervised Growing Neural Gas*.
SRNG	Algoritmo *Supervised Relevance Neural Gas*.
STRs	Microssatélites, do inglês: *Short Tandem Repeats*.
VNTRs	Minissatélites, do inglês: *Variable Number of Tandem Repeats*.

Sumário

1

Introdução

Na década de 90, ocorreu a união de vários países em um consórcio internacional, o Projeto Genoma Humano, para desenvolver um projeto que tinha como objetivo sequenciar o código genético humano. Devido ao grande número de instituições de pesquisa investindo esforços para a conclusão deste projeto, ocorreu um amplo desenvolvimento de técnicas de sequenciamento. Assim, em 2003, o projeto foi concluído antecipadamente, com o sequenciamento de 99% do genoma humano (Consortium, 2004).

Consequentemente, surgiu uma nova área de conhecimento, a **Bioinformática**, que se tornou uma área de destaque que envolve profissionais de diferentes áreas de conhecimento: Medicina, Ciência da Computação, Biologia, Química entre outros. Essa fusão de esforços, buscou armazenar, analisar e desvendar os mistérios por trás das informações codificadas no alfabeto genético representado pelas letras A, T, C, G e U, que são as bases nitrogenadas que compõem o material do genoma humano.

No início do século XXI, formas cada vez mais sofisticadas de sequenciamento do DNA levaram à produção de grandes volumes de dados genômicos. Com este advento, abriram-se portas para uma gama de novas pesquisas fazendo uso da grande massa de informação biológicas que se tornou disponível.

Através do estudo desses dados genômicos, tem sido possível compreender como diferentes regiões do genoma, tais como os genes, influenciam na incidência de uma determinada doença ou característica. No entanto, para que isso seja possível, neste tese lançamos mão de estudar e propor uma metodologia capaz de identificar os principais fatores genéticos e ambientais associados a uma determinada doença. Com esse conhecimento em mãos, abre-se um leque de possibilidades que propiciam o desenvolvimento de diagnósticos mais específicos e rápidos, assim como a produção de fármacos ou tratamentos personalizados para diferentes perfis genéticos.

1.1 Caracterização do Problema

Embora duas pessoas compartilhem semelhanças em mais do que 99% do seu DNA, existem inúmeras diferenças entre os genomas de dois indivíduos (Feuk *et al.*, 2006) que com o avanço tecnológico do sequenciamento do DNA, foi possível detectar diferenças em até uma única base nitrogenada, revelando assim a presença dos *Single Nucleotide Polymorphism* (SNP).

Os **SNPs** são modificações pontuais no genoma que geram diferentes formas de um mesmo gene. A ocorrência dessa modificação tem uma frequência mínima de 1% na população, caso contrário, a variação é considerada uma simples mutação (Schwender e Ickstadt, 2008). Essas variações pontuais podem ocorrer em diferentes regiões nas sequências de DNA, mas ao ocorrer especialmente em regiões bem conservadas, podem provocar mudanças fenotípicas mais significativas, tais como, anormalidades no metabolismo ao alterar importantes atuadores das vias metabólicas celulares (Schwender e Ickstadt, 2008).

Estudos envolvendo a análise de genoma de pacientes casos e controles (***Genome-wide Association Study*** - GWAS), utilizam dados com centenas de milhares de SNPs. O enorme volume dos dados influencia na forma como as diferentes aborsagens computacionais lidam com os dados. Abordagens computacionais mais tradicionais se detém a avaliar a influência de um único SNP sobre o fenótipo. Esse tipo de abordagem é útil e computacionalmente eficiente para identificar casos relacionados com doenças monogenéticas (Zhao *et al.*, 2011), tais como Anemia Falciforme, Doença de Huntington, Fibrose Cística, entre outras.

No entanto, casos em que um único SNP atua de forma individual sobre uma doença são raros, sendo as interações entre os SNPs mais fortemente correlacionadas com uma grande gama de doenças complexas, como doenças cardiovasculares, neurológicas, endócrinas, entre outras (NLM, 2017). No entanto, identificar as diferentes combinações de interações entre os fatores de risco de uma doença complexa é um problema computacionalmente muito custoso, e por vezes inviável, dependendo da forma de busca utilizada, tal como uma busca exaustiva das interações entre SNPs.

Ao entender a relevância da identificação dos SNPs em função de sua correlação com o aumento da incidência de inúmeras doenças, apresentamos aqui os desafios relacionados a sua identificação utilizando os dados biológicos:

- Um primeiro desafio que nos deparamos ao analisar as interações entre SNPs se dá pelo fato de que esses dados de GWAS são de **alta dimensionalidade**, ou seja, contém uma quantidade grande de atributos para cada instância (ou indivíduo) (na ordem de 100 mil). Assim, o número de testes que precisariam ser realizados para detectar as interações leva a um problema conhecido na computação como a "maldição da dimensionalidade" (Keogh e Mueen, 2010).

- Um segundo desafio ainda em aberto é a identificação das **interações de alta ordem entre SNPs**, ou seja, envolvendo mais de dois SNPs. Vários métodos, tais como

BOOST (Wan *et al.*, 2010a), SNPHarvester (Yang *et al.*, 2009), MDR (Ritchie e Motsinger, 2005), já realizam a inferência das interações SNP-SNP, ou seja interações entre um par de SNPs, com eficiência, por ser um problema relativamente mais fácil de se resolver. No entanto, à medida que o número de SNPs interagindo cresce, as abordagens tornam-se mais restritivas quanto ao número de análises realizadas, ignorando a existência de interações de alta ordem, as quais podem conter informações valiosas para caracterizar diferentes tipos de doenças (Wang *et al.*, 2015).

- Um terceiro desafio está associado à grande quantidade de dados, unidos com **baixo poder estatístico das interações**. Essa combinação faz com que várias interações apresentem uma baixa probabilidade de representar uma doença, levando assim a uma alta taxa de falsos positivos (Fang *et al.*, 2012).

1.2 Motivação, Problema e Objetivos

Para lidar com os desafios relatados acima, nesta tese foram considerados dois tipos de abordagens. O primeiro, utilizando **técnicas de seleção de características** com o objetivo de reduzir a dimensionalidade dos dados, removendo uma parte dos SNPs irrelevantes do conjunto de dados e tornando assim a tarefa de identificação das interações menos árdua para uma outra metodologia; e o segundo tipo de abordagem experimentando diferentes métodos computacionais não exaustivos e robustos, capazes de avaliar os conjuntos de dados do tipo GWAS.

Para a escolha das técnicas de seleção de características, foram considerados os filtros de seleção de caracteríistivas mais citados e adequados para lidar com o problema da inferência das interações entre os SNPs. A utilização dessas técnicas é justificada pela tentativa de lidar com o primeiro desafio descrito: a alta dimensionalidade dos dados. Para isso, serão realizados experimentos com as diferentes técnicas utilizando conjuntos de dados com diferentes características, de forma a obter informações sobre suas capacidades e limitações ao lidar com o problema.

Após a avaliação das técnicas de seleção de características, serão estudadas diferentes abordagens computacionais já conhecidas, propostas para lidar com o problema das interações entre SNPs, verificando as diferentes limitações encontradas ao lidar com o problema. Após esta revisão, serão investigadas formas alternativas capazes de lidar com a inferência de SNPs relevantes, para isso serão avaliadas **abordagens computacionais baseadas em aprendizagem de máquina**.

Os algoritmos baseados em *Learning Vector Quantization* (LVQ), originados do LVQ1 proposto por Kohonen (1997), são abordagens tradicionalmente desenvolvidas para resolver problemas de classificação de dados, capazes de lidar com conjuntos de dados com alta dimensionalidade. Dentre os algoritmos pertencentes a essa família, como os algoritmos baseados em *Relevance Learning Vector Quantization* (RLVQ), tais como, o algoritmo *Generalized* RLVQ (GRLVQ) parece ser um candidato promissor, por ser capaz de lidar com o problema de dados

de alta dimensionalidade ao utilizar um vetor de relevâncias que atribui um peso à importância de uma dimensão na classificação dos dados.

Muitas abordagens que realizam a inferência de interações epistáticas são limitadas quanto ao número de interações que avaliam, ou precisam que o número de interações a ser avaliado precise ser informado previamente. No entanto, para os conjuntos de dados reais, não é possível saber com precisão o número de SNPs relevantes que estão presentes no conjuntos de dados. Assim, a utilização de uma metodologia mais adaptável como o uso do vetor de relevância, pode ser uma alternativa para lidar com essas limitações. Com o uso do vetor de relevâncias, não há necessidade de informar o número de interações que se deseja inferir, assim o método fica livre para identificar um número n de SNPs relevantes, sem cair num problema de ordem exponencial como encontrado em abordagens de busca exaustiva.

Apesar disso, os algoritmos baseados em RLVQ são propostos para lidar com o problema de classificação de dados, e não com o problema da inferência de SNPs relevantes. Para avaliar suas limitações, são realizados experimentos para melhor investigá-los, e apartir das análises dos resultados são propostas adaptações e modificações com a proposição de uma nova abordagem baseada no GRLVQ mais adequada ao problema.

Diante disso, nesta tese serão realizados experimentos utilizando dados simulados e reais de pacientes casos e controles, para avaliar o desempenho tanto de técnicas de inteligência computacional como técnicas de seleção de características, frente à inferência das interações entre os SNPs, apontando suas deficiências e sugerindo soluções implementadas de uma nova ferramenta para lidar com a inferência das interações epistáticas.

1.3 Organização da Tese

Neste primeiro capítulo foi apresentada uma introdução do problema de inferência de interações entre SNPs avaliando-o sob um foco computacional. Além disso, foram apresentados a motivação e os objetivos a serem alcançados. A seguir, é apresentada a estruturação do documento desta tese.

No Capítulo 2, são apresentadas informações biológicas essenciais para que o leitor compreenda aspectos biológicos básicos e relevantes para a compreensão do problema, apresentando em detalhes diferentes moléculas, estruturas e processos biológicos relacionados com os SNPs, assim como uma descrição detalhada sobre os SNPs, suas interações e influência sobre o fenótipo. Ainda neste capítulo, como estudo de caso, serão apresentadas informações sobre a doença de Alzheimer, sua prevalência, fatores de risco e tratamento.

No Capítulo 3, são descritos os conjuntos de dados simulados e real utilizados no estudo das interações entre SNPs, apontando o embasamento biológico que deve ser levado em consideração quando se deseja criar dados simulados que representem adequadamente as características relevantes das informações genômicas reais.

No Capítulo 4, são apresentadas ferramentas computacionais que tratam do problema da

interação entre os SNPs com suas especificações e características particulares. Neste capítulo também são apresentadas técnicas de seleção de características e algoritmos de aprendizagem de máquina alvo de experimentações para verificar suas capacidades e limitações em lidar com o problema tratado.

No Capítulo 5, são descritos os experimentos iniciais realizados com as diferentes abordagens descritas no Capítulo 4. A partir dos resultados obtidos e experimentos realizados são levantadas hipóteses como fontes de inspiração para a proposição de uma nova metodologia, apresentada no Capítulo 6. Ainda neste capitulo, são apresentados os experimentos utilizando dados simulados e reais, comparando sua acurácia e performance com diferentes ferramentas computacionais.

Por fim, no Capítulo 7, são apresentadas as considerações finais desta tese.

2

Conceitos Básicos em Biologia Molecular

Neste capítulo, são apresentados conceitos biológicos importantes para a compreensão da inferência das interações entre SNPs. Para se entender os SNPs, sua importância e como ocorrem suas interações, serão apresentadas diferentes moléculas e processos biológicos relacionados. Adicionalmente, será apresentada a doença de Alzheimer, incluindo seus sintomas, prevalência, características genéticas e base de dados.

2.1 Conceitos Básicos

O **Projeto Genoma Humano**, surgiu na década de 90 com a união de vários países em um consórcio internacional. O objetivo desse projeto era identificar os genes presentes no genoma humano.

Assim em 2001, um primeiro esboço do genoma, contendo cerca de 90% de toda a informação genética, foi publicado nas revistas Nature (Consortium, 2001) e Science (Venter *et al.*, 2001), e em 2004 foi anunciada à imprensa a conclusão total do sequenciamento do genoma com 99,9% de precisão (Consortium, 2004).

Doze anos foram necessários para a conclusão do sequenciamento do genoma, com um gasto de cerca de 3 bilhões de dólares. Com este primeiro esboço do genoma, muitos questionamentos levantados sobre o funcionamento do DNA foram revelados, mas muitos outros ainda estão por ser entendidos.

Incentivos para uma rápida finalização desse projeto, proporcionaram o surgimento de eficientes sequenciadores automáticos de DNA. Atualmente, o *Next-Generation Sequencing* (NGS) é capaz de sequenciar todo genoma humano em apenas um dia, com um gasto médio de apenas 1.000 dólares americanos (Muzzey *et al.*, 2015). Esse avanço propiciou o barateamento e rapidez no sequenciamento, permitindo avanços nos estudos do genoma humano.

Nas próximas seções, serão apresentadas as moléculas que compõem o genoma, suas propriedades e funções em uma célula eucariótica. Também serão apresentadas as formas de replicação e expressão gênica do material genético. Ao final, serão apresentados os tipos de variações genéticas existentes no genoma, dando um maior foco para os polimorfismos de

nucleotídeo único, alvo do estudo nesta tese.

2.1.1 DNA e RNA

Nas células dos organismos vivos, existem duas moléculas essenciais para manutenção da vida, o ácido desoxirribonucleico (DNA) que tem como principal função armazenar informações essenciais para manutenção do organismo e hereditariedade, e o ácido ribonucleico (RNA) que tem um papel fundamental na síntese de proteínas e outras moléculas de RNA (Miko e LeJeune, 2009).

Essas duas moléculas, representadas na Figura 2.1, são compostas por componentes básicos e fundamentais denominadas de **nucleotídeos**, unidos por ligações covalentes fosfodiéster, formando uma cadeia de nucleotídeos (Nature-Education, 2010).

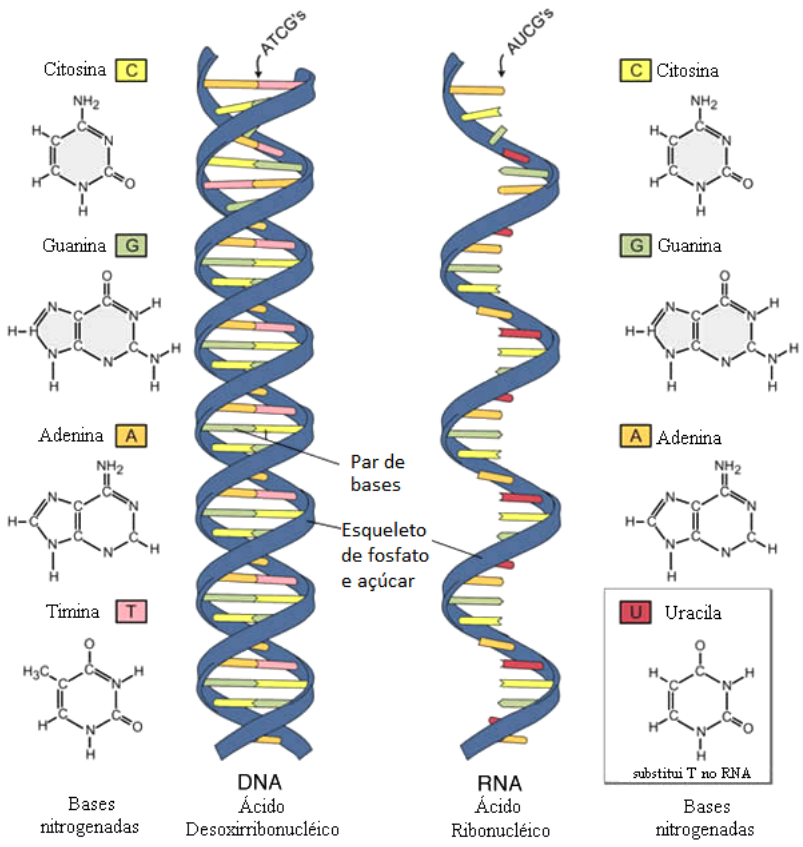


Figura 2.1 Fragmento da molécula de DNA fita dupla e RNA fita simples e suas respectivas bases nitrogenadas. Fonte: Traduzida de Wikipédia, (disponível sob licença *Creative Commons*).

A molécula de RNA é composta por uma única cadeia de nucleotídeos, enquanto a molécula de DNA possui duas cadeias de nucleotídeos ligadas através de ligações do tipo ponte de hidrogênio. Essas ligações dão estabilidade à estrutura da molécula, permitindo que a combinação das duas cadeias adquira uma forma helicoidal com giro para a direita (Nature-Education, 2010).

Em uma célula eucariótica, a molécula de DNA situa-se no interior do núcleo celular e é compactada e organizada por proteínas estruturais (histonas) em uma estrutura chamada **cromossomo**. Ao conjunto de cromossomos, cujo número e morfologia são característicos de uma espécie ou de seus gametas (espermatozóides e ovócitos), dá-se o nome de **cariótipo**.

Em sua maioria, as células dos seres humanos, com exceção dos gametas, hepatócitos e hemácias, possuem um cariótipo composto por 46 cromossomos combinados em 23 pares unidos pelo centrômero, sendo 22 pares de cromossomos autossomos e 1 par de cromossomos sexuais, sendo XX para as mulheres e XY para os homens. Os dois cromossomos de cada par, contudo, não são totalmente idênticos, sendo um oriundo da mãe e outro do pai (NIH, 2010).

O processo de replicação é uma característica essencial para manutenção dos organismos vivos, através dela é possível fazer cópias de suas células e assim dar continuidade à espécie. Esse processo se deve à capacidade que as células têm de replicarem seu material genético e passarem suas informações para seus descendentes.

2.1.2 Replicação do Material Genético

A replicação do DNA é um processo complexo dividido em duas fases: interfase e fase mitótica. A união dessas fases, permite que uma célula seja capaz de replicar 3 bilhões de pares de bases dos cromossomos em poucas horas e assim manter o ciclo celular (Miko e LeJeune, 2009).

A interfase é a etapa do ciclo celular que corresponde ao período entre o final de uma divisão celular e o início da próxima. Geralmente a célula encontra-se em interfase durante a maior parte da sua vida, sendo esta fase dividida em três sub-fases:

- G1 é a fase de alta síntese de proteínas, enzimas e RNA,
- S é a fase na qual ocorre a replicação de todo o material genético e consequente aumento da massa celular e,
- G2 é a fase na qual são sintetizadas as moléculas necessárias para a divisão celular.

Na fase mitótica, ocorre a organização, separação e compactação de todo o material genético para divisão da célula mãe, formando duas células filhas.

Todo o processo de replicação do DNA, deve ser preciso para manter a integridade do material genético. Para isso, as células possuem mecanismos de verificação de erros durante o processo do ciclo celular que permitem, ou não, o início de nova fase no ciclo celular.

Apesar disso, alguns erros podem permanecer no DNA. Os erros podem ocorrer em qualquer tipo de célula, mas em células germinativas, esses erros são perpetuados para as futuras gerações (Miko e LeJeune, 2009).

Este é um processo fundamental para a evolução de uma espécie, pois os erros perpetuados podem promover a produção de novos genes, que ao serem expressos produzem produtos

gênicos capazes de alterar vias metabólicas. A consequência disso é o surgimento da diversidade dentro da espécie, sendo algumas mais bem adaptadas do que outras.

2.1.3 Mutações e Polimorfismos

As **alterações genéticas** podem ocorrer pelo resultado da remoção, adição, substituição ou duplicação da informação genética contida no DNA. Estas alterações podem ser decorrentes de **polimorfismos** ou **mutações**. As alterações genéticas ocasionadas por mutações são divididas em:

- Mutações espontâneas: ocorrem quando há erros durante a replicação do DNA, gerando cópias de células com alterações. Essas mutações podem provocar diferentes anomalias ao metabolismo celular normal, como também, um crescimento descontrolado de células podendo resultar em câncer.

- Mutações induzidas: são provocadas por agentes externos, oxidativos, alquilantes, radiações eletromagnéticas de alta energia, vírus, entre outros agentes mutagênicos, que alteram a estrutura da molécula de DNA (Miko e LeJeune, 2009).

Uma alteração genética bem estabelecida é conhecida como polimorfismo (do grego: poli = muitas, e morfo = formas) que ocorre quando as modificações na sequência de DNA produzem duas ou mais formas alternativas de um mesmo gene (alelos).

O tipo mais comum de polimorfismo, com cerca de 10 milhões de variações já identificadas (Feuk *et al.*, 2006), é o *Single Nucleotide Polymorphism* (SNP, lê-se: "snip") que é caracterizado por apresentar uma variação pontual da base nitrogenada em uma posição específica no DNA.

A identificação dessa variação é o principal alvo de estudo nesta tese, devido à sua alta prevalência, influência na susceptibilidade de diversas doenças, como hipertensão arterial, diabetes mellitus, asma e câncer, entre outras (Cordell, 2002) e por provocar diferentes respostas na atuação dos medicamentos.

No entanto, é preciso considerar também a interação existente entre a carga genética e a influência de fatores ambientais, como tabagismo, hábitos alimentares, alcoolismo, medicamentos, fatores ocupacionais e radiação solar. Esses fatores, também influenciam de forma significativa nas manifestações clínicas de diferentes doenças. Mas, encontrar quais genes influenciam em uma determinada característica ou doença é um grande desafio para muitos pesquisadores.

É preciso deixar claro que, embora raros, alguns desses genes podem contribuir quase que totalmente para o aparecimento de uma doença (Figura 2.2 A), sendo conhecidos como **genes de "alta penetrância"**. Neste caso, o componente genético é muito forte em relação aos fatores ambientais, sendo mais facilmente detectados através de análise do genoma.

Casos de pacientes que possuem genes de alta penetrância, como BRCA1 e BRCA2, possuem também uma alta herdabilidade, ou seja, a proporção da variabilidade fenotípica que pode ser herdada. No entanto, esses casos representam apenas 5 a 10% do total desses casos de câncer de mama (Cancer-Society, 2010).

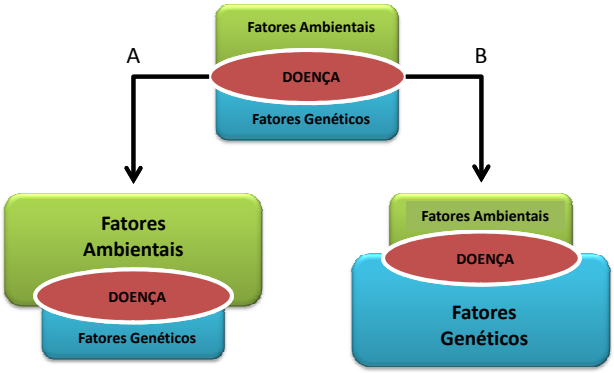


Figura 2.2 Esquema simplificado da influência dos fatores ambientais e genéticos na ocorrência de doenças. Em A) uma maior influência dos fatores ambientais sobre uma doença. Em B) maior influência dos fatores genéticos sobre uma doença.

No entanto, em sua grande maioria, os genes só podem levar ao desenvolvimento de uma doença quando associados a um conjunto de outros fatores de risco e/ou associados com outros genes (Figura 2.2 B). Esses genes são conhecidos como **genes de "baixa penetrância"**. Nesse caso, a interação de vários fatores, que incluem a presença de SNPs e fatores ambientais influenciam sobre a ocorrência de uma doença (Pinto, 2007).

Diante disso, os SNPs são um importante alvo de pesquisa, pois a sua identificação pode revelar um maior entendimento, sobre a influência dessas alterações pontuais no desenvolvimento de uma doença.

Polimorfismo de Nucleotídeo Único

Geneticistas tentaram por décadas encontrar as diferenças genéticas entre os indivíduos. As primeiras diferenças identificadas estavam relacionadas aos raros casos de variação de número e estrutura dos cromossomos, possíveis de serem detectadas utilizando um microscópio eletrônico (Feuk *et al.*, 2006). Com o avanço tecnológico do sequenciamento do DNA, foi possível detectar um número muito maior, de pequenas diferenças a um nível submicroscópico. Tais diferenças incluem variações em uma única base nitrogenada, revelando assim a presença dos SNPs no genoma humano.

Atualmente, as informações oriundas dos SNPs, podem ser empregadas nas mais diversas áreas, como medicina forense, evolução, definição de marcadores de predisposição a determinadas patologias, prognóstico a diferentes tratamentos, conservação e manejo de recursos genéticos, farmacogenética e desenvolvimento de vacinas, entre outras áreas. Dentre estes usos, temos o desafio de conseguir identificá-los e relacionar a presença dessas diferenças genéticas ao fenótipo, assim como, determinar o risco de certas doenças e a resposta a terapias.

As semelhanças entre o DNA genômico de duas pessoas é de 99,9% e as variações, embora representem uma parte muito pequena (0,1%) de todo o genoma, são regiões de grande interesse, pois são capazes de explicar as diferenças entre os indivíduos. Dentre as variações que ocorrem no DNA genômico, os SNPs são os mais comuns entre humanos, caracterizados por uma mudança pontual de um nucleotídeo. Estima-se cerca de 10 milhões dessas diferenças individuais (SNPs) ocorrendo a cada 300 pares de bases, ao longo dos aproximadamente três bilhões de pares de base existentes no genoma humano (Feuk *et al.*, 2006).

Os experimentos que detectam os SNPs, também podem detectar outros tipos de variações estruturais que ocorrem no genoma, como *copy-number variant* que incluem pequenas inserções, deleções e duplicações de pequenos segmentos de DNA (de 1 a 50 kilobases) (Stankiewicz e Lupski, 2010). Outros tipos menos comuns envolvem variantes de diferentes categorias de sequências repetitivas, tais como:

- Os microssatélites ou *Short Tandem Repeats* (STRs) que se caracterizam por apresentarem de 3 a 10 repetições de 1 a 4 pares de bases no DNA.
- Os minissatélites ou *Variable Number of Tandem Repeats* (VNTRs) são repetições de 20 a 200 pares de bases.
- *Short Interspersed Nuclear Elements* (SINEs) são caracterizados por duplicação em tandem de segmentos ricos em citosina e guanina (CG) separados por segmentos de adenina (A).
- o *Long Interspersed Nuclear Elements* (LINEs) são sequências repetidas de até 6500 pares de bases.

Para que uma variação genética seja considerada um SNP, a variante menos comum, ou seja, o alelo de menor frequência deve ocorrer na população acima de algum limiar preestabelecido, tal como em 1% ou mais da população. A sua ocorrência é independente da região do genoma, podendo ser encontrado em regiões codificantes (genes) ou mesmo em regiões não-codificantes, por exemplo, em sítios de ligação de fatores de transcrição, de *splicing*, de ancoragem de RNA*i*, entre outros.

SNPs Sinônimo e Não-Sinônimo

A presença de um SNP em uma região codificante no genoma humano, não necessariamente indica que ocorrerá uma modificação sobre o produto gênico formado, a este tipo de SNP dá-se o nome de **SNP sinônimo**. O SNP sinônimo ocorre quando, a mudança de uma única base nitrogenada altera a sequência, mas não altera o aminoácido formado.

Isto ocorre porque todos os aminoácidos, com exceção da metionina e do triptófano, são codificados por duas ou mais sequências distintas, esta propriedade é conhecida como a degeneração do código genético (Schwender e Ickstadt, 2008).

No entanto, mesmo SNPs sinônimos podem causar danos na expressão gênica, pois mesmo que a mudança do códon não altere o aminoácido, existem moléculas como o RNA*i*, que se ligam aos RNA*m* em posições de ancoragem específicas, que ao serem alterados impedem o acoplamento dessas moléculas podendo afetar a expressão dos RNA*m* na célula (Clancy, 2008).

Um outro tipo de SNP, denominado **SNP não-sinônimo (SNPns)**, é causado pela mudança de uma única base nitrogenada modificando o aminoácido formado. Em experimentos realizados por Gorre *et al.* (2001), foi identificado que a presença de um único SNP, alterava a interação entre proteína e o fármaco.

Um novo medicamento proposto para o tratamento da leucemia mieloide crônica, a droga STI571, que é um inibidor específico da atividade da proteína híbrida tirosina-quinase BCR-ABL, responsável pelo desenvolvimento dessa doença. Observou-se que um grupo de pacientes resistente à medicação, devido a um polimorfismo na posição 315 da proteína alvo. Isso ocorria devido à substituição de um aminoácido treonina por uma isoleucina, que nesta posição impedia a formação de uma ponte de hidrogênio essencial para a ligação do medicamento inibidor à proteína BCR-ABL (Gorre *et al.*, 2001).

Em alguns casos, mesmo ocorrendo a mudança do aminoácido, pela presença de SNPns, ainda assim pode acontecer de a proteína formada não ser especialmente sensível à variação naquele aminoácido específico, mantendo ou reduzindo ligeiramente a capacidade funcional (Teng *et al.*, 2009). Em outros casos, a mudança de uma base nitrogenada pode ser ainda mais nociva ao alterar a estrutura da proteína a tal nível que a torne inativa, ocasionando alteração de vias metabólicas essenciais ao funcionamento normal da célula.

Estudos realizados por Teng *et al.* (2009) mostraram que 70% das alterações causadas por SNPns provocam doenças que estão associadas a alterações que afetam o domínio conservado da proteína, região responsável pela sua função dentro da célula. Em outro estudo (Iughetti *et al.*, 2001), mostrou que a presença do SNPns aumenta em duas vezes e meia do risco de se desenvolver câncer de próstata.

Apesar da presença de um único SNP poder alterar o risco de desenvolvimento de uma doença, são raros os casos em que isso ocorre. Em vez disso, supõe-se que as interações, ou seja a combinação de diferentes SNPs, expliquem as diferenças existentes entre grupos de pessoas que apresentam um baixo ou um alto risco a desenvolver uma doença (Onay *et al.*, 2006).

Surge assim, a necessidade de estudar as interações entre os SNPs, também denominada de epistasia, como forma de verificar a sua influência sobre uma determinada característica física, disfunção metabólica ou susceptibilidade a uma doença.

Epistasia

A princípio, as análises em torno dos SNPs, concentravam-se em pesquisar a influência de um único SNP sobre uma doença, ou seja, SNP com **efeito principal** (Yang *et al.*, 2010). Este tipo de variação genética que possui influência individual suficiente para alterar um fenótipo ocorre nas doenças monogênicas (Yang *et al.*, 2010).

Em doenças complexas ou multifatoriais, tipo mais comum de doença na população, ocorre a influencia de múltiplas variações genéticas. Este tipo de efeito combinado entre as variantes genéticas é conhecido como **efeito epistático** (Yang *et al.*, 2010), e é caracterizado por SNPs, com pouco ou nenhum efeito individual, mas que apresentam forte influência sobre o fenótipo quando estão atuando em conjunto.

Na Figura 2.3 é apresentada uma ilustração da interação epistática entre SNPs. É possível observar que ambos SNP_1 e SNP_2, possuem uma distribuição balanceada com 4000 pacientes sendo 2000 casos e 2000 controles. O efeito epistático entre os SNP_1 e SNP_2 pode ser observado nos 9 quadros centrais da Figura 2.3. Nestes quadros, as distribuições entre casos e controles torna-se variável de acordo com as combinações dos genótipos, tendo os genótipos AABB, AAbb, AaBb e aaBb um maior número de pacientes do tipo caso.

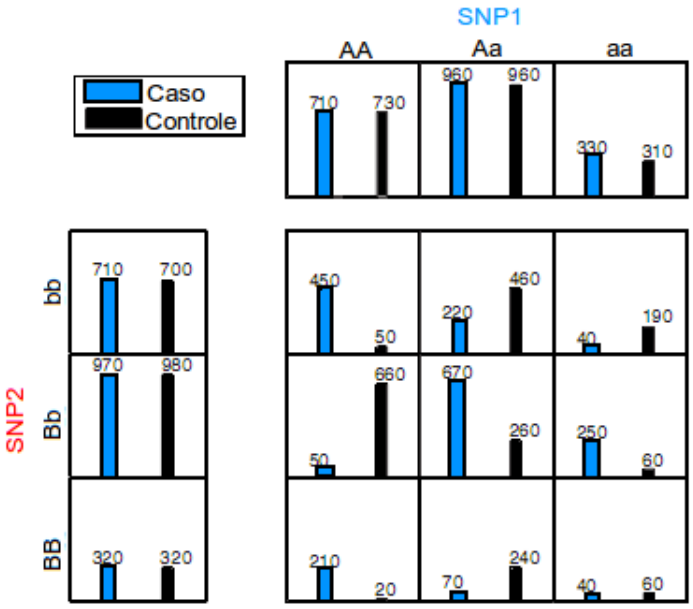


Figura 2.3 Ilustração dos SNP_1 e SNP_2, exibindo uma distribuição de pacientes casos e controles de acordo com os diferentes alelos: homozigoto dominante (AA ou BB), heterozigoto (Aa ou Bb) e homozigoto recessivo (aa ou bb). O efeito combinado entre os SNPs exibe uma distribuição diferenciada entre casos e controles, caracterizando o efeito epistático.

Desequilíbrio de Ligação

Além da verificação da atuação da interação entre os SNPs sobre um fenótipo. Também é estudado o princípio do Desequilíbrio de Ligação (LD, do inglês: *Linkage Disequilibrium*). O princípio de LD é dado pela correlação existente entre SNPs, ou seja, é uma associação não-aleatória de SNPs (Ardlie *et al.*, 2002).

O desequilíbrio de ligação, ocorre quando, dois ou mais alelos específicos, em loci distintos de um mesmo cromossomo, são mais vezes encontrados em conjunto do que separados.

Quando isso acontece, considera-se que os loci estejam em desequilíbrio e que a identificação de um SNP em um locus fornece informações sobre SNPs em outros loci.

As análises de LD são especialmente utilizadas para o mapeamento dos dados genômicos. Sendo mais mais efetivas quando são utilizados dados de populações isoladas, pois possuem menor heterogeneidade alélica, ou em análises de doenças causadas por mutações mais antigas, como displasia distrófica em finlandeses (Botstein e Risch, 2003).

Estas diferentes características e processos biológicos apresentados, são fundamentais para compressão da atuação dos SNPs sobre as doenças e construção de diferentes modelos de dados, capazes de simular o problema da inferência de SNPs relevantes. Na seções seguintes será apresentado a doença de Alzheimer, descrevendo suas características e fatores genéticos já identificados, relacionando seus principais SNPs e genes causadores da doença.

2.2 Doença de Alzheimer

Em 1906, Alois Alzheimer, um médico alemão, realizou uma autopsia em um mulher de 55 anos de idade com um histórico de deterioração mental progressiva. Seu córtex cerebral, parte do cérebro responsável pelo raciocínio e memória, apresentava uma redução significativa da massa celular. Foram identificados estranhos pacotes de fibras, nomeados de **emaranhados neurofibrilares** e uma acumulação de restos celulares denominados de **placas senis** (Grimm *et al.*, 2016).

2.2.1 Sintomas

A doença de Alzheimer é a forma mais comum de demência, a doença se caracteriza pela perda e/ou redução progressiva das capacidades cognitivas (Brookmeyer *et al.*, 2007). Até o momento, é uma doença incurável, mas que pode e deve ser tratada o mais precocemente possível, a fim de retardar o seu avanço e assim ter um maior controle sobre os sintomas.

Os primeiros sintomas são muito comuns na velhice e surgem de forma muito sutil, por esta razão, são facilmente confundidos com sinais relacionados com a idade ou estresse. Nos primeiros estágios da doença, o sintoma mais comum é a dificuldade em recordar eventos recentes, ou seja, perda de memória de curto prazo.

A doença progressivamente passa para estágios mais severos, apresentando sintomas como confusão mental, irritabilidade, alterações de humor, comportamento agressivo, dificuldades com a linguagem e perda de memória de longo prazo. Por fim, a morte de células neurais é tão severa, que o corpo vai perdendo a capacidade de realizar funções normais para sua manutenção, o que acaba por levar à morte (Grimm *et al.*, 2016).

2.2.2 Alterações Cerebrais

O motivo pelo qual, a doença se manifesta ainda não é conhecido, no entanto como relatado em 1906 por Alois Alzheimer, as principais alterações cerebrais que marcantemente representam a doença de Alzheimer (AD) podem ser macro ou microscópicas:

- Sob o ponto de vista macroscópico, a redução do número das células nervosas (neurônios) e das ligações entre elas (sinapses) provocam uma redução progressiva do volume cerebral. Essa redução da quantidade de células nervosas, pode ser visualmente observado em estágios mais severos da doença, como pode ser observado na Figura 2.4.

- As alterações microscópicas são provocadas pela presença das placas senis, sendo estas decorrentes da deposição da **proteína β-amiloide**. E dos emaranhados neurofibrilares que são formados pela hiperfosforilação da **proteína tau**.

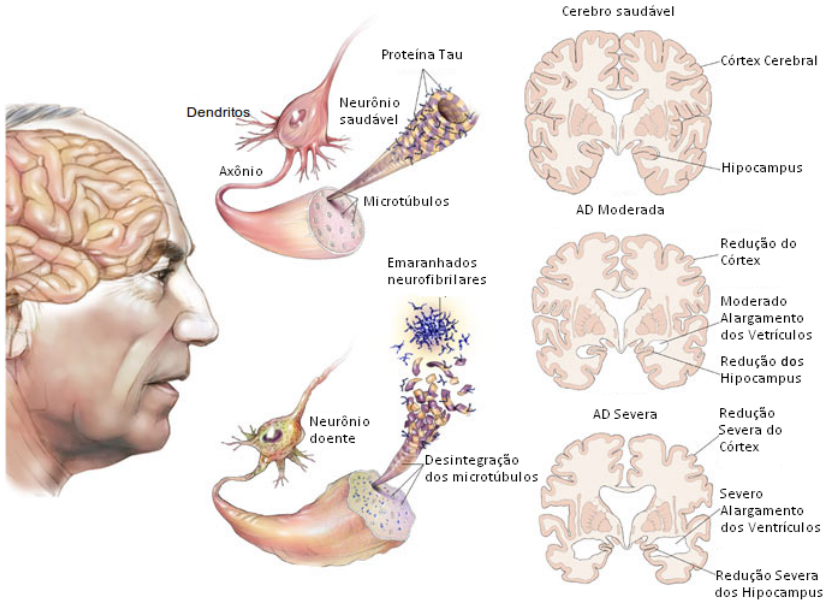


Figura 2.4 Cérebro humano, comparando o córtex cerebral em um cenário normal, com avanços moderado e severo da doença de Alzheimer (AD). Fonte: Alzheimer's-Association (2016)

As placas senis, também conhecidas como placas amiloides ou neuríticas, são placas extracelulares produzidas através de um erro na clivagem da proteína precursora da amiloide (APP). Esse erro da clivagem, é fortemente associado a mutações nos genes PSEN1 e PSEN2, produzindo fragmentos da proteína β-amiloide.

Devido ao caráter insolúvel da proteína β-amiloide, ocorre sua deposição no tecido neural e seu acúmulo promove a formação das placas senis. Estas placas causam bloqueios na comunicação entre neurônios e promovem a ativação do sistema imunológico contra o próprio corpo, causando a destruição dos neurônios.

Em contrapartida, os emaranhados neurofibrilares são formados a partir de uma alteração na proteína tau. Esses emaranhados, tendem a ser mais abundantes em áreas cerebrais com maior destruição neuronal. Iniciando na região do hipocampo e nas zonas adjacentes ao lóbulo temporal, sendo estas as principais estruturas cerebrais responsáveis pela memória.

Na Figura 2.5 é mostrada em tons de azul a progressão da deposição das placas senis e emaranhados neurofibrilares no córtex cerebral ao longo do tempo. A velocidade de progressão da doença varia de pessoa para pessoa, mas em geral, inicia-se na região do hipocampo e com a progressão da doença uma porção cada vez significativa do cérebro vai sendo acometido pela doença, sendo a expectativa média de vida em torno de 8 anos, podendo chegar alcançar 20 anos.

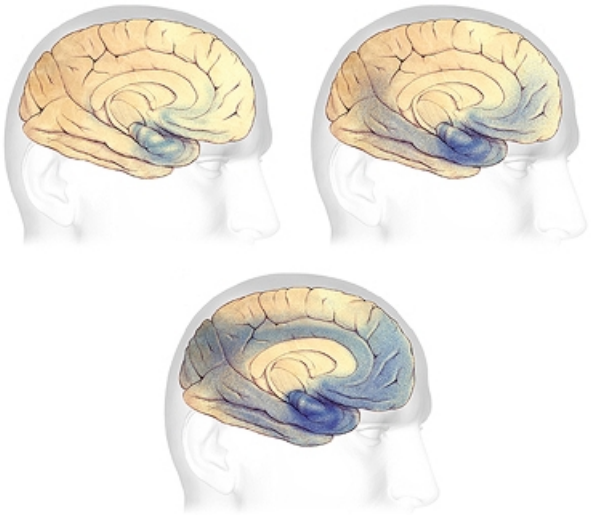


Figura 2.5 Em azul a progressão da deposição das placas senis e emaranhados neurofibrilares no córtex cerebral. Fonte: Alzheimer's-Association (2016)

2.2.3 Prevalência

Com o envelhecimento da população mundial, um maior conhecimento sobre a biologia molecular da doença é essencial para se introduzir novas metodologias e técnicas mais efetivas no combate, prevenção ou retardo da doença. Estima-se que no mundo existam cerca de **35,6 milhões de pessoas com AD** e prevê-se que em 2050 a doença afete uma em cada 85 pessoas em uma escala mundial (Brookmeyer *et al.*, 2007).

A doença pode ser caracterizada em dois tipos: o primeiro, e sendo este o tipo mais comum com a 95% dos casos, ocorre em pessoas com idade >60-65 anos, também denominada **AD *late onset***. O outro tipo, acomete apenas 5% dos casos, e ocorre em pessoas idade inferior a 65 anos, denominada **AD *early onset***.

Embora a doença afete apenas 6 a 10% dos idosos entre os 65 e 70 anos, a sua prevalência aumenta exponencialmente com a idade, sendo 30% aos 80 anos e mais de 60% depois dos 90 anos (Brookmeyer *et al.*, 2007).

2.2.4 Fatores de risco na AD

O histórico familiar é um importante indicador do risco da doença, estima-se que a herdabilidade da doença é de 79% e que aproximadamente 25% dos casos de AD é de herança familiar, os 75% restantes são causadas pela interação entre genética e fatores ambientais (Gatz *et al.*, 2006).

Vários genes e SNPs vêm sendo identificados como fatores de risco da doença. Dentre estes, mutações em quatro genes: **Presenilin 1** (PSEN1) localizado no cromossomo 14, **Presenilin 2** (PSEN2) localizado no cromossomo 1, gene da **Proteína Precursora da Amilóide** (APP) localizado no cromossomo 21 são associados com a ocorrência da doença e polimorfismos específicos no gene **Apolipoproteína E** (APOE) localizado no cromossomo 19 são identificados como fatores de risco associados com a AD.

No gene PSEN1, podem existir mutações associados com a AD, sendo responsável por cerca de 30 a 75% dos casos de AD *early-onset* familiar. No gene PSEN2, as mutações presentes promovem menos de 5% de todos os casos de AD *early-onset* familiar. Enquanto que no gene APP, aproximadamente 10 a 15% dos casos de AD *early-onset* familiar são promovidas por mutações nesse gene. No gene APOE, a variante $\varepsilon 4$ é a principal responsável por aumentar a susceptibilidade de casos de AD *late-onset*. Assim, estes genes juntos são responsáveis por menos de $\frac{1}{4}$ da proporção de casos de AD. Os $\frac{3}{4}$ restantes, são fatores de risco que estão ainda por ser identificados (Williamson *et al.*, 2009).

Indivíduos com síndrome de Down costumam desenvolver AD *early-onset*, pois possuem uma cópia extra do cromossomo 21 que contém o gene APP responsável pelo controle da produção da proteína β-amiloide.

A maioria das variações genéticas descritas são correlacionadas a modificações no DNA nuclear, sendo amplamente estudadas devido a sua influência no desenvolvimento de diferentes doenças, mas o advento de técnicas mais sofisticadas de sequenciamento como o NGS permitiu que o DNA mitocondrial (DNA*mt*) também fosse alvo de estudo (Devall *et al.*, 2016).

Atualmente, poucos são os estudos que avaliam o material genético mitocondrial, mas é crescente o interesse da comunidade científica sobre os padrões de herança e metilação[1] na influência do DNA*mt* nas doenças. Apenas recentemente, o padrão de metilação do DNA*mt* tem sido associado a diferentes doenças neurodegenerativas como esclerose lateral amiotrófica e AD (Devall *et al.*, 2016).

Outros fatores de risco foram identificados em estudos epidemiológicos, mostrando que as mulheres representam $\frac{2}{3}$ dos casos de AD, sendo a deficiência hormonal que ocorre pós menopausa, apontada com principal fator (Grimm *et al.*, 2016). Assim, a combinação entre o envelhecimento e a redução de hormônios circulantes, especialmente em mulheres podem representar fatores de risco adicionais da AD. No entanto, os mecanismos que envolvem o

[1] O padrão de metilação do DNA é de grande interesse na epigenética, pois a posição em que o radical metil é inserido, pode alterar a expressão gênica.

aumento do risco da doença devido a esses fatores ainda não são bem compreendidos, incluindo também as implicações do DNA*mt* neste processo (Grimm *et al.*, 2016).

Assim, para identificar e compreender os diferentes fatores de risco e as variantes genéticas relacionadas com a doença de Alzheimer. Várias bases de dados buscam unir informações de diferentes pacientes com etnias e idades diferentes.

2.2.5 Bases de Dados

A base de dados Alzgene (Bertram *et al.*, 2007) disponibiliza informações de 1395 diferentes estudos com dados de GWAS relacionados com a AD. Sendo descritos 695 genes e 2973 SNPs, pertencentes aos cromossomos do genoma humano, incluindo cromossomos sexuais e mitocondrial que estão relacionados com a doença.

Uma outra base de dados, a ADNI (*Alzheimer's Disease Neuroimaging Initiative*) é utilizada como fonte do conjunto de dados reais. A base de dados é composta por 757 pacientes (449 homens e 308 mulheres) com um total de 620.901 SNPs.

Conclusão

Neste capítulo foram apresentados vários conceitos biológicos para permitir ao leitor uma maior compreensão sobre as moléculas do DNA e RNA, e com o uso dessa informação compreender sobre sua estrutura e funções dentro da célula. São nessas moléculas que os SNPs, variações genéticas de interesse, estão inseridos e as modificações provocados por eles podem determinar um maior risco a determinadas doenças.

Na Doença de Alzheimer, por exemplo, mostramos que existem diversos fatores genéticos e ambientais que contribuem para aumentar o risco de ocorrência. Dentre estes fatores, destacamos a presença do polimorfismo no gene APOE que pode aumentar em até 10 vezes a incidência da doença. Para estudar esses fatores, são verificadas as interações entre eles, utilizando conjuntos de dados de pacientes casos e controles. Esses conjuntos de dados são utilizados em diferentes experimentos nesta tese e são apresentados no próximo capítulo.

3

Conjuntos de Dados

Neste capítulo, são apresentados conjuntos de dados simulados e reais, baseados em dados de pacientes dos tipos casos e controles. Esses dados são utilizados na avaliação do desempenho de diferentes abordagens computacionais, que realizam a inferência das interações entre SNPs.

Inicialmente serão apresentadas características e a estrutura dos dados, e nas seções seguintes, serão apresentados os três conjuntos de dados simulados, especificando as diferenças entre eles e ao final, o conjunto de dados real, composto por pacientes acometidos com a doença de Alzheimer.

3.1 Características dos Dados de Polimorfismo

Nesta tese, serão utilizados dados de **estudo de caso-controle** sem relação parental, este é o tipo de dado de mais fácil obtenção, devido ao grande número de casos que podem ser investigados, pois, não há relação de dependência dos pacientes com graus de parentesco. No entanto, estes dados estão sujeitos a uma maior diversidade genética, oriunda por exemplo, da variabilidade existente entre as diferenças entre as raças: caucasiana, negra e asiática.

Em alguns estudos, para se reduzir a influência da miscigenação dos dados populacionais, são selecionados grupos de pessoas de uma mesma raça, ou são feitos estudos de *family-based design*. Este último tem vantagens, pois reduz de forma significativa a variabilidade genética e os efeitos de fatores ambientais, pois são analisados membros de uma mesma família. No entanto, a principal desvantagem de *family-based design*, é acumular uma quantidade de amostras estatisticamente suficiente de famílias bem caracterizadas, e por isso, este tipo de estudo representa uma minoria das investigações que avaliam variantes genéticas em doenças complexas.

Uma outra forma de avaliar os métodos computacionais, é utilizando dados simulados que incorporam características genéticas. Os **dados simulados** possuem vantagens, pois garantem ao pesquisador, um conhecimento *a priori* das características dos dados estudados. Assim como, pode-se simular a presença de erros ou de dados faltantes, tornando-os mais previsíveis e

confiáveis para realizar testes de avaliação e capacidade de diferentes ferramentas.

Para simular as características biológicas em dados simulados, são utilizadas funções de penetrância. O papel da função de penetrância sobre os dados simulados, é modelar a relação existente entre as variações genéticas e o risco de doença, ou seja, um modelo de penetrância é definido como a probabilidade de ocorrência de uma doença dada uma combinação de genótipos.

Estas funções penetrância podem ser contínuas ou discretas, e podem incluir medidas de co-variáveis envolvendo mais de um locus. Todas as outras variáveis que não estão associadas explicitamente ao modelo de penetrância são assumidas como um distribuição normal e aleatória entre os indivíduos. Sendo assim, os modelos que utilizam uma função de penetrância são aplicados sobre um conjunto de SNPs, aos quais se deseja simular uma doença, enquanto os demais são aleatorizados.

Após definidos os SNPs relevantes, os dados simulados utilizados nos experimentos, são representados como matrizes M *x* N, onde M é a quantidade de pacientes e N o número de SNPs. Cada paciente é representado por uma variável dicotômica de tamanho M, que representa a classe, e diferencia os pacientes casos, dos controles.

Nessas matrizes, os SNPs são apresentados em três formas distintas: homozigoto dominante, heterozigoto e homozigoto recessivo, representadas pelos valores 0, 1 e 2, respectivamente. E a classe que divide os pacientes em casos e controles, atribui-se 0 para pacientes controle e 1 para caso. Dessa forma, a matriz de dados simulados pode ser representada de acordo com a Tabela 3.1.

Tabela 3.1 Representação da matriz de pacientes casos e controles, exibindo os SNPs pelos valores 0, 1 e 2; e a variável classe definida como 0 para os pacientes controles e 1 para casos.

Paciente	SNP_0	SNP_1	SNP_2	SNP_3	SNP_4	...	SNP_n	**Classe**
1	1	1	0	2	0	...	1	0
2	0	0	2	1	1	...	2	1
3	0	0	0	0	2	...	1	0
4	0	1	0	2	2	...	1	1
5	1	0	1	2	1	...	1	0
6	0	0	1	0	0	...	2	1
7	1	0	2	1	0	...	1	1
8	0	1	0	2	1	...	0	0
9	0	2	1	2	1	...	2	1
10	1	2	2	1	2	...	2	0

Outros tipos de informação, pode ser utilizada para facilitar na identificação das interações epistáticas, como sexo, idade, outros fatores de risco e também informações conhecidas como *prior knowledge* (Minelli *et al.*, 2013). As informações do tipo *prior knowledge*, incorporam informações biológicas a respeito dos SNPs ou genes estudados, como sua posição no DNA, o tipo do SNP (SNP sinônimo ou SNP não sinônimo), se modifica uma região de ligação da proteína, se é um SNP já conhecido por apresentar associação com a doença, entre outras

informações. Esse conhecimento prévio sobre os SNPs, pode ajudar a identificar mais facilmente os SNPs relevantes e suas interações epistáticas.

3.2 Dados Simulados

Nesta seção, serão apresentados três conjuntos de dados simulados com diferentes características biológicas, sendo os dois primeiros, compostos por interações entre 2 SNPs e diferentes características biológicas e o terceiro com interações de alta ordem entre os SNPs (3, 4 e 5 interações entre SNPs).

3.2.1 Interações entre 2 SNPs - Velez

Os conjuntos de dados simulados com interações entre 2 SNPs, possuem quatro diferentes valores de número de SNPs (20, 50, 100 e 1000 SNPs) e dois valores para o tamanho da população de pacientes casos e controles (800 e 1600 pacientes).

Cada conjunto de dados é formado por **70 modelos epistáticos**, com diferentes funções de penetrância construídas por Velez *et al.* (2007). Cada modelo epistático é caracterizado por uma combinação de valores de herdabilidade[1] (HERD) e frequência do menor alelo[2] (MAF), sendo sete valores para a HERD 0.01, 0.025, 0.05, 0.1, 0.2, 0.3, e 0.4 e os dois valores para o MAF 0.2 e 0.4.

Para cada uma dessas combinações entre MAF e herdabilidade temos 5 subconjuntos de modelos epistáticos. Na Tabela 3.2, são descritos os 70 modelos epistáticos em grupos com 5 modelos, de acordo com a combinação de valores entre a HERD e MAF. Cada modelo é construído com 100 amostras totalizando assim 56.000 amostras ($2 \times 4 \times 70 \times 100$). Os dados simulados foram disponibilizados por Velez *et al.* (2007).

Em cada amostra, estão presentes um par de SNPs relevantes, que ao serem analisados, devem ser ordenados como os mais relevantes dentre os demais SNPs. Como cada modelo epistático possui 100 amostras, o número de respostas corretas para cada modelo, deve ser um valor entre 0 a 100. Sendo 100, quando o método identificar todos os pares de SNPs relevantes em todas as amostras analisadas.

3.2.2 Interações entre 2 SNPs - Shang

O conjunto de dados proposto por Shang *et al.* (2016), é composto por duas interações relevantes entre SNPs. Este conjunto de dados possui dados com 100 e 10000 SNPs e uma população balanceada de 2000 pacientes casos e 2000 controles.

As combinações dos genótipos são representadas por seis modelos, com diferentes funções epistáticas combinadas com o número de SNPs e pacientes. Dentre os seis modelos epis-

[1] Informa o quanto o fator genético influencia o fenótipo.
[2] Frequência com que o alelo menos comum ocorre na população.

Tabela 3.2 Modelos Epistáticos combinados aos valores da herdabilidade e MAF.

Modelos	Herdabilidade	MAF
0-4	0.4	0.2/0.8
5-9	0.2	0.2/0.8
10-14	0.1	0.2/0.8
15-19	0.05	0.2/0.8
20-24	0.3	0.2/0.8
25-29	0.4	0.4/0.6
30-34	0.2	0.4/0.6
35-39	0.1	0.4/0.6
40-44	0.05	0.4/0.6
45-49	0.3	0.4/0.6
50-54	0.025	0.2/0.8
55-59	0.01	0.2/0.8
60-64	0.025	0.4/0.6
65-69	0.01	0.4/0.6

táticos selecionados, os modelos 1 e 2 são caracterizados pela combinação de SNPs com efeito principal, ou seja, quando um único SNP influencia o fenótipo, e com efeitos epistáticos, quando dois ou mais SNPs influenciam o fenótipo. Os demais modelos, do 3 ao 6, são representados apenas por efeitos epistáticos.

Cada modelo consiste de um total de 100 amostras que estão em equilíbrio de Hardy-Weinberg. Os valores do MAFs são diferentes, para cada SNP em cada modelo epistático. O modelo 1 é composto por MAF(a) 0.3, MAF(b) 0.2, o modelo 2 pelo MAF(a) 0.4, MAF(b) 0.2. Os modelos 3, 4, 5 e 6 tem os mesmos valores de MAF para ambos SNPs relevantes, sendo o modelo 3 com MAF 0.2, o modelo 4 com MAF 0.4 e os modelos 5 e 6 com MAF 0.5.

3.2.3 Interações de Alta Ordem - Himmelstein

Assim como, nos conjuntos de dados com interação entre pares de SNPs, os dados simulados com interações de alta ordem, proposto por Himmelstein *et al.* (2011), possuem diferentes combinações no número de SNPs (20, 100 e 1000 SNPs) e uma população de pacientes casos e controles com 800 e outra com 1600 pacientes. No entanto, neste caso, os dados não são criados a partir de modelos genéticos predefinidos, mas sim, como dados de modelagem livre ou *model free*.

Para a criação deste conjuntos de dados, Himmelstein *et al.* (2011) desenvolveram um algoritmo para que as amostras de dados criadas tivessem uma maior diversidade, sem a limitação de diversidade imposta pelos modelos genéticos pré-definidos. Assim, os conjuntos de dados foram construídos considerando reduzir a preditividade de SNPs individuais e pares de SNPs em relação ao fenótipo, enquanto que a preditividade de interações de terceira, quarta e quinta ordem eram maximizadas.

Os dados são compostos por 100 amostras em equilíbrio de Hardy-Weinberg e possuem variações no número de interações 3, 4 e 5 interações. Este modelo possui 1.800 amostras ($2 \times 3 \times 3 \times 100$) a serem analisadas.

Após a introdução das diferentes características dos conjuntos de dados simulados, será então apresentado, o conjunto de dados reais envolvendo pacientes acometidos com a doença de Alzheimer.

3.3 Dados Reais

O conjunto de dados real, foi obtido a partir do banco de dados **Alzheimer's Disease Neuroimaging Initiative** (ADNI) através do link: adni.loni.usc.edu. O conjunto de dados da doença de Alzheimer, disponibilizado pelo ADNI (Saykin *et al.*, 2010), é composto por 757 pacientes da raça Caucasiana (449 homens e 308 mulheres) com um total de 620.901 SNPs. Os pacientes tem idades entre 55 a 90 anos, oriundos de 50 locais diferentes dos Estados Unidos e Canadá.

Dos 757 pacientes, 214 são pacientes saudáveis e os demais pacientes acometidos com alguma doença. Dentre os pacientes doentes, temos três grupos: os 175 pacientes acometidos pela Doença de Alzheimer (AD), os 230 pacientes acometidos com MCI (*mild cognitive impaiment*) e os 138 pacientes acometidos com LMCI (*late mild cognitive impaiment*).

Considerando o objetivo proposto na tese em estudar os pacientes com AD. Do conjunto de dados original, foram removidos os pacientes e SNPs relacionados com MCI e LMCI. O conjunto de dados final, contém um total de **389 pacientes (214 controles e 175 casos) e 329.983 SNPs**. Além dessa seleção dos pacientes, vale ressaltar que conjuntos de dados reais podem ser acometidos com vários problemas inatos. Diante disso, antes de realizar qualquer tipo de análise é necessário que esses dados sejam previamente processados através de metodologias de controle de qualidade, para remoção de erros de genotipagem, tratamento de dados faltantes, dentre outros (Turner *et al.*, 2011). O pré-processamento dos dados será detalhado no Capítulo 6.

Genes e SNPs Relevantes

Avaliando estudos sobre a doença de Alzheimer, foi verificado que o gene APOE (SNP rs429358) e TOMM40 (SNP rs2075650), vem sendo apontados com os principais marcadores associados com as doenças MCI e AD. Outros SNPs considerados relevantes incluem o rs10932886 (EPHA4), rs7610017 (TP63) e rs6463843 (NXPH1) Shen *et al.* (2010).

Assim, foi realizada uma pesquisa bibliográfica, para enumerar os SNPs reportados como relevantes fatores de risco da AD. Informações sobre genes e SNPs, foram coletadas na base de dados **AlzGene** (Bertram *et al.*, 2007), oriundas de diferentes estudos científicos associados com a AD Shen *et al.* (2010). O objetivo dessa pesquisa, é utilizar essas informações coletadas para avaliar a qualidade dos resultados obtidos com a inferência das ferramentas nos dados reais.

Através da base de dados AlzGene, foram identificados 674 genes associados com a AD. A distribuição desses genes pelos cromossomos, pode ser observada nas barras verticais em azul na Figura 3.1. Desses 674 genes, 296 genes, representados pela barra vertical em vermelho, foram relacionados positivamente em pelo menos um estudo científico com a AD e 102 genes, representados pelas barras verticais em verde, em mais de um estudo.

Observando o gráfico da Figura 3.1, verifica-se que o cromossomo 10, possui o maior número de genes reportados em estudos com a AD. Os estudos realizados por Myers *et al.* (2000), mostram que existem evidências de um fator de risco para AD *late onset*, localizado no cromossomo 10. Este fator de risco está associado ao metabolismo da proteína β-amiloide, que é conhecidamente a causadora da destruição de células neuronais na AD. Os demais cromossomos com uma grande quantidade de genes (≥ 40) são, os cromossomos 1, 6, 11, 12 e 19. Destes, no cromossomo 1, é encontrado o gene PSEN2 e no cromossomo 19 o gene APOE, ambos são fatores de risco amplamente conhecidos por estarem associados a AD.

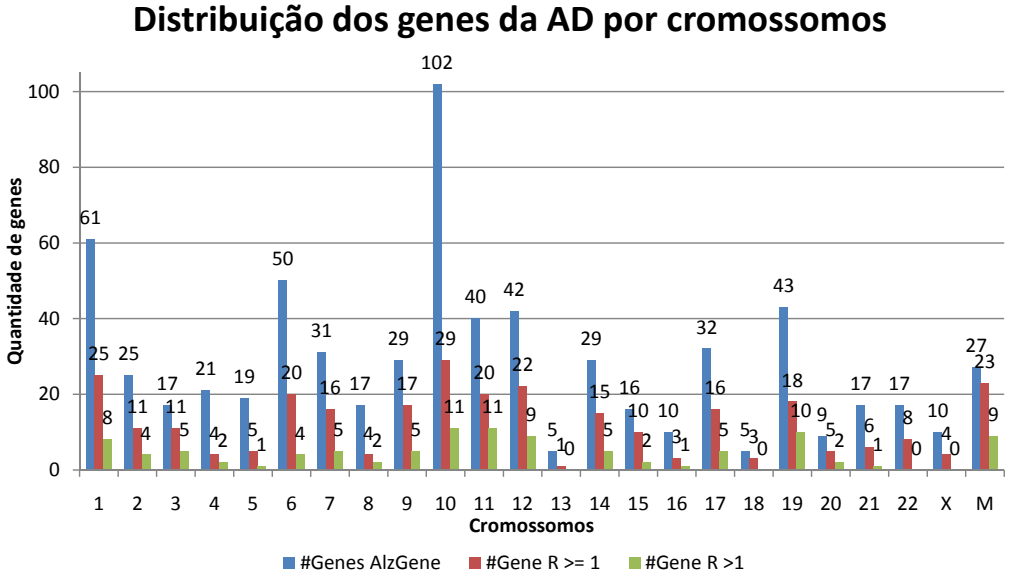


Figura 3.1 Gráfico de barras com a distribuição dos genes: Em azul 674 genes; Em vermelho 296 genes relevantes em pelo menos 1 estudo; E em verde 102 genes relevantes em mais de um estudo. Os genes estão distribuídos nos 22 cromossomos, incluindo o cromossomo sexual (X) e o DNA Mitocondrial (M).

Os genes com maiores associações positivas relacionados com a AD, são apresentados na Figura 3.2. As barras em azul representam a quantidade de vezes que o gene foi associado positivamente com a AD, enquanto que as barras em vermelho indicam o número de estudos realizados, nos quais o gene não teve alta correlação com a doença, sendo então considerado como uma associação negativa.

Outras investigações, utilizando o conjunto dos 296 genes relevantes, foram realizados para identificar outras doenças relacionadas com estes genes. Para isso, foram utilizadas informações contidas em Goh *et al.* (2007) e omim.org, e os resultados obtidos são apresentados na Tabela 3.3. Diferentes tipos de doenças, tais como cardiopatias e diabetes tem sido frequentemente associadas a pacientes com AD e nos resultados obtidos, foi obtido uma maior ocorrência

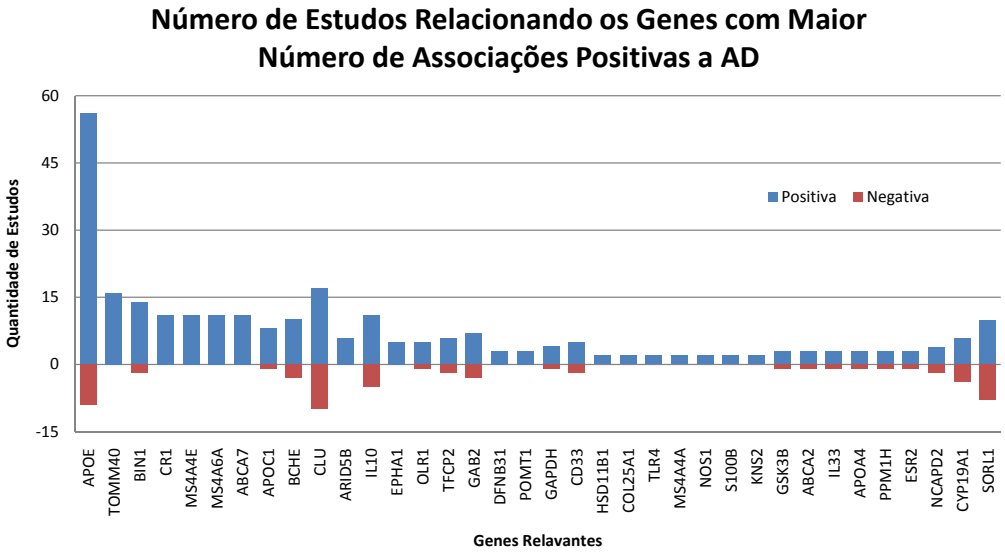


Figura 3.2 Gráfico de barras exibindo os 36 genes com maiores associações positivas com AD dentre os 296 citados em estudos caso e controle. As barras em azul indicam o número de estudos que identificaram o gene positivamente associado a AD. Barras em vermelho indicam a quantidade de estudos que investigaram o gene no entanto, este não foi considerado relevante no estudo.

dessas doenças, podendo representar possíveis fatores de risco relacionados a AD.

A partir dos 296 genes relevantes, também foram identificados 116 SNPs desses, 76 foram identificados no conjunto de dados reais do ADNI sem nenhum dado faltante. Esses SNPs serão utilizados em futuras consultas para avaliar o desempenho dos métodos de inferência das interações epistáticas. A lista dos genes e SNPs identificados é apresentado no Apêndice desta tese.

Conclusão

Foram apresentados neste capítulo quatro diferentes conjuntos de dados, sendo um deles dados reais de pacientes com Alzheimer (ADNI) e os outros três simulados com diferentes quantidades de pacientes, número de SNPs, modelos epistáticos e características biológicas, como frequência do menor alelo, epistasia e herdabilidade.

Os dados simulados com duas interações são produzidos com diferentes funções epistáticas, no conjunto proposto por Velez *et al.* (2007), sendo funções gradualmente modificadas, tornando cada vez mais difícil detectar as interações, à medida que a herdabilidade e o MAF diminuem. Nos dados propostos por Shang *et al.* (2016) as funções epistáticas aplicam dois tipos de efeitos sobre os SNPs relevantes, o efeito principal e o epistático. O último conjunto simulado, proposto por Himmelstein *et al.* (2011) insere interações de alta ordem nas diferentes amostras.

No conjunto de dados real, foi realizada uma seleção de pacientes e uma extensa pesquisa bibliográfica. Esta pesquisa teve como objetivo, dar embasamento aos resultados obtidos,

Tabela 3.3 Doenças Relacionadas com os Genes Relevantes na AD.

Classe da doença	Cr[1]	Gene	Doença
Cardiovascular	1	PSEN2	Cardiomiopatia
	6	HFE	Microvascular Complicações da Diabetes
	7	NOS3	Hipertensão
	7	NOS3	Espasmos Coronários
	7	NOS3	Derrame
	14	PSEN1	Cardiomiopatia
	17	ACE	Derrame
	17	ACE	Infarto do Miocárdio
	17	ACE	Microvascular Complicações da Diabetes
	19	APOE	Hiperlipoproteinemia
	19	APOE	Infarto do Miocárdio
Endócrina	17	ACE	Diabetes Mellitus
	17	ACE	Conversão da Angiotensina
Hematológica	6	HFE	Porfiria
	10	PLAU	Desordem Plaquetária
	19	APOE	Histiócitose
Imunológica	10	ACE	Síndrome Respiratória Aguda Grave
	10	MPO	Deficiência na Mieloperoxidase
Metabólica	6	HFE	Índice de Transferrina no soro
	6	HFE	Hemocromatose
	12	A2M	Deficiência da A2M
	19	APOE	Hiperlipoproteinemia
Neurológica	14	PSEN1	Doença de Pick
	14	PSEN1	Demência
	21	APP	Amiloidoses
Oftalmológica	19	APOE	Degeneração Macular
Psiquiátrica	1	AD7CNTP	Transtorno Bipolar
	21	APP	Esquizofrenia
Renal	17	ACE	Disgenesia renal tubular
	19	APOE	Glomerulopatia
Não-Classificada	7	NOS3	Descolamento de placenta

[1] Cr é abreviatura para Cromossomo.

identificando genes e SNPs relevantes, a serem confrontados com os resultados da inferência das interações epistáticas pela ferramenta de inferência proposta.

No Capítulo a seguir, serão revisadas diferentes abordagens computacionais, que serão avaliadas com os diferentes conjuntos de dados aqui apresentados.

4

Conceitos Básicos em Computação

Neste capítulo serão apresentadas diferentes abordagens computacionais, estudadas para lidar com os dados de polimorfismos. Inicialmente será feita uma revisão sobre diferentes filtros de seleção de características. Na Seção 4.2 serão apresentadas ferramentas de inferência de epistasia, classificadas de acordo com a sua forma de busca. Na Seção 4.3, serão apresentados algoritmos baseados em *Learning Vector Quantization* (LVQ), investigados com o propósito de avaliar suas capacidades ao lidar com um problema diferente do que classificação de dados.

4.1 Seleção de Caraterísticas

Com o objetivo de reduzir a dimensionalidade dos dados, métodos de seleção de características (do inglês: *Feature Selection*) vem sendo utilizados em muitos problemas que envolvem dados de alta dimensionalidade, como em análise de dados de *microarray*, classificação de texto e imagens.

As técnicas de seleção de características podem ser categorizados em três grupos principais (ver Figura 4.1):

filters: estes calculam a relevância das características observando suas propriedades, produzindo um subconjunto com as melhores características encontradas.

wrappers: realizam um processo iterativo, produzindo um subconjunto de características que é avaliado por um classificador e seu resultado é retornado para o *wrapper* até que seja encontrado um subconjunto ótimo.

embedded: possuem um processo semelhante aos *wrappers*, no entanto são algoritmos mais rápidos pois estas técnicas de filtragens são construídas dentro dos classificadores.

Os ***filters*** usualmente calculam a relevância de cada característica utilizando uma métrica, e um limiar mínimo, remove do conjunto de dados as características que estão abaixo desse limiar. Os filtros são técnicas computacionalmente simples, rápidas. Além disso, são independentes de um classificador, sendo consequentemente os mais utilizados.

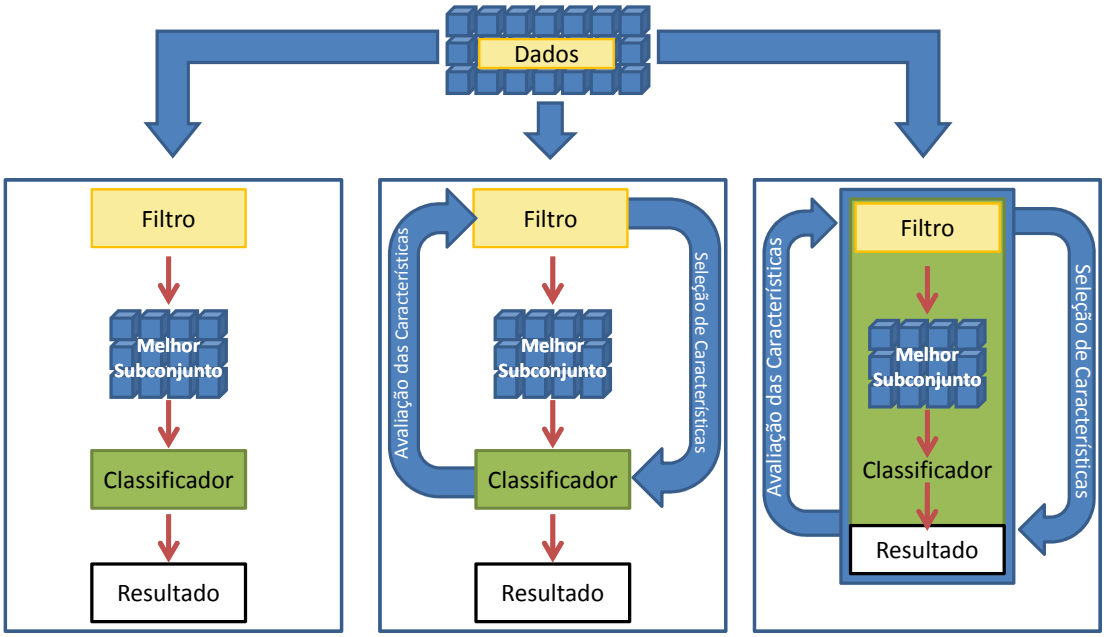


Figura 4.1 Esquema da seleção de características realizado pelas técnicas: *filters*, *wrappers* e *embedded*.

As técnicas ***wrappers***, em sua maioria, consideram as interações entre as características. Além disso, o procedimento para avaliar e construir o subconjunto de características, é feito através de treinamento e teste em um classificador específico. Por isso, os *wrappers* requerem recursos computacionais maiores que os filtros, especialmente quando o espaço de características cresce. O procedimento de treino e teste, também tem um agravante, pois pode levar ao *overfitting*, fazendo com que o classificador seja muito específico para o problema treinado.

O último grupo de técnicas, são os ***embedded***, estes são similares aos *wrappers* mas a diferença entre eles está relacionada ao classificador. Nos *embedded*, os filtros são construídos dentro dos classificadores, esta é uma vantagem por ser computacionalmente menos custoso que os *wrappers*. No entanto, por conta dessa características, são filtros dependentes do classificador (Saeys *et al.*, 2007) e por isso, não é possível avalia-los individualmente.

A alta dimensionalidade, é um dos desafios apontados na análise de dados genômicos, assim nesta tese, é feito um estudo sobre diferentes filtros de seleção de características. Os filtros foram selecionados por serem capazes de realizar remoção de informações irrelevantes em dados de alta dimensionalidade, serem computacionalmente eficientes e de fácil implementação. Sendo estudados com a finalidade de serem aplicados em um processo de pré-processamento dos dados.

4.1.1 Filtros

Baseados em sua característica principal, os filtros podem ser classificados em dois grupos: filtros **univariados** e **multivariados**. A diferença essencial entre esses dois grupos se dá pelo fato de que os filtros univariados não consideram as interações existentes entre as características, enquanto que os multivariados o fazem.

Filtros Univariados

As técnicas de filtragem univariadas podem ser divididas em duas classes: **Filtros Paramétricos** e **Filtros não Paramétricos** ou *model-free*. Os filtros paramétricos assumem uma distribuição de probabilidade para os dados, ou então, são baseados em estimativas de uma distribuição de parâmetros desconhecida.

Entre as **técnicas paramétricas**, estão incluídos o *t*-test, χ^2 e *Fisher Score*. Embora sejam técnicas amplamente utilizadas, a incerteza sobre a verdadeira distribuição dos conjuntos de dados genéticos, e as dificuldades para validar as suposições de distribuição fazem com que métodos não-paramétricos ou *model-free* sejam uma alternativa atraente para esses estudos (Saeys *et al.*, 2007).

Técnicas não paramétricas como *Information Gain* e *Gini Index* são referenciados como métodos livres de distribuição. Em outras palavras, estas técnicas não assumem um modelo de distribuição para os dados. Vale ressaltar, que estas técnicas baseiam-se em um menor número de suposições sobre os dados e também possuem uma menor dependência em testes de hipóteses do que os filtros paramétricos, assim são consideradas mais aplicáveis e robustos.

No entanto, são testes com menos poder estatístico, ou seja, podem existir casos em que testes paramétricos sejam mais adequados. Nesses casos, para se ter conclusões com o mesmo grau de confiança com o uso de técnicas não paramétricas, uma amostra de dados maior pode ser necessária.

Tabela 4.1 Técnicas de Filtragem de Seleção de Características

Univariado		**Multivariado**
Paramétrico	**Model-Free**	
t-test	Gini Index	ReliefF
Chi-Square Score	Information Gain	CFS
Fisher Score		FCBF
		INTERACT

Na Tabela 4.1, é apresentado um resumo das diferentes técnicas de filtragem discutidas, sendo separadas de acordo com suas características.

Filtros Multivariados

As técnicas de filtragem multivariadas podem ser classificadas em dois grupos, de acordo com o tipo de saída retornado. Assim, existem os filtros, os quais, a saída é um *ranking* com as características mais relevantes no topo da lista, e os filtros que retornam um subconjunto ótimo, no qual as características mais relevantes são agrupadas.

No primeiro caso, a filtragem ocorre quando as características após ordenadas não alcançam um limiar mínimo, sendo então removidas. No segundo caso, o filtro retorna um subconjunto das características relevantes identificadas no conjunto de dados, sendo as demais características todas eliminadas.

ReliefF

O filtro Relief desenvolvido por Kira e Rendell (1992) considera a dependência entre os atributos e estima a qualidade das características. Para isso, o filtro faz uso de um algoritmo de vizinhança, apresentado no Algoritmo 1, que verifica a relevância de uma característica a partir de um vetor de pesos, calculado através de uma métrica entre a amostra analisada e o vizinho da mesma classe e de classe diferente da amostra.

Algoritmo 1: Cálculo do Vetor de Pesos pelo Relief

Entrada: Conjunto de dados X

Saída: Peso dos Atributos (W) de -1 (pior) a +1 (melhor)

```
1 inicializa o vetor de pesos W = 0;
2 para i ← 1 até m faça
3     Randomicamente seleciona um amosta R_i;
4     Encontra o vizinho da mesma classe (H) e o vizinho da outra classe (M);
5     para todo W faça
6         W_i = W_i − diff(R_i,H_i)^2 + diff(R_i,M_i)^2;
7     fim
8     Relevancia_i = (1/m) × W_i
9 fim
```

Onde, W é o vetor de pesos é atualizado usando o vizinho mais próximo da classe (H) e da outra classe (M). A função a diff() retorna 0, se o atributo e o vizinho forem iguais ou 1, se forem diferentes.

Uma versão derivada do Relief, foi desenvolvida por Kononenko (1994) sendo denominada de ReliefF. Este filtro também captura as interações entre os atributos, no entanto, utiliza os n vizinhos mais próximos da amostra em vez de apenas o mais próximo como faz o Relief. Esta modificação, reduz a influência de dados redundantes e ruído nos dados. Além disso, o ReliefF também trata problemas como dados faltantes e lida com problemas com mais de duas classes.

Utilizando o vetor de atributos, é possível calcular a relevância para cada característica, e assim, criar um *ranking* com as melhores características no topo da lista. A remoção das características irrelevantes, se dá através da definição de um limite inferior, sendo valores de relevância menores que este limite descartados.

Correlation-Based Feature Selection

O *Correlation-Based Feature Selection* (CFS), desenvolvido por Hall (1999) é um filtro que realiza a inferência de um subconjunto de características relevantes. Para que este filtro seja capaz de encontrar um subconjunto ótimo (ou próximo do ótimo), o filtro realiza uma busca heurística denominada *forward best first search*. Este tipo de busca, também denominada de busca gulosa, é capaz de encontrar um subconjunto dos dados em um tempo razoável. No entanto, este tipo de busca não há garantias de encontrar o melhor subconjunto possível.

Para avaliar a correlação das características com a classe e a interação entre as características, o CFS estima esta associação com a medida *Symmetrical uncertainty*. Esta medida, utiliza a entropia individual e a entropia condicional para obter o *Information Gain*,

$$\begin{aligned} InfoGain &= H(X) - H(X|Y) \\ H(X) &= -\sum_{x \in X} p(x) log_2 p(x) \quad e \quad H(X|Y) = -\sum_{x,y} p(x,y) log \frac{p(x,y)}{p(y)}. \end{aligned} \tag{4.1}$$

onde, $p(x)$ é a probabilidade de x, $H(X)$ a entropia individual e $H(X|Y)$ a entropia conjunta das *features*.

Com a normalização do *Information Gain* é obtido a métrica *Symmetrical uncertainty*,

$$SymUnc = 2.0 \times \frac{InfoGain}{H(Y) + H(X)}. \tag{4.2}$$

onde, H(Y) e H(X) são a entropia individual das características descritas na equação (4.1).

Em seguida, o CFS realiza uma avaliação dos subconjuntos de características relevantes, realizando uma busca até encontrar o menor número de características capazes de separar as classes de dados. Para isso, utiliza informações oriundas da característica sozinha para predizer sua classe e a correlação entre outras características através da métrica MeritS,

$$MeritS = \frac{Kr_{cf}}{\sqrt{K + K(K-1)r_{ff}}}, \tag{4.3}$$

onde S é o subconjunto das características, K a quantidade de características em S, r_{cf} é a média da relação *feature-classe*, e r_{ff} é a média da correlação *feature-feature*. O denominador indica o quão preditivo é um grupo de características; e o numerador indica a redundância existente entre as características (Hall, 1999).

Fast Correlation-based Feature Selection

O filtro *Fast Correlation-based Feature Selection* (FCBF) proposto por Yu e Liu (2004), utiliza a métrica *Symmetrical uncertainty* para avaliar as características relevantes.

O algoritmo do FCBF, é apresentado em Algoritmo 2, sendo dividido em duas etapas.

Na primeira etapa, realiza o *ranking* das características utilizando a equação (4.2) para estimar a correlação entre as características e a classe. Os valores obtidos com esta métrica, são utilizados como um limiar, para selecionar apenas as características altamente relevantes. Na segunda etapa, o subconjunto das características selecionadas na primeira etapa, é avaliado com o objetivo de remover as redundâncias existentes nesse subgrupo.

Algoritmo 2: Fast Correlation-Based Filter.

Entrada: Conjunto de dados X
Saída: Melhor subconjunto de features sem redundância

```
1  S_listF = NULL; //Lista das features selecionadas;
2  S_listR = NULL; //Lista dos Ranks;
3  para i ← 1 até N faça
4      Calcula SU_i,c de uma amosta r_i;
5      se SU_ri,c > γ então
6          S_listF ← r_i ;
7          S_listR ← SU_ri,c;
8      fim
9  fim
10 Ordena S_listF decrescentemente pelo valor da S_listR;
11 Ordena S_listR decrescentemente;
12 enquanto r_i ← GetFeature(R,i) ≠ NULL faça
13     j ← i + 1;
14     enquanto r_j ← GetFeature(R,j) ≠ NULL faça
15         j ← j + 1;
16         f ← CalculaCorrelacao(r_i,r_j);
17         se f < S_listRj então
18             Remove a feature r_j de S_listF;
19         fim
20     fim
21     i ← i + 1
22 fim
23 retorna S_listF;
```

onde, γ é um *threshold* definido pelo usuário para remoção das características irrelevantes.

INTERACT

O método INTERACT, desenvolvido por Zhao e Liu (1991), é um filtro que emprega o uso de uma métrica denominada $c-contribution$ (*Consistency Contribution*) que avalia a importância da característica no conjunto de dados. A métrica $c-contribution$ é um indicador de significância para remoção de uma característica, ou seja, quando o valor de $c-contribution$ de

uma característica tende a zero, significa que é uma característica irrelevante, sendo descartada.

O algoritmo do INTERACT, apresentado em Algoritmo 3, também é dividido em duas etapas principais. Na primeira etapa, é realizado um *ranking* das características utilizando a equação (4.2) para estimar a relevância entre a amostra e a classe, assim como é feito no FCBF e CFS. Na segunda etapa, as características são avaliadas uma por uma utilizando a métrica $c-contribution$, começando das características menos relevantes para as mais relevantes, de acordo com o método (*backward elimination*).

Algoritmo 3: Seleção de subconjunto ótimo pelo INTERACT

Entrada: Conjunto de dados X

Saída: Melhor subconjunto de features sem redundância

```
1  S_list = NULL;
2  para i ← 1 até N faça
3      Calcula SU_{i,c} Randomicamente seleciona uma amosta R_i;
4      Adiciona R_i a uma lista S_list;
5  fim
6  Ordena S_list em ordem decrescente pelo valor de SU_{i,c};
7  counter = N;
8  repita
9      F = S_list[counter];
10     p = c-contribution(F, S_list);
11     se p ≤ δ então
12         remove F de S_list;
13     fim
14     counter = counter - 1;
15 até counter = 0;
16 S_best = S_list;
17 retorna S_best;
```

Conclusão

Vários desses filtros, já foram utilizados em diferentes problemas biológicos, como o CFS, aplicado em dados de câncer de microarray por Wang *et al.* (2005), combinando o uso desse filtro com outros classificadores. Outros filtros, como ReliefF, Information Gain, e o teste de χ^2, foram utilizados com sucesso em Wang *et al.* (2004), sendo produzida uma ferramenta para identificar gene relevantes. Esses trabalhos estimularam os testes desses filtros nesta tese, sendo aplicado em uma etapa de pré-processamento dos dados. Na seção seguinte, serão apresentadas diferentes ferramentas propostas para lidar com a inferência das interações entre os SNPs, verificando suas capacidades e limitações.

4.2 Ferramentas de Inferência de Epistasia

As primeiras formas de avaliação dos SNPs utilizavam **métodos estatísticos tradicionais**, como o teste de χ^2, teste exato de Fisher e *t*-test. Estes métodos realizam testes de hipótese sobre cada um dos SNPs, verificando sua relevância para determinar o fenótipo.

A análise da relevância de um único nucleotídeo é efetivo em descobrir SNPs fortemente relacionados em aumentar o risco de uma determinada doença. No entanto, essas novas descobertas proporcionadas por este tipo de análise, explicam apenas uma pequena fração (menos de 20%) da herdabilidade da doença (Chen *et al.*, 2009).

Isso acontece pois, a maioria das doenças possuem mecanismos complexos de regulação celular, que são codificados no genoma humano. Assim, a herdabilidade também é explicada pelo efeito da interação (**epistasia**) de vários SNPs e outros fatores. Sabendo-se disso, um grande esforço é feito nesta área, para avaliar as interações entre SNPs em doenças multifatoriais.

Para tanto, algumas considerações estatísticas precisam ser levadas em conta, quando estamos tratando de interações entre SNPs. O tamanho do conjunto de dados e o número de interações a serem identificadas, por exemplo, afetam diretamente o desempenho das abordagens computacionais, tornando o processo da inferência muito mais difícil devido ao maior número de avaliações a serem executadas.

Diante disso, várias abordagens computacionais vêm buscando detectar de forma eficiente (computacional e biologicamente), as interações entre SNPs. Para isso, utilizam diferentes formas de busca, modernas estruturas tecnológicas, como o uso de clusters de computadores equipados com unidades de processamento gráfico (GPU), paralelização e computação bit a bit (*bitwise*) (Wei *et al.*, 2014).

Nesta tese, foram selecionadas diferentes ferramentas, baseando sua escolha, nas suas diferentes estratégias de busca utilizadas. De acordo com Yang *et al.* (2009), as estratégias de busca dividem as ferramentas em: métodos de busca exaustiva, métodos de busca gulosa e métodos de busca estocástica.

Seguindo esta divisão, foram identificadas diversas abordagens, sendo escolhidas cinco ferramentas com base em seus desempenhos avaliados em estudos anteriores (Araujo e Guimaraes, 2011), (Araujo *et al.*, 2011) e (Shang *et al.*, 2011) ao lidar com o problema da inferência das interações epistáticas, popularidade da ferramenta medida através do número de citações e a disponibilidade do código fonte ou da ferramenta para testes. Nas próximas seções, as ferramentas são descritas detalhadamente.

4.2.1 Métodos de Busca Exaustiva

Os métodos de busca exaustiva, são uma forma simples de avaliar todas as interações entre SNPs. Essas abordagens têm demonstrado excelentes resultados, no entanto devido a busca exaustiva limitam-se a utilizar conjuntos de dados com até 1000 SNPs ou avaliar um número

limitado de interações (duas ou no máximo, três interações). No entanto, sua utilização em conjuntos muito grandes, como dados de GWAS, é computacionalmente inviável quando se deseja analisar interações de alta ordem entre SNPs (Yang *et al.*, 2009).

Multifactor Dimensionality Reduction (MDR)

O *Multifactor Dimensionality Reduction* (MDR) proposto por Ritchie *et al.* (2001), é uma abordagem de mineração de dados não-paramétrica, que não assume nenhum modelo genético para detectar e caracterizar as interações entre as variáveis genéticas e ambientais.

O MDR foi proposto para detectar interações na ausência de efeitos marginais (efeitos de um único polimorfismo, quando não há interação) e, portanto, complementa abordagens estatísticas, como regressão logística e métodos de aprendizagem de máquina, como florestas randômicas e redes neurais (Moore *et al.*, 2010).

Muitas modificações e extensões têm sido propostas ao MDR, mas a ideia principal da abordagem é baseada na redução do espaço de representação dos dados. Essa redução, torna mais fácil a utilização de métricas para detecção das interações entre as variáveis. A ferramenta foi implementada e descrita por Ritchie *et al.* (2003), é uma ferramenta *open-source*, escrita na linguagem Java, que é capaz de tratar tanto de dados do tipo, caso e controle como também, dados *family-based*.

Observando o esquema na Figura 4.2, pode-se observar que o MDR pode ser dividido em 4 etapas: validação cruzada, interação, classificação e predição. Na validação cruzada, o conjunto de dados é dividido em 10 subconjuntos, dos quais 9 subconjuntos são selecionados para formar os dados de treinamento, e 1 para o teste. O processo é repetido 10 vezes permutando todos os subconjuntos. A cada rodada da validação cruzada todas as combinações entre SNPs são testadas.

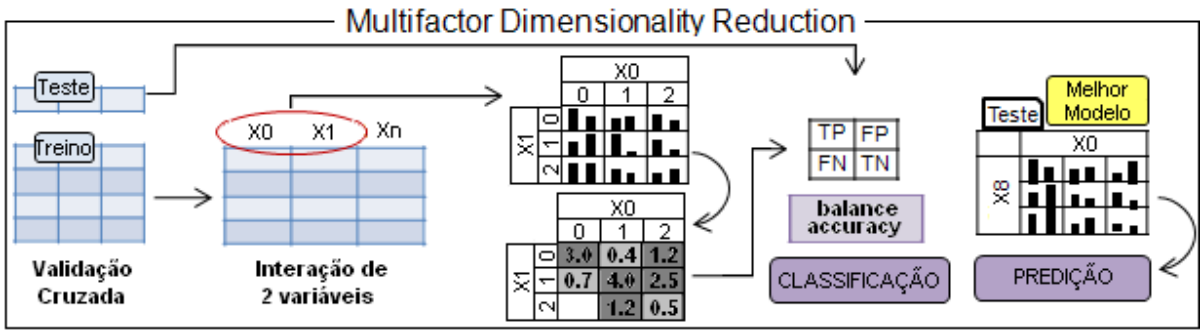


Figura 4.2 Esquema do processamento da análise das interações no MDR. Em uma validação cruzada, o conjunto de dados é divido em treinamento e teste, utilizado para classificação e predição do método.

Após a separação dos dados, utilizando o conjunto de treinamento, n variáveis genéticas são selecionadas para verificar suas interações. A Figura 4.2 é apresentado um exemplo de interação entre um par de SNPs X_0 e X_1 que são representados em uma tabela genótipo/fenótipo *N-dimensional*. Como cada SNP X_0 e X_1 é representado por três genótipos (homozigoto dominante,

heterozigoto e homozigoto recessivo) teremos 9 classes de combinações de genótipos.

Utilizando a tabela genótipo/fenótipo é calculada a razão entre o número de casos e controles de cada classe, sendo a classe rotulada como de "alto risco" (em cinza escuro), se a razão exceder um limite $\geq = 1$, ou de "baixo risco" (em cinza claro), se o limite não é ultrapassado. Dessa forma, o espaço passa a ter apenas duas classes "baixo risco" e "alto risco".

Após o cálculo da tabela de contingência, utilizando o conjunto de treinamento, os rótulos de alto e baixo risco obtidos são comparados com os genótipos do conjunto de teste. A cada rodada da validação cruzada, são calculados os valores dos verdadeiros positivos, verdadeiros negativos, falso positivos e falsos negativos. Sendo calculada o *balanced accuracy* da interação:

$$BalancedAccuracy = \frac{\frac{TP}{TP+FN} + \frac{TN}{FP+TN}}{2}, \tag{4.4}$$

onde, TP são os verdadeiros positivos, quando o paciente é caso e seu genótipo está numa célula de alto risco, TN são os verdadeiros negativos quando o paciente é controle e o genótipo em uma célula de baixo risco, FP são os falsos positivos quando o paciente é caso e está numa célula de baixo risco e FN são os falsos negativos quando o paciente é controle e está numa célula de alto risco.

Utilizando a equação (4.4) são calculadas todas as n combinações possíveis do conjunto de dados durante o ciclo da validação cruzada, utilizando os conjuntos de treinamento e teste. Ao final da validação cruzada, a interação que apresentou o maior número de vezes com maior valor do *balanced accuracy*, é definida como o modelo que possui o menor número de indivíduos erroneamente classificados, sendo selecionado dentre os demais modelos para fins preditivos de uma doença.

Esta abordagem é amplamente aceita pela comunidade científica e mesmo diante de seu caráter exaustivo é bastante utilizada e referenciada por vários grupos de pesquisa (Ritchie *et al.*, 2001; Hahn *et al.*, 2003; Ritchie e Motsinger, 2005; He *et al.*, 2009), devido ao excelente poder de detecção das interações entre SNPs. Uma outra vantagem da ferramenta, foi proposta por (Hahn *et al.*, 2003), tornando-a capaz de lidar com dados faltantes, sendo necessário que o dado faltante seja representado por um valor diferente dos genótipos (0, 1 e 2).

O método também é capaz de analisar dados com mais de duas interações, no entanto, é preciso informar previamente o número de interações que se deseja analisar. Tornando essa uma desvantagem, especialmente ao lidar com conjuntos de dados reais, os quais não se sabe ao certo a quantidade de SNPs que estão interagindo. Um outro limitante está em sua análise combinatorial exaustiva, o que impossibilita o estudo de conjunto de com alta dimensionalidade e inferência de interações maiores que 2, já que o custo computacional cresce de forma combinatorial com o número de SNPs interagindo. Uma versão posterior (Bush *et al.*, 2006) torna a abordagem paralelizável ampliando sua capacidade de processamento para conjuntos de dados maiores, mas a limitação referente ao número de interações não é suplantada.

Boolean Operation based Screening and Testing (BOOST)

O *Boolean Operation based Screening and Testing* (BOOST) é uma ferramenta desenvolvida por Wan *et al.* (2010a). O BOOST apresenta uma estratégia de busca dividida em dois estágios: *Screening* e *Testing*. Uma etapa de pré-processamento dos dados é realizada, modificando a representação dos dados para uma representação Booleana, tornando o método muito eficiente, pois são realizadas apenas operações Booleanas.

No estágio *Screening* é utilizado um método não-iterativo para aproximar a estatística da razão de verossimilhança na avaliação de todos os pares de SNPs e selecionar aqueles que superam um limiar, eliminando as interações não significativas. No estágio *Testing* os pares de SNPs significativos são submetidos a um teste clássico de razão de verossimilhança para medir os efeitos da interação.

Para a detecção das interações entre SNPs, a abordagem BOOST considera a interação como sendo baseada em modelos de regressão logística. A representação do modelo de regressão logística que considera apenas os efeitos principais (LM) é definida na equação:

$$log\frac{P(Y=1|X_p=i,X_q=j)}{P(Y=2|X_p=i,X_q=j)} = \beta_0 + \beta_i^{X_p} + \beta_j^{X_q} \tag{4.5}$$

onde Y é a variável resposta, ou seja a classe e X_p e X_q são os SNPs da interação testados.

Já o modelo de regressão logística completo, considera tanto as interações como também os efeitos principais (LF) é definido na equação:

$$log\frac{P(Y=1|X_p=i,X_q=j)}{P(Y=2|X_p=i,X_q=j)} = \beta_0 + \beta_i^{X_p} + \beta_j^{X_q} + \beta_{ij}^{XqXq} \tag{4.6}$$

Sabendo que LM e LF são o log-verossimilhança do modelo de efeito principal e do modelo completo, respectivamente e de acordo com o teste da razão de verossimilhança, os efeitos da interação são definidos pela diferença entre os logs-verossimilhança desses dois modelos, ou seja, $2(\hat{L}_F - \hat{L}_M)$.

No entanto, utilizar regressão logística para calcular todos os pares de interação em estudo de *Genome-Wide* é inviável, de forma que percebendo a equivalência entre o modelo de regressão logística e seu modelo log-linear correspondente, são utilizados os modelos de associação homogênea conforme a equação:

$$log_{\mu ijk} = \lambda + \lambda_i^X p + \lambda_j^X q + \lambda_k^Y + \lambda_{ij}^X pXq + \lambda_{ik}^X pY + \lambda_{jk}^X qY \tag{4.7}$$

e de associação saturada:

$$log\frac{P(Y=1|X_p=i,X_q=j)}{P(Y=2|X_p=i,X_q=j)} = \beta_0 + \beta_i^X p + \beta_j^X p + \beta_{ij}^{XqXq} \tag{4.8}$$

O teste de razão de verossimilhança é a diferença entre o log-verossimilhança e modelos log-linear homogêneo (LH) e saturado (LS) definido na equação:

$$\hat{L_S} - \hat{L_H} = \sum_{i,j,k} \left[n_{ijk} log \frac{n_{ijk}}{\hat{\mu}^H_{ijk}} - n_{ijk} + \hat{\mu}^H_{ijk} \right] \quad (4.9)$$

Para testar o efeito de interação entre dois SNPs (X_p e X_q) e a classe Y utilizando os modelos log-linear, uma tabela de contingência (Tabela 4.2) é construída com três variáveis i, j e k, a partir de uma representação Booleana dos dados de genótipos. Esta representação Booleana dos dados, torna o método mais eficiente por tratar apenas dados do tipo 0 e 1. Permitindo o uso de apenas operações lógicas, que são realizadas de forma rápida pelo processador.

Tabela 4.2 Tabela de contingência Booleana, onde, Y = 0 representa os casos e Y = 1 os controles.

Y = 0	X_q **= 0**	X_q **= 1**	X_q **= 2**	**Y = 1**	X_q **= 0**	X_q **= 1**	X_q **= 2**
$X_p = 0$	n_{111}	n_{121}	n_{131}	$X_p = 0$	n_{112}	n_{122}	n_{132}
$X_p = 1$	n_{211}	n_{221}	n_{231}	$X_p = 1$	n_{212}	n_{222}	n_{232}
$X_p = 2$	n_{311}	n_{321}	n_{331}	$X_p = 2$	n_{312}	n_{322}	n_{332}

No método BOOST, há uma transformação Booleana dos dados. Cada SNP que antes era representando por 0, 1 e 2, passa a ser representando apenas por 0 e 1. Para solucionar a separação entre homozigoto dominante (0), homozigoto recessivo (2) e heterozigoto (1). Um SNP passa a ser representado em 3 linhas, onde cada linha é um genótipo específico (0, 1 e 2) e cada coluna representa a separação da classe de casos e controles, como exibido na Tabela 4.3.

Tabela 4.3 Representação da matriz Booleana construída no método do BOOST. X0, X1 e X2, são SNPs representados em três linhas da matriz, cada linha corresponde a um genótipo AA = 0, Aa = 1 e aa = 2, separados em duas colunas, sendo formada por 16 (P_n) pacientes, 8 casos e 8 controles.

SNPs	**Pacientes controles**								**Pacientes casos**							
	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}
$X_0 = 0$	1	0	0	0	0	0	1	1	0	0	0	0	1	0	1	1
$X_0 = 1$	0	0	1	1	0	0	0	0	1	1	0	0	0	1	0	0
$X_0 = 2$	0	1	0	0	1	1	0	0	0	0	1	1	0	0	0	0
$X_1 = 0$	0	0	0	0	1	1	1	0	0	0	1	1	1	0	1	0
$X_1 = 1$	0	1	1	1	0	0	0	1	0	0	0	0	0	1	0	1
$X_1 = 2$	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
$X_2 = 0$	1	1	0	0	0	0	0	1	1	0	0	0	1	1	1	0
$X_2 = 1$	0	0	0	0	1	1	1	0	0	1	1	1	0	0	0	0
$X_2 = 2$	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1

O método BOOST, esquematizado na Figura 4.3, transforma os dados em Booleanos, em seguida constrói a tabela de contingência e a utiliza durante a fase de *Screening* para avaliar todas as interações aos pares entre os SNPs.

O método, por ser implementado de forma eficiente, apresenta vantagens frente a métodos exaustivos. Além disso, é um método que por dividir a busca em duas etapas, é conceitualmente simples e fácil de implementar. No entanto, a ferramenta é limitada por não realizar inferência

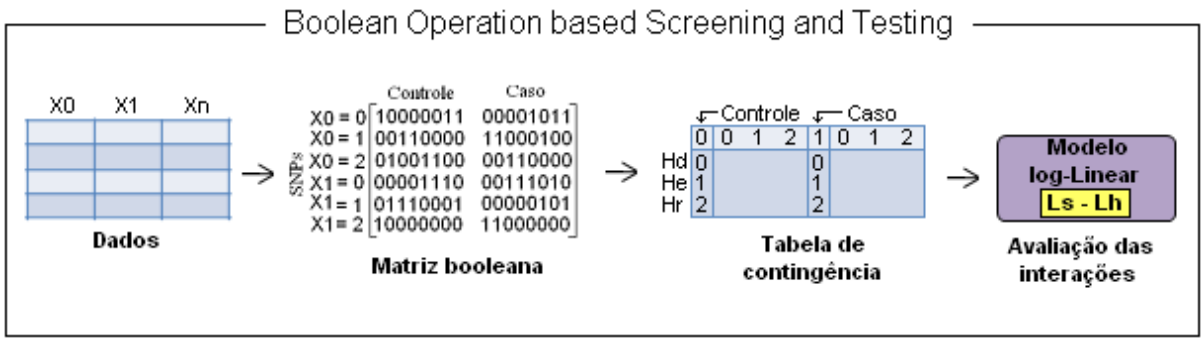


Figura 4.3 Esquema do processamento da análise das interações no BOOST. Os dados de entrada são armazenados em uma nova matriz Booleana que é utilizada para construção de uma tabela de contingência para avaliar as interações entre os SNPs.

de interações de alta ordem e por isso, pode não detectar todas as interações entre SNPs devido a sua busca incompleta.

Testes comparando o desempenho do BOOST com o método PLINK (Purcell *et al.*, 2007) foram realizados por Wan *et al.* (2010a) com populações de 800 e 1600 casos e controles, demonstrando o melhor desempenho do BOOST. Apesar disso, algumas deficiências não foram solucionadas, como o tratamento de dados faltantes e o desbalanceamento de casos e controles.

4.2.2 Métodos de Busca Gulosa

Abordagens computacionais gulosas, são aquelas que buscam uma solução ótima, tomando decisões em direção a um novo ótimo local, em cada passo de execução. Com isso, espera-se, que ao final da busca seja alcançado o ótimo global, sem que para isso, seja necessário analisar todas as situações possíveis.

É importante ressaltar, que não há garantias de que o ótimo global será alcançado. O algoritmo guloso pode ser utilizado como um algoritmo de seleção, para priorizar as opções dentro de uma pesquisa, ou como algoritmo *branch and bound* para finalizar o crescimento de um ramo ou na divisão dos nós ao encontrar uma solução ótima.

No problema das interações entre SNPs, os métodos de busca gulosa executam uma filtragem baseada nos SNPs ou interações não-epistáticas de ordem inferior para filtrar SNPs que não exibem nenhum efeito principal. Em geral, os métodos que realizam esse tipo de busca utilizam dois ou mais passos para avaliar as interações.

Uma estratégia como o CART (*classification and regression tree*) seleciona um SNP que melhore o valor de uma métrica, por exemplo, *Gini Index* ou Entropia, e em seguida, segue no crescimento da árvore. No entanto, o sucesso da aplicação dessa estratégia depende da natureza das interações presentes no conjunto de dados, ou seja, é necessário que os SNPs tenham interações puramente epistáticas, caso contrário, são susceptíveis de serem descartados da interação.

Apesar disso, métodos de busca gulosa permitem uma inferência das interações em um

conjunto de dados muito maior, pois diferentemente das abordagens que utilizam uma busca exaustiva, o fato de não analisar todo o conjunto de dados, permite sua utilização em dados de GWAS.

SNPRuler

O SNPRuler (Wan *et al.*, 2010b) é um algoritmo *branch and bound* de busca não exaustiva, sendo caracterizado por buscar uma solução ótima enumerando alguns candidatos como solução e descartando o restante através de um limite superior e inferior.

A ferramenta tem como objetivo de encontrar as interações epistáticas e para isso é aplicada uma aprendizagem de regras preditivas (*predictive rule learning*). As regras preditivas, tem como objetivo descrever a relação entre os SNPs e as classes caso e controle, e assim pode facilitar na detecção das interações mais significativas.

Estas regras preditivas, são propostas considerando que as interações contém algum tipo de regra que pode ser aprendida. Ao se considerar isso, realizar uma busca destas regras é mais fácil e menos custosa, do que avaliar cada uma das interações entre SNPs.

No entanto, o método não dá garantias de que a regra preditiva seja capaz de detectar as interações epistáticas. Assim é proposta inicialmente, uma aprendizagem das regras preditivas para selecionar regras com boa confiabilidade, e em seguida é utilizado testes estatísticos para identificar as interações selecionadas pelas regras.

Observando a Figura 4.4, podemos entender basicamente o funcionamento da abordagem.

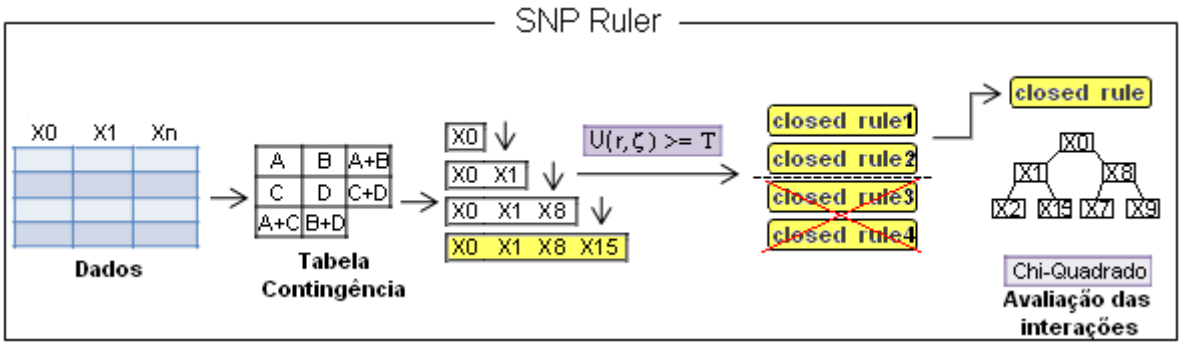


Figura 4.4 Esquema do processamento da análise das interações no SNPRuler. Uma tabela de contingência é criada e utilizada para construção das regras fechadas, em seguida as regras são avaliadas pela estatística χ^2.

A aprendizagem das regras, é feita com o objetivo de encontrar todas as regras preditivas do tipo fechada $((r,\zeta)_i)$, ou seja, regras que não podem ser melhoradas pela adição de novos SNPs, sendo r é o conjunto de vários SNPs e ζ a classe (1 para caso e 0 para controle).

Os SNPs s são formados por um par de variáveis (i,v), onde i é o índice e v um valor que pode ser 0, 1 ou 2 representando os genótipos homozigoto dominante, heterozigoto ou homozigoto recessivo respectivamente.

Com o objetivo de encontrar as melhores regras preditivas, uma medida de relevância

$U(.)$ é utilizada para ordenar as regras que contenham interações verdadeiras. A partir da medida da regra preditiva $U(.)$, um limite superior é definido para evitar a expansão desnecessária da regra, evitando uma busca exaustiva das interações.

O cálculo da medida da regra preditiva $U((r,\zeta)_i)$

$$U(r,\zeta) = \frac{(R-\delta^2)}{(1+\delta)(\gamma-\delta-1)} \tag{4.10}$$

é feito utilizando uma tabela de contingência analisada através da *Rule Utility*, desenvolvida a partir da estatística χ^2. Onde $\delta = \frac{b}{a}$, $\gamma = \frac{a+b+c+d}{a}$ e $R = \frac{b+d}{a+c}$. As variáveis a, b, c e d são obtidas através de uma tabela de contingência (Tabela 4.4) construída a partir do conjunto de dados para uma dada regra (r,ζ)

As regras preditivas são ditas ruins e eliminadas da análise quando a medida de relevância $U((r,\zeta)_i) \leq T$, onde T é um limite inferior definido pelo usuário.

Tabela 4.4 Tabela de contingência para um dada regra (r,ζ).

	$\zeta = 0$	$\zeta \neq 0$	**Total**
r	a	b	$a+b$
$\neg r$	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Para a construção de uma regra preditiva (r,ζ), o algoritmo SNPRuler inicia basicamente com um único SNP (S_j) e genótipo g_i. Cria e expande uma regra preditiva adicionando outros SNPs ($S_j = g_i$) até que esta regra se torne uma regra fechada, fazendo este processo para todos os SNPs. A fim de evitar expansões desnecessárias das regras, o algoritmo utiliza o limite superior $U_{max}(r \bigcap (S_j = g_i), \zeta)$, da medida da regra $U(.)$, definido na equação:

$$U_{max} = \frac{Rm - (b - min(b,d'))^2}{(m + (b - min(b,d')))(\gamma m - (b - min(b,d') - m))} \tag{4.11}$$

onde, R e γ são definidos na equação (4.10), $m = min(a,a')$, sendo a' e d' variáveis do novo SNP adicionado à regra, obtidas a partir de uma nova tabela de contingência como descrita na Tabela 4.4.

Após obtidas as melhores regras utilizando a medida $U(r,\zeta)$, o algoritmo então constrói uma árvore de busca para cada regra selecionada, onde cada protótipo representa um SNP e cada ramo que liga os protótipos representa uma possível interação que é avaliada utilizando um método de busca em profundidade (*depth-first transversal*) que gera e avalia as possíveis interações utilizando a estatística χ^2 ajustada pelo uso da correção de Bonferroni (Abdi, 2007). Ao final o algoritmo exibe em sua saída uma lista de interações ordenadas através da estatística χ^2.

Apesar de ser um método de aplicação mais ampla para conjunto de dados do GWAS, o fato de realizar uma busca das interações de forma gulosa não garante que o método encontre as

interações relevantes presentes no conjunto de dados. A seguir, serão apresentados os métodos de busca estocásticas.

4.2.3 Métodos de Busca Estocástica

Os algoritmos de busca estocástica realizam uma investigação aleatória do espaço de busca. Alguns começam com um modelo composto por um conjunto aleatório de SNPs e tentam melhorar a sua precisão de classificação, enquanto outros usam métodos caros em um subconjunto pequeno dos dados previamente selecionados aleatoriamente. Apesar de eficientes, esses métodos muitas vezes dependem do acaso, para selecionar as interações de SNPs que exercem influência sobre uma doença. À medida que os conjuntos de dados crescem em número de SNPs, as chances de encontrar os dados corretos diminuem devido ao crescimento do espaço de busca.

Vários métodos utilizam esta abordagem, como *Genetic Evolution Optimized Neural Network* (GENN) que é uma abordagem de aprendizagem de máquina desenvolvida por Motsinger-Reif *et al.* (2008) que utiliza evolução gramatical para otimizar uma rede neural na detecção das interações entre SNPs. Outras abordagens que também utilizam busca estocástica para analisar as interações entre os SNPs são a *Bayesian Epistasis Association Mapping* (Zhang e Liu, 2007) (BEAM) e a abordagem *Co-Information Based Method for Detecting and Visualizing n-order Epistatic Interactions* (CINOEDV) (Shang *et al.*, 2016), uma das ferramentas foco dessa dissertação que serão melhor descritas nas seções seguintes.

Bayesian Epistasis Association Mapping (BEAM)

A ferramenta *Bayesian Epistasis Association Mapping* (BEAM) calcula, através de um modelo estatístico, tanto o efeito de um único SNP, como também das interações entre os SNPs associados a doenças. Para isso, utiliza o método de Monte Carlo via cadeias de Markov (MCMC) para interrogar cada um dos SNPs. A ferramenta considera em suas análises a utilização de metodologias de diferentes escolas da estatística para avaliar as interações entre SNPs, ao utilizar inferência Bayesiana e o teste estatístico B, de forma a complementar as informações obtidas.

Ao final, a abordagem exibe a probabilidade a posteriori[1] de cada SNP; e a probabilidade de cada interação entre SNPs estar associada com a doença, assim como também avalia a importância de cada SNP ou da interação utilizando um teste de hipótese através da estatística B. A ferramenta desenvolvida por Zhang e Liu (2007), programada em linguagem C, está disponível em http://www.fas.harvard.edu/ junliu/BEAM/ para Windows e Linux incluindo seu código-fonte.

Neste método, todos os L SNPs são divididos em três grupos ($l_0 + l_1 + l_2 = L$), o grupo 0 é formado por SNPs que não estão associados com uma doença, o grupo 1 por SNPs que

[1] A probabilidade a posteriori informa a chance de ocorrer um determinado SNP, dado que já se conhece sua probabilidade a priori.

contribuem com a doença apenas com efeitos principais e o grupo 2 por SNPs que contribuem com a doença através de interação. O principal objetivo do método é encontrar os SNPs pertencentes aos grupos 1 e 2.

Dessa forma, considerando que os SNPs da classe caso, tenham uma distribuição diferente dos SNPs da classe controle, um modelo de verossimilhança é descrito assumindo independência entre os SNPs da população controle.

A informação da população caso, formada pelos grupos D_0, D_1 e D_2 é utilizada para obter a distribuição a posteriori de cada SNP I, de acordo com a equação:

$$P(I|D,U) \propto P(D_1|I)P(D_2|I)P(D_0,U|I)P(I) \tag{4.12}$$

onde D, representa o genótipo da população caso dividido nos grupos D_0, D_1 e D_2. E U representa os genótipos dos SNPs nos grupos 0, 1 e 2 respectivamente.

Um limitante é imposto para determinar o número máximo de interações que podem ser analisadas pelo método, sendo definido como $l_2 \leq log_3(N_d) - 1$. Os valores da probabilidade a priori p_1 (probabilidade do SNP pertencer ao grupo 1) e p_2 (probabilidade do SNP pertencer ao grupo 2), são definidos de acordo com o conhecimento prévio dos dados, e os valores podem ser modificados livremente, mas por padrão foram definidos como $p_1 = 0,01$, ou seja a probabilidade do SNP possuir efeitos marginais e $p_2 = 0,01$ probabilidade do SNP possuir efeitos epistáticos.

Então, com o objetivo de definir a probabilidade a posteriori dos SNPs, basicamente o algoritmo de *Metropolis-Hastings* (método Monte Carlo) tem um contador de iterações inicializado e um valor inicial é definido para I pela probabilidade a priori $P(I)$:

$$P(I) \propto p_1^{l_1} p_2^{l_1} (1 - p_1 - p_2)L - l_1 - l_2 \tag{4.13}$$

Em seguida, um novo valor da distribuição é calculando utilizando a equação (4.12). Para isso, dois métodos são utilizados para fazer as trocas dos SNPs entre os grupos 0, 1 e 2, podendo ser mudanças aleatórias do grupo ao qual o SNP pertence ou trocas aleatórias de dois SNPs entre os grupos. A mudança proposta é aceita de acordo com a razão de *Metropolis-Hastings*. A saída é a distribuição a posteriori dos SNPs e as interações associadas com à doença.

Apesar do método fazer inferências estatísticas diretamente a partir das probabilidades a posteriori, os resultados também são analisados utilizando um teste de hipótese para verificar cada marcador ou conjunto de marcadores para as associações significativas com a doença. Para testar cada M conjunto de SNPs, assumindo a hipótese nula H_0 como M não associado a doença é utilizada a estatística B:

$$B_M = ln\frac{P_A(D_M,U_M)}{P_O(D_M,U_M)} = \frac{P_{join(D_M[Pind(U_M)+P_{join}(U_M)]}}{P_{ind}(D_M,U_M) + P_{join}(D_M,U_M)} \tag{4.14}$$

onde D_M e U_M são os genótipos para M casos e controles. $P0$ (DM, UM) e PA (DM, UM) são os fatores de Bayes sob a hipótese nula e alternativa respectivamente. Na hipótese nula diz que

os genótipos de caso e controles seguem a mesma distribuição, enquanto a hipótese alternativa diz que seguem distribuições diferentes.

Versões mais recentes da abordagem, como BEAM3 (Zhang, 2012) desenvolvida para GPU, tem como objetivo de tornar-se um método viável para o uso em conjunto de dados de GWAS, embora ainda não seja completamente viável para análise de interações de alta ordem.

Co-Information Based Method for Detecting and Visualizing n-order Epistatic Interactions (CINOEDV)

A *Co-information based method for detecting and visualizing n-order epistatic interactions* (CINOEDV) é uma ferramenta proposta por Shang *et al.* (2016) e destaca-se por ser proposta para avaliar interações do tipo *n*-order. O CINOEDV possui duas diferentes estratégias de busca para identificar interações, sendo uma delas uma busca exaustiva e a outra estocástica baseada em *particle swarm optimization* (PSO).

A busca exaustiva foi proposta para investigar conjunto de dados com baixa dimensionalidade ou para avaliar interações entre SNPs de baixa ordem. Assim, para conjuntos de dados com interações de alta ordem e/ou alta dimensionalidade deve-se utilizar a estratégia de busca PSO.

O CINOEDV utiliza a métrica *co-information* para avaliar a relevância das interações.

$$CI(S_1;...;S_n;C) = -\sum_{T\subseteq V}(-1)^{n+1-|T|}H(T) \tag{4.15}$$

onde S representa os SNPs e C a classe ao qual pertence.

Essa métrica é capaz de avaliar as dependências multivariáveis. Para isso, calcula a entropia conjunta de todos os subconjuntos de T em V, sendo $H(T)$ a entropia conjunta de T dada pela equação (4.16).

$$H(T) = -\sum_{t\in T} p(t)logp(t) \tag{4.16}$$

No entanto, a métrica *co-information* (CI) tem dois problemas principais: o primeiro diz respeito ao valor não normalizado obtido pela métrica, quando se lida com problemas bivariados ou multivaridos. E o segundo esta relacionado a ordem como as interações são apresentadas influenciar no resultado final. Para resolver esses dois problemas, foi feita uma normalização da equação CI utilizando uma média ordenada dos valores de CI obtidos.

No entanto, com apenas essa modificação a métrica CI não era capaz de avaliar a contribuição de uma interação. Para isso, é utilizada a equação (4.17).

$$CCI(S_1;...;S_n;C) = \sum_{Z\subseteq CS\cap Z\subseteq S_1;...;S_n} NCI(Z';C) \tag{4.17}$$

onde Z representa todos os SNPs no conjunto Z, C representa o fenótipo e CS é o conjunto de

interações de SNP que tem seu NCI maior que um determinado limiar.

O CINOEDV é uma abordagem eficiente e adaptável para diferentes tipos de problemas, seja para tratar conjunto de dados com baixa ou alta dimensionalidade. Como proposto, o CINOEDV é capaz de avaliar interações com *n* combinações, uma característica inovadora. No entanto, para realizar esse tipo de procedimento faz-se necessário configurar o número de interações que se deseja pesquisar previamente e utilizar a forma de busca PSO, informando o número de partículas e a quantidade de iterações desejada.

Conclusão

Na Tabela 4.5, é apresentada uma visão geral das ferramentas de inferência epistáticas estudadas, incluindo informações como suas estratégias de busca e sobre suas características gerais, tais como número de interações que realizam inferência e algumas desvantagens da abordagem.

Tabela 4.5 Características Gerais das Ferramentas de Inferência de Epistasia

Ferramenta	**Busca**	**Interações**	**Desvantagens**
MDR	Exaustiva	*n-order*	Análise exponencial
BOOST	Exaustiva	Apenas 2	Limite de 2 interações
SNPRuler	Gulosa	*n-order*	-
BEAM	Estocástica	*n-order*	-
CINOEDV	Estocástica e Exaustiva*	*n-order*	*Análise exponencial

Na Seção 4.3, será apresentada uma visão geral dos algoritmos baseados em aprendizagem de máquina, dando um maior enfoque aos algoritmos baseados em *Relevance Learning Vector Quantization*, sendo estes, alvos de estudo nesta tese.

A escolha dessa classe de algoritmos, se dá pelas capacidades já demostradas em estudos anteriores (Araujo e Guimaraes, 2011) e (Araujo *et al.*, 2011). Em Strickert *et al.* (2006), foram capazes de lidar com problemas biológicos, e em Kietzmann *et al.* (2008) utilizados em dados de alta dimensionalidade. Assim também é levantada a hipótese, que o uso do vetor de relevâncias por esses algoritmos, pode ser uma característica promissora, para que sejam capazes de lidar com o problema da inferência de SNPs relevantes.

4.3 Algoritmos de Aprendizagem de Máquina

Os algoritmos de aprendizagem de máquina são capazes de lidar com inúmeros problemas, tais como, identificação de padrões, classificação e clusterização dos dados. A vantagem do uso desses algoritmos, se dá pelo fato de serem capazes de aprender com seus erros, e realizar inferências ou predições sobre os dados.

Em geral, os dados não fornecem informações facilmente observáveis. Para isso, esses algoritmos são empregados, na tentativa de identificar correlações ou padrões nos dados. Com o aprendizado das informações contidas nos dados, são capazes de classificar e/ou tomar decisões. De acordo, com o tipo de dados utilizado, esses algoritmos podem ser divididos em dois grupos principais:

Não Supervisionados: neste caso, os algoritmos realizam um agrupamento dos dados sem a presença de rótulos. Para realizar este agrupamento, os algoritmos, tais como, *Self-Organizing Maps* (SOM) (Kohonen, 1997) e *Growing Neural Gas* (GNG) (Fritzke, 1995) que buscam identificar similaridades ou padrões nos dados.

Supervisionados: neste grupo, os algoritmos recebem como entrada, dados que possuem rótulos ou classes. Os conjuntos de dados analisados nesta tese, contém os rótulos, sendo as classes dos pacientes casos e controles. Neste caso, os algoritmos, tais como, *Learning Vector Quantization* (LVQ) (Kohonen, 1997), *Supervised Growing Neural Gas* (Garcia e Forster, 2012), *Supervised Relevance Neural Gas* (Hammer *et al.*, 2005), fazem uso de uma informação extra para classificar os dados.

Nesta tese, os algoritmos de aprendizagem de máquina selecionados, precisam atender uma série de propriedades para que estejam aptos a lidar com o problema da inferência das interações entre SNPs. Para isso, além de seus resultados serem de fácil interpretação, os algoritmos selecionados devem ser capazes de:

- Lidar com dados com alta dimensionalidade;
- Considerar as interações entre os dados sem que para isso precisem realizar análises exaustivas; e
- Realizar o *ranking* das dimensões relevantes.

Tendo isso em mente, algoritmos supervisionados e baseados em *Learning Vector Quantization* (LVQ) apresentados na Figura 4.5 foram selecionados. Os algoritmos baseados em LVQ, compõem uma família que utilizam os protótipos[2] (*codebook vectors*) para representar a região de uma classe de dados, denominada de região de Voronoi.

[2] Vetores que representam um subconjunto de amostras de mesma classe dos dados.

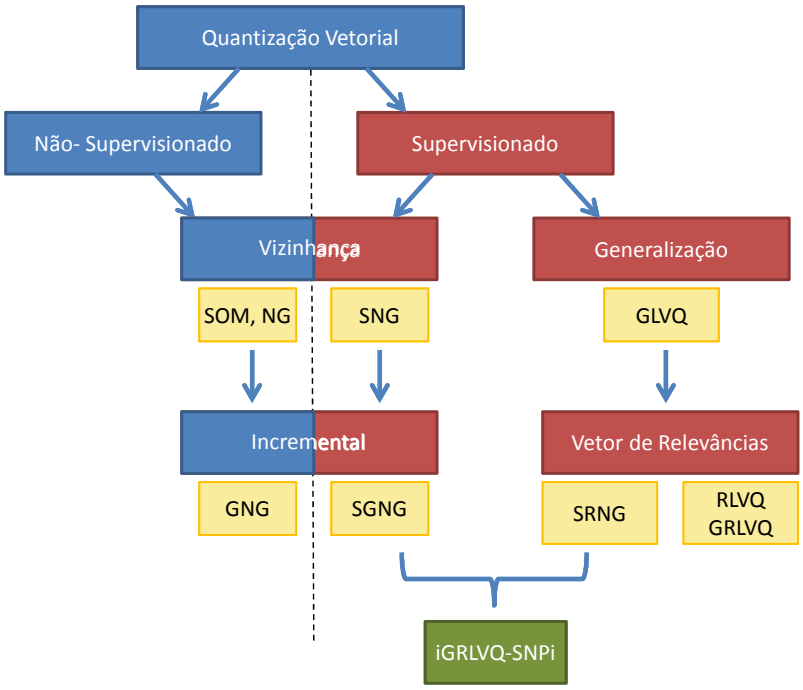


Figura 4.5 Características dos classificadores baseados em quantização vetorial. Nas caixas em vermelho, são apresentadas relevantes características dos algoritmos estudadas neste tese. Nas caixas em amarelo, são apresentadas as siglas dos algoritmos estudados. Na caixa em verde, está identificado o modelo proposto nesta tese. E nas caixas em azul, são exemplos de algoritmos não supervisionados citados ao longo da tese, mas não utilizados nos experimentos.

Na Figura 4.6, é apresentado um esquema do comportamento dos algoritmos baseados em LVQ. No espaço de dados, é mostrado os dados representados pelas bolas em verde, as estrelas pretas representam os protótipos que serão os representantes de um grupo de amostras e a separação entre esses grupos limitadas pelas regiões de Voronoi (linha azul tracejada).

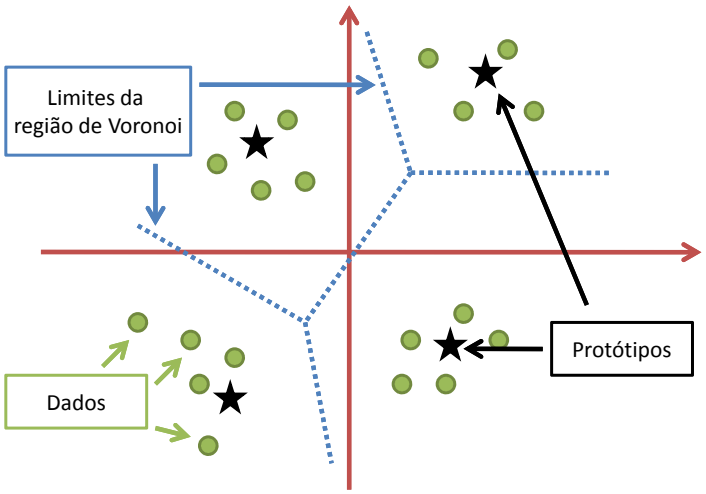


Figura 4.6 Visualização em duas dimensões da representação da estrutura dos algoritmos baseados em LVQ. Os círculos verdes representam os dados, As estrelas representam os protótipos e a linha tracejada em azul indica as delimitações entre as classes de dados, representada pela região de Voronoi.

A ideia básica da **teoria da quantização vetorial**, é realizar uma codificação de um

grande conjunto de dados em um pequeno grupo de representantes ou protótipos. Durante o processo de aprendizagem, amostras são selecionadas e os protótipos mais próximos são atualizados, tornando-se o representante de um subconjunto.

Os classificadores baseados em LVQ, são particularmente simples e fáceis de entender. Basicamente, o algoritmo possui os protótipos que são os representantes de uma classe de dados que compartilham similaridades entre si. Cada protótipo está associado a apenas uma das possíveis classes (rótulos), porém, pode haver mais de um protótipo associado a uma mesma classe. Assim, cada dado que é agrupado em um determinado protótipo, é classificado com o mesmo rótulo deste protótipo.

O custo computacional dos algoritmos baseados em LVQ depende do número de protótipos, da dimensionalidade dos dados e do número de épocas[3]. Nas próximas seções, serão apresentados com mais detalhes, diferentes variantes do LVQ1, incluindo o modelo proposto, utilizados nos experimentos desta tese.

4.3.1 *Learning Vector Quantization*

Na década de 80, Teuvo Kohonen apresentou a primeira versão do algoritmo LVQ (LVQ1). Após a proposição do LVQ1, muitas outras variantes surgiram, pois os primeiros algoritmos baseados em LVQ apresentavam muitos problemas, como sensibilidade na inicialização, problemas de baixa convergência do algoritmo e instabilidades.

Para resolver alguns desses problemas, foi proposto por Sato e Yamada (1996), o *Generalized Learning Vector Quatization* (GLVQ). Nesse algoritmo, foi introduzida uma função de custo a ser minimizada através de um gradiente estocástico descendente.

$$C_{GLVQ} = \sum_{i=1}^{m} sgd(\mu(x_i)) \tag{4.18}$$

onde a função $sgd(x) = (1 + exp(-x))^{-1}$ é uma sigmoide logística e $\mu(x_i)$ e dada por:

$$\mu(x_i) = \frac{d^+(x_i) - d^-(x_i)}{d^+(x_i) + d^-(x_i)}, \tag{4.19}$$

onde $\mu(x_i)$ pode ter um valor entre -1 a 1. Quando $\mu(x_i)$ é negativo a amostra é classificada corretamente. E $d^+(x_i)$ e $d^-(x_i)$ são respectivamente as distâncias Euclidianas quadráticas da amostra para o protótipo da mesma classe ω^+ e da classe oposta ω^-. Diante disso, a função de custo (4.18) tem como objetivo ser minimizada para que uma melhor classificação seja obtida, para isso o algoritmo é treinado por n épocas ou até a convergência do algoritmo.

Durante cada ciclo de treinamento do algoritmo, uma amostra x_i, é aleatoriamente selecionada, e os protótipos mais próximos da mesma classe ω^+ e da classe distinta ω^- são identificados utilizando as distâncias Euclidianas quadráticas $d^+(x_i)$ e $d^-(x_i)$, como mostra na Figura 4.7.

[3]Ciclos de execução do algoritmo.

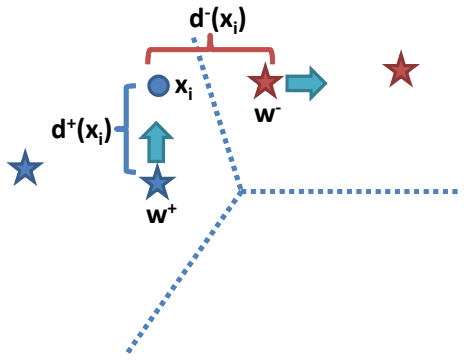


Figura 4.7 Representação da identificação dos protótipos mais próximos da amostra através da menor distância obtida com $d^+(x_i)$ e $d^-(x_i)$ da amostra x_i (círculo azul) para o protótipo da mesma classe da amostra (estrela azul) e da classe oposta (estrela vermelha). Os protótipos mais próximos selecionados, terão seus valores atualizados de acordo com as Equações 4.20.

onde, a $d^+(x_i) = (x_i - \omega^+)^2$ e $d^-(x_i) = (x_i - \omega^-)^2$ são as distâncias Euclidianas da amostra para os protótipos.

Sabendo que ω^+ e ω^- são os protótipos mais próximos da amostra x_i, sendo ω^+ da mesma classe e ω^- de uma classe distinta. Esses, protótipos são atualizados de acordo com as equações:

$$\begin{aligned} \Delta\omega^+ &= \varepsilon^+ \times sgd'_{\mu(x^i)} \times \xi^+ \times 2 \times (\omega^+ - x^i) \\ \Delta\omega^- &= -\varepsilon^- \times sgd'_{\mu(x^i)} \times \xi^- \times 2 \times (\omega^- - x^i), \end{aligned} \tag{4.20}$$

onde ε^+ e ε^- são respectivamente, as taxas de aprendizagem positiva e negativa (TAP e TAN) e ξ^+ e ξ^- as distâncias derivativas:

$$\xi^+ = \frac{2 \times d^+(x_i)}{(d^+(x_i) + d^-(x_i))^2} \quad e \quad \xi^- = \frac{2 \times d^-(x_i)}{(d^+(x_i) + d^-(x_i))^2}. \tag{4.21}$$

Ao final de cada época, a taxa de aprendizagem positiva e negativa sofrem decaimento de acordo com:

$$\varepsilon(t) = \frac{\varepsilon(0)}{1 + \tau \times (t - t0))}, \tag{4.22}$$

a fim de que o processo de aprendizagem dos protótipos se estabilize, e ocorra a convergência do algoritmo.

O algoritmo básico do GLVQ é apresentado no Algoritmo 4.

Algoritmo 4: Algoritmo básico do GLVQ

Entrada: Conjunto de dados X

Saída: Conjunto de protótipos W

Inicialize os parâmetros: ε^+, ε^-, τ, n, m

Inicialize os protótipos: W = $\omega^1, \omega^2, ..., \omega^m$ com posições aleatórias e classes alternadas

para $j \leftarrow 1$ ***até*** *o número de épocas n* **faça**

  para $i \leftarrow 1$ ***até*** *o tamanho do conjunto de dados* **faça**

    Selecione uma amosta aleatória x^i do conjunto de dados;

    Encontre os protótipos mais próximos da mesma classe ω^+ e da classe distinta ω^-

    Atualize os protótipos de acordo com (4.20):

    $\omega^+ \leftarrow \omega^+ + \Delta\omega^+$

    $\omega^- \leftarrow \omega^- + \Delta\omega^-$

  fim

  Realize o decaimento das taxas de aprendizagem de acordo com (4.22):

  $\varepsilon^+ \leftarrow \varepsilon^+ \times \tau$

  $\varepsilon^- \leftarrow \varepsilon^- \times \tau$

fim

No entanto, apesar desse avanço na implementação da função de custo, a maioria dos algoritmos baseados em LVQ não fazem uma discriminação entre as características mais ou menos relevantes, e por isso, não são capazes de realizar a inferência dos SNPs relevantes.

Com a introdução do vetor de relevância no algoritmo do LVQ, proposto por Bojer *et al.* (2001) no algoritmo RLVQ e a sua generalização (GRLVQ) proposta por Hammer e Villmann (2002), tornou possível, a avaliação das dimensões relevantes. Considerando a aplicação no problema da inferência dos SNPs, no vetor de relevâncias serão armazenados, o quão importante é um SNP para a classificação dos dados. Na Seção 4.3.2 será apresentado com detalhes o algoritmo do GRLVQ.

4.3.2 *Generalized Relevance Learning Vector Quantization*

O GRLVQ é uma modificação mais robusta do RLVQ e mais amplamente utilizado como algoritmo baseado em *Relevance Learning Vector Quantization*. O GRLVQ tem sido aplicado em diferentes tipos de problemas, como classificação de imagens (Kietzmann *et al.*, 2008), séries temporais (Strickert *et al.*, 2001) e inclusive em problemas biológicos, como análise de expressão de dados de *microarray* (Strickert *et al.*, 2006) e espectrometria de massa (Schneider *et al.*, 2007).

A principal diferença entre o GRLVQ e o GLVQ, reside no fato que no GRLVQ é incrementada uma métrica adaptativa no treinamento do classificador que utiliza informações oriundas do vetor de relevâncias.

A adaptação da métrica de distância, é obtida através da introdução do vetor de relevância $\mu_\lambda(x_i)$ no cálculo da distância. Obtendo assim, uma distância ponderada pela relevância de cada dimensão (equação (4.23)). Esta modificação, confere ao método a habilidade de lidar com conjuntos de dados com alta dimensionalidade, pois dados irrelevantes e/ou ruidosos tem sua influência significativamente reduzida, com o uso do vetor de relevâncias durante o treinamento do algoritmo.

A função objetivo do GRLVQ é muito semelhante a apresentada no GLVQ, no entanto, é utilizada uma função de distância com pesos λ em sua equação:

$$\begin{aligned} C_{GRLVQ} &= \sum_{i=1}^{m} sgd(\mu_\lambda(x_i)) \\ \mu_\lambda(x_i) &= \frac{d_\lambda^+(x_i) - d_\lambda^-(x_i)}{d_\lambda^+(x_i) + d_\lambda^-(x_i)} \end{aligned} \tag{4.23}$$

A cada novo padrão x_i apresentado, são atualizados o protótipo mais próximo de mesma classe ω_{r^+}, de acordo com:

$$\Delta\omega_{r^+} = \varepsilon^+ \times sgd'_{\mu_\lambda(x_i)} \times \xi^- \times \frac{\partial d_\lambda^+(x_i)}{\partial \omega_{r^+}}, \tag{4.24}$$

o protótipo mais próximo da classe oposta ω_{r^-}, de acordo com:

$$\Delta\omega_{r^-} = -\varepsilon^- \times sgd'_{\mu_\lambda(x_i)} \times \xi^+ \times \frac{\partial d_\lambda^-(x_i)}{\partial \omega_{r^-}}, \tag{4.25}$$

e o vetor de relevâncias, de acordo com:

$$\Delta\lambda = -\varepsilon^\lambda \times sgd'_{\mu_\lambda(x_i)} \times \frac{\frac{2\times\partial d_\lambda^+(x_i)}{\partial\lambda\times d_\lambda^-(x_i)} - \frac{2\times\partial d_\lambda^+(x_i)\times\partial d_\lambda^-(x_i)}{\partial\lambda}}{(d_\lambda^+(x_i) + d_\lambda^-(x_i))^2} \tag{4.26}$$

onde ε^+, ε^- e ε^λ são respectivamente, as taxas de aprendizagem positiva (TAP), negativa (TAN) e do vetor de relevâncias (TAW). As distâncias derivativas com pesos são calculadas através das equações:

$$\xi^- = \frac{2\times d_\lambda^-(x_i)}{(d_\lambda^+(x_i) + d_\lambda^-(x_i))^2} \quad e \quad \xi^+ = \frac{2\times d_\lambda^+(x_i)}{(d_\lambda^+(x_i) + d_\lambda^-(x_i))^2} \tag{4.27}$$

O pseudo algoritmo do GRLVQ é apresentado no Algoritmo 5.

Algoritmo 5: Algoritmo do GRLVQ

Entrada: Conjunto de dados X

Saída: Conjunto de protótipos W

Inicialize os parâmetros: ε^+, ε^-, ε^λ, τ, n, m

Inicialize os protótipos: W = $\omega^1, \omega^2, ..., \omega^m$ com posições aleatórias e classes alternadas

para $j \leftarrow 1$ ***até*** *o número de épocas n* **faça**

  para $i \leftarrow 1$ ***até*** *o tamanho do conjunto de dados* **faça**

    Selecione uma amosta aleatória x^i do conjunto de dados;

    Encontre os protótipos mais próximos da mesma classe ω_{r^+} e da classe distinta ω_{r^-}

    Atualize os protótipos de acordo com:

    $\omega_{r^+} \leftarrow \omega^+ + \Delta\omega_{r^+}$

    $\omega_{r^-} \leftarrow \omega^- + \Delta\omega_{r^-}$

    Atualize o vetor de relevâncias de acordo com:

    $\lambda \leftarrow \lambda + \Delta\lambda$

  fim

  Realize o decaimento das taxas de aprendizagem:

  $\varepsilon^+ \leftarrow \varepsilon^+ \times \tau$

  $\varepsilon^- \leftarrow \varepsilon^- \times \tau$

  $\varepsilon^\lambda \leftarrow \varepsilon^\lambda \times \tau$

fim

No total o GRLVQ possui 6 diferentes parâmetros descritos na Tabela 4.6.

Tabela 4.6 Descrição dos Parâmetros do GRLVQ com os intervalos utilizados

Parâmetros	Descrição
TAP	Taxa de Aprendizagem Positiva
TAN	Taxa de Aprendizagem Negativa
TAW	Taxa de Aprendizagem do Vetor de Pesos
TAU	Taxa de Decaimento das Taxas de Aprendizagem
EPOCHS (n)	Número de Ciclos de Aprendizagem do Algoritmo
NNODES (m)	Número de Nodos ou Protótipos

Apesar do GRLVQ ser uma abordagem robusta, ainda apresenta alguns problemas como sensibilidade a inicialização dos protótipos. Alguns protótipos tendem a coletar uma grande quantidade de dados, deixando outros protótipos sem classificar nenhum padrão e também o algoritmo possui certa tendência de ficar preso em ótimos locais, não encontrando o melhor resultado possível.

Uma variante do GRLVQ proposta por Hammer *et al.* (2005) é o *Supervised Relevance Neural Gas* (SRNG). O SRNG é baseado no algoritmo não supervisionado *Neural Gas* (NG) (Martinetz *et al.*, 1993), o qual introduziu um esquema de cooperação de vizinhança entre os

protótipos, tornando o método menos sensível a inicialização dos protótipos. Na Seção 4.3.3 será apresentado com detalhes o SRNG.

4.3.3 *Supervised Relevance Neural Gas*

A ideia do SRNG é inserir a cooperação da vizinhança na função de custo do GRLVQ. A vizinhança inserida, ajuda os protótipos a se espalharem sobre o espaço de busca do conjunto de dados, sem que sejam tão fortemente afetados pelos ruídos nos dados. A vizinhança reduz também, a possibilidade dos protótipos ficarem presos em ótimos locais.

A função de custo utilizada pelo SRNG é descrita como:

$$C_{SRNG} = \sum_{i=1}^{m} \sum_{\omega_{r^+} \in W^{x_i}} \frac{h_\gamma(r^+(x_i, W^{x_i})) \times sgd(\mu_\lambda(x_i))}{C(\gamma, K^{x_i})} \tag{4.28}$$

onde W^{x_i} é um subconjunto dos protótipos de W de mesma classe que a amostra x_i. $h_\gamma(r^+(x_i, W^{x_i}))$ é o grau de cooperatividade da vizinhança, dado pela equação:

$$h_\gamma(r^+(x_i, W^{x_i})) = exp\left(-\frac{r^+(x_i, W^{x_i})}{\gamma}\right), \tag{4.29}$$

onde $r^+(x_i, W^{x_i})$ fornece o *rank* dos protótipos da vizinhança e γ (GAMMA) é o parâmetro que determina o intervalo da cooperação da vizinhança. A função $sgd(\mu_\lambda(x_i))$, é uma função sigmoide logística dada por:

$$sgd(\mu_\lambda(x_i)) = (1 + exp(-\mu_\lambda(x_i)))^{-1}, \tag{4.30}$$

onde, $\mu_\lambda(x_i)$ é:

$$\mu_\lambda(x_i) = \frac{d_\lambda^+(x_i) - d_\lambda^-(x_i)}{d_\lambda^+(x_i) + d_\lambda^-(x_i)}, \tag{4.31}$$

e $d_\lambda^+(x_i) = (x_i - \omega_{r^+})^2$, as distâncias adaptativas da amostra para o protótipo de mesma classe e $d_\lambda^-(x_i) = (x_i - \omega_{r^-})^2$ a distância adaptativa da amostra para uma classe distinta. E $C(\gamma, K^{x_i})$, uma constante de normalização, dependente do intervalo de cooperação da vizinhança γ e da cardinalidade K de W.

Para a minimização da função de custo (equação (4.28)), enquanto que no GRLVQ são selecionados apenas os dois protótipos mais próximos da amostra, sendo um da mesma classe da amostra e o outro da classe distinta. No SRNG a atualização dos protótipos considera, não só o protótipo mais próximo de mesma classe da amostra, como também, a vizinhança do protótipo selecionado.

Assim, todos os protótipos de W^{x_i} (de mesma classe da amostra x_i), o protótipo mais próximo da classe distinta e o vetor de relevâncias são atualizados de acordo com as equações

abaixo, que levam em consideração a função de vizinhança introduzida:

$$\Delta\omega_{r^+} = \varepsilon^+ \times \frac{sgd'_{\mu_\lambda(x_i)} \times \xi^+ \times h_\gamma(r^+(x_i, W^{x_i}))}{C(\gamma, K^{x_i})} \times \frac{\partial d^+_\lambda(x_i)}{\partial \omega^{i+}}$$

$$\Delta\omega_{r^-} = -\varepsilon^- \times \sum_{\omega_{r^+} \in W^{x_i}} \frac{sgd'_{\mu_\lambda(x_i)} \times \xi^- \times h_\gamma(r^+(x_i, W^{x_i}))}{C(\gamma, K^{x_i})} \times \frac{\partial d^+_\lambda(x_i)}{\partial \omega^{i+}}$$

$$\Delta\omega_{r^-} = -\varepsilon^\lambda \times \sum_{\omega_{r^+} \in W^{x_i}} \frac{sgd'_{\mu_\lambda(x_i)} \times \xi^- \times h_\gamma(r^+(x_i, W^{x_i}))}{C(\gamma, K^{x_i})} \times \left(\varepsilon^+ \times \frac{\partial d^+_\lambda(x_i)}{\partial \lambda^{i+}} - \varepsilon^- \times \frac{\partial d^-_\lambda(x_i)}{\partial \lambda^{i+}} \right) \tag{4.32}$$

onde ξ^- e ξ^+ são calculadas pelas equações (6.6) e $C(\gamma, K^{x_i})$ é uma constante de normalização da vizinhança, dependente da cooperação da vizinhança γ e da cardinalidade K.

No total, o SRNG possui 7 parâmetros, contendo os mesmos 6 do GRLVQ mais o parâmetro da cooperação da vizinhança γ. Este parâmetro, esta relacionado a quantidade de protótipos que serão atualizados durante o processo de treinamento, quanto mais próximo de 1, maior será o raio da vizinhança, englobando assim um maior número de protótipos.

Além da inserção do vetor de relevâncias e da vizinhança dos protótipos, uma terceira característica relevante dos algoritmos baseados em RLVQ, foi implementada no algoritmo *incremental* GRLVQ (iGRLVQ) proposto por Kietzmann *et al.* (2008). A introdução de uma abordagem com adição automática dos protótipos ao mapa, favoreceu aos algoritmos adequarem o número de protótipos à complexidade do conjunto de dados utilizado. Assim, dados desbalanceados ou classes de dados mais complexas, tem o número de protótipos ajustados para cada classe de dados, favorecendo uma melhor classificação dos dados.

4.3.4 *Incremental Generalized Relevance Learning Vector Quantization*

O iGRLVQ (Kietzmann *et al.*, 2008) é um algoritmo proposto para identificar características relevantes em objetos 3D. No iGRLVQ a quantidade de protótipos é ajustada automaticamente de acordo com a complexidade dos dados. Em algoritmos como o GRLVQ ou o SRNG o número de protótipos é definido de forma fixa e são em geral fornecidos a mesma quantidade de protótipos para cada classe.

No iGRLVQ, a inserção de um novo protótipo ocorre quando a quantidade de erros durante o treinamento ultrapassa um determinado limiar g_{max}. Para contabilizar os erros ocorridos durante o processo de treinamento, um parâmetro de erro de classificação é incrementado toda vez que uma amostra é classificada incorretamente. Como mostrado na Figura 4.8, uma amostra

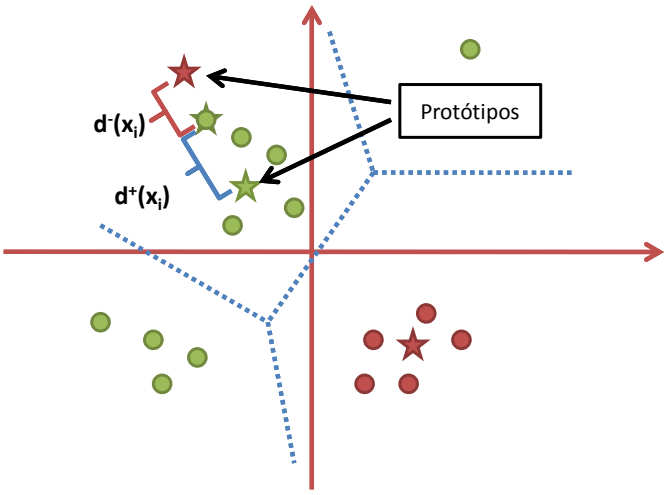


Figura 4.8 Esquema de inserção de novos protótipos. Um novo protótipo é inserido sobre a amostra (verde) que tem uma distância $d_\lambda^+(x_i)$ maior do que $d_\lambda^-(x_i)$.

é classificada incorretamente quando a distância $d_\lambda^+(x_i)$ é maior do que a distância $d_\lambda^-(x_i)$, isso significa dizer que o protótipo da classe oposta da amostra está mais próximo da amostra do que o protótipo da mesma classe. Sempre que este tipo de erro de classificação ocorre, o contador g é incrementado e a amostra incorretamente classificada é armazenada no conjunto S juntamente com a distância $d_\lambda^-(x_i)$ obtida.

Ao final de uma época, durante o processo de treinamento, se o valor de g ultrapassar o limiar do g_{max}, um novo protótipo é adicionado para cada classe. A posição de inserção do novo protótipo será dada pela posição da amostra no conjunto S que tem a menor distância $d_\lambda^-(x_i)$ para que o protótipo fique posicionado nos limites da região de decisão, e sua classe será a mesma classe da amostra sobre a qual será posicionado, essa amostra tem a menor distância $d_\lambda^-(x_i)$. A posição de inserção do novo protótipo faz com que este fique posicionado próximo as limites das regiões de decisão. Após a adição do novo protótipo o contador g é setado para 0.

Um outro algoritmo estudado foi o *Supervised Growing Neural Gas* (SGNG) proposto por Garcia e Forster (2012). O SGNG, assim como o iGRLVQ, também realiza a inserção automática dos protótipos. No entanto, as regras para a inserção desses novos protótipos é diferente da utilizada no iGRLVQ, sendo melhor detalhada na seção a seguir.

4.3.5 *Supervised Growing Neural Gas*

O *Supervised Growing Neural Gas* (SGNG) proposto por Garcia e Forster (2012) é um versão supervisionada do algoritmo *Growing Neural Gas* (GNG), que realiza, durante o processo de treinamento, a inserção e remoção de novos protótipos. Assim como, o SRNG, o GNG e o SGNG, o SGNG tem a cooperação da vizinhança entre seus protótipos. No entanto, no SGNG, os protótipos são ligados através de arestas independente da classe que representam, fazendo com que seja construída uma rede que conecta os protótipos no mapa.

O treinamento do SGNG, é inicializado com um único protótipo para cada classe do

conjunto de dados. Inicialmente, todos os protótipos são conectados e posicionados aleatoriamente. Para cada nova amostra apresentada, são atualizados os dois protótipos mais próximos e o protótipo mais próximo da mesma classe.

Os protótipos da mesma classe da amostra são aproximados, e os de classe oposta são afastados, assim como ocorre no algoritmo do GLVQ. As conexões entre os protótipos são inseridas e/ou removidas de acordo com um limiar do número máximo de conexões por protótipo. Além disso, são contabilizados os erros de classificação para cada classe.

Quando o critério de inserção de um novo protótipo é alcançado, sendo este critério relacionado com a estabilização do erro de classificação ou com uma quantidade fixa de épocas, um novo protótipo é adicionado.

Para adicionar um novo protótipo, é realizada uma busca do protótipo com o maior número de erros de classificação. Esse protótipo é duplicado, e o novo protótipo duplicado é posicionado sobre o protótipo original, no entanto, o novo protótipo tem sua classe modificada.

Conclusão

Neste capítulo foram apresentadas diferentes abordagens computacionais, sendo divididas em três grupos principais: os filtros de seleção de características; as ferramentas referenciadas na literatura para realizar inferência das interações entre SNPs; e por último os algoritmos baseados em LVQ.

Dentre os filtros de seleção de características, foram selecionados os multivariados, mais representativos e amplamente discutidas na literatura e capazes de lidar com dados com alta dimensionalidade. Um estudo comparativo entre esses filtros pode ser visto no Capítulo 5.

As ferramentas de inferência das interações, foram selecionadas considerando exemplares com diferentes formas de busca exaustiva, gulosa e estocástica, assim como, suas acurácias e *performances* já observadas na literatura (Araujo e Guimaraes, 2011), (Araujo *et al.*, 2011) e (Shang *et al.*, 2011).

Os algoritmos baseados em LVQ, foram selecionados considerando suas capacidades em lidar com os dados e os desafios do problema de inferência das interações. Para avaliá-los, foram realizados experimentos, apresentados nos Capítulo 5 e Capítulo 6, para verificar suas limitações e capacidades em lidar com os diferentes conjuntos de dados.

5

Experimentos com Abordagens na Literatura

Neste capítulo serão apresentados experimentos e resultados obtidos com a análise de diferentes abordagens computacionais ao lidar com dados de SNPs. Inicialmente, na Seção 5.1, serão apresentados os resultados obtidos com as quatro técnicas de filtragem multivariadas: CFS, FCBF, ReliefF e INTERACT, descritas na Seção 4.1.1.

Na Seção 5.2 será apresentado o desempenho de dois algoritmos baseados em RLVQ: o GRLVQ e SRNG, avaliando e verificando a influência dos parâmetros sobre sua acurácia. Assim, como a importância do ajuste dos parâmetros no desempenho dos algoritmos. A partir da análise dos resultados resultados obtidos, foram criadas hipóteses sobre o comportamento dos algoritmos no problema de inferência das interações entre SNPs.

As hipóteses são testadas na Seção 5.3, incluindo estudos sobre a adição automática de protótipos com o uso do iGRLVQ, a verificação da convergência do algoritmo utilizando o vetor de relevâncias e avaliação das interações entre SNPs com o uso de métricas estatísticas.

5.1 Filtros Multivariados

Os métodos de seleção de características, vêm sendo utilizados para a redução da dimensionalidade dos dados e minimização da complexidade de muitos problemas, como análise de dados de *microarray*, classificação de texto e imagens. Os experimentos descritos a seguir, tem por objetivo, avaliar capacidade dos filtros em lidar com dados de interação entre 2 SNPs (Seção 5.1.1) e de Alta Ordem (Seção 5.1.2). Em cada seção, os experimentos são descritos e os resultados apresentados e discutidos.

5.1.1 Experimentos com Dados com Interações entre 2 SNPs

Os conjuntos de dados simulados utilizados possuem apenas dois SNPs relevantes dentre um total de 20, 100 e 1000 SNPs e uma população de pacientes casos e controles (800 e 1600 pacientes). Cada conjunto de dados é formado pela combinação de valores da herdabilidade

(0.01, 0.025, 0.05, 0.1, 0.2, 0.3, e 0.4) e os valores da frequência do menor alelo (MAF) (0.2 e 0.4) formando **70 modelos epistáticos** com diferentes funções de penetrância (Velez *et al.*, 2007).

Devido à natureza multivariada dos dados, a utilização de filtros paramétricos univariados, tais como, *t*-Test, χ^2 Score e Fisher Score, assim como, os filtros não paramétricos, tais como, Gini Index e Information Gain, não são adequadas para o problema, pois não consideram a dependência existente entre as características nos conjuntos de dados. Ainda assim, esses filtros foram testados, mas como esperado não foram capazes de identificar os SNPs relevantes.

Vale ressaltar que, os filtros univariados são utilizados de forma eficiente na literatura. No entanto, para dados de GWAS, estas técnicas de filtragem devem ser evitadas quando o objetivo é identificar as interações relevantes entre as características. As demais técnicas multivariadas estudadas, ReliefF, INTERACT, CFS e FCBF, foram avaliadas e os seus resultados são exibidos na Tabela 5.1 e na Tabela 5.2.

Tabela 5.1 Acurácia dos filtros em conjuntos de dados com 800 pacientes. Abreviações dos métodos: R (ReliefF), I (INTERACT), C (CFS) e F (FCBF).

	Acurácia (%)											
Modelos	**20 SNPs**				**100 SNPs**				**1000 SNPs**			
Herd / MAF	**R**	**I**	**C**	**F**	**R**	**I**	**C**	**F**	**R**	**I**	**C**	**F**
0.4 / 0.4	100	71	16	1	87	2	6	0	0	0	0	0
0.4 / 0.2	100	68	14	0	100	2	2	0	4	0	0	0
0.3 / 0.4	100	73	17	1	50	1	5	0	0	0	0	0
0.3 / 0.2	100	67	16	0	99	3	5	0	3	0	0	0
0.2 / 0.4	100	67	14	0	23	1	4	0	0	0	1	0
0.2 / 0.2	100	66	18	0	67	2	6	0	0	0	1	0
0.1 / 0.4	92	68	13	0	3	1	8	0	0	0	1	0
0.1 / 0.2	88	63	17	1	9	4	6	0	0	0	1	0
0.05 / 0.4	80	65	12	0	2	0	6	0	0	0	0	0
0.05 / 0.2	38	60	20	1	1	3	5	0	0	0	0	0
0.025 / 0.4	56	65	16	0	3	1	5	0	0	0	0	0
0.025 / 0.2	22	58	15	1	0	2	6	0	0	0	0	0
0.01 / 0.4	15	67	16	0	1	1	4	0	0	0	0	0
0.01 / 0.2	3	57	16	1	0	3	7	0	0	0	0	0

Considerando que cada modelo possui 100 arquivos de amostras, o acerto máximo obtido será de 100, indicando que o filtro identificou os dois SNPs relevantes em todas as 100 amostras.

Nas Tabela 5.1 e Tabela 5.2, o ReliefF e o INTERACT, são os filtros com melhor desempenho identificados. Também é facilmente observável que o desempenho dos filtros é melhorado quando se utiliza um conjunto de dados com um número maior de pacientes.

De fato, realizando um teste estatístico *T*-Student a um nível de significância de 5%, nos resultados com 1600 pacientes e 20 SNPs. É observado que a hipótese nula $H_0 : \mu_0 = \mu_1$ não é rejeitada, apenas ao comparar os filtros ReliefF e INTERACT com p-value 2,15. Assim, os resultados dos filtros ReliefF e INTERACT não são estatisticamente diferentes.

Tabela 5.2 Acurácia dos filtros em conjuntos de dados com 1600 pacientes. Abreviações dos métodos: R (ReliefF), I (INTERACT), C (CFS) e F (FCBF).

Modelos	**Acurácia (%)**											
	20 SNPs				**100 SNPs**				**1000 SNPs**			
Herd / MAF	**R**	**I**	**C**	**F**	**R**	**I**	**C**	**F**	**R**	**I**	**C**	**F**
0.4 / 0.4	100	87	19	0	100	2	9	0	1	0	1	0
0.4 / 0.2	100	85	15	0	100	3	5	0	47	0	1	0
0.3 / 0.4	100	87	16	1	93	2	8	0	3	0	0	0
0.3 / 0.2	100	80	16	0	100	4	6	0	21	0	1	0
0.2 / 0.4	100	87	15	0	65	1	7	0	0	0	1	0
0.2 / 0.2	100	77	19	1	98	4	7	0	0	0	2	0
0.1 / 0.4	100	85	16	1	19	1	7	0	0	0	0	0
0.1 / 0.2	100	77	19	0	40	2	7	0	0	0	0	0
0.05 / 0.4	97	82	15	0	20	2	5	0	0	0	1	0
0.05 / 0.2	80	78	21	0	6	2	6	0	0	0	1	0
0.025 / 0.4	79	78	14	0	9	3	5	0	0	0	0	0
0.025 / 0.2	42	74	19	0	4	3	8	0	0	0	0	0
0.01 / 0.4	39	84	15	0	0	3	6	0	0	0	0	0
0.01 / 0.2	6	79	22	0	0	3	6	0	0	0	1	0

No entanto, os filtros só conseguem ter um bom desempenho em conjuntos de dados com a menor quantidade de características testadas (20 SNPs). A medida que o número de características aumenta, há uma queda acentuada no desempenho dos métodos de filtragem.

Devido à configuração de alguns parâmetros para criar os conjuntos de dados simulados, como o MAF e a herdabilidade, quanto menores esses valores, menor é a influência genética sobre a doença, assim é natural que os métodos apresentem resultados com baixa acurácia nessas situações (modelos com herdabilidade menores que 0.1). No entanto, o INTERACT e o CFS são filtros que, embora não tenham obtido resultados com 100% de acerto, assim como o ReliefF, não foram afetados pelos dados, o que é uma característica desejável para tratar dados de polimorfismos.

Na seção seguinte, serão apresentados os experimentos com estes filtros, utilizando conjuntos de dados com interações em alta ordem.

5.1.2 Experimentos com Dados com Interação de Alta Ordem

Os conjuntos de dados simulados de alta ordem proposto por Himmelstein *et al.* (2011), possuem 3, 4 ou 5 interações entre SNPs relevantes dentre um total de 20, 100 e 1000 SNPs e uma população de pacientes casos e controles (800 e 1600 pacientes). Os resultados obtidos, podem ser observados na Tabela 5.3, sendo omitidos os filtros CFS e FCBF pois não foram capazes de identificar simultaneamente todas características relevantes em nenhum dos conjuntos de dados com interações de alta ordem. Assim, na tabela são apresentados dois valores para os resultados dos filtros ReliefF e INTERACT. O valor entre parênteses representa o resultado quando o filtro

encontrou pelo menos uma das características relevante e o outro resultado refere-se a acurácia do filtro em identificar corretamente todas as características relevantes.

Tabela 5.3 Acurácia dos filtros ReliefF e INTERACT com conjuntos de dados com interações de alta ordem, com 3, 4 e 5 interações. As abreviações: R (ReliefF), I (INTERACT), Pop. (População).

Interações	Pop.	Acurácia (%)					
		20 SNPs		100 SNPs		1000 SNPs	
		R	I	R	I	R	I
3	800	100 (100)	53 (98)	1 (67)	0 (16)	0 (7)	0 (1)
	1600	100 (100)	69 (99)	9 (94)	0 (8)	0 (5)	0 (0)
4	800	98 (100)	31 (100)	1 (61)	0 (16)	0 (7)	0 (0)
	1600	100 (100)	50 (100)	1 (86)	0 (6)	0 (9)	0 (0)
5	800	32 (98)	17 (100)	0 (56)	0 (27)	0 (7)	0 (1)
	1600	98 (100)	25 (100)	0 (64)	0 (10)	0 (8)	0 (0)

Em parênteses estão os resultados quando encontrados ao menos uma atributo relevante.

Dentre os filtros testados, o ReliefF mostrou-se novamente o filtro mais promissor, sendo seus resultados estatisticamente melhores em comparação com os demais filtros, no teste T-Student a um nível de significância de 5%, com a rejeição da hipótese nula $H_0 : \mu_0 = \mu_1$, seguido em desempenho pelo INTERACT.

Os demais filtros, FCBF e CFS utilizam a mesma métrica: *Symmetrical uncertainty* para avaliar os SNPs relevantes, e este pode ser o motivo de ambas terem tido desempenho semelhante ao lidarem com estes dados.

Embora os filtros multivariados testados tenham sido propostos para detectar correlação entre variáveis nos conjuntos de dados, os seus desempenhos são afetados com o aumento do número de características relevantes. À medida que a quantidade dessas características cresce, há um decréscimo acentuado no desempenho dos filtros. Este resultado é um problema muito significativo, pois desejamos filtrar SNPs relevantes em conjuntos de dados de GWAS, que costumam ser conjuntos de dados com alta dimensionalidade e com interações entre os SNPs, o que de fato os torna ineficazes para serem aplicados a este tipo de problema.

5.2 Algoritmos Baseados em RLVQ

Com o objetivo de avaliar o desempenho dos algoritmos baseados em *Relevance Learning Vector Quantization* (RLVQ) em lidar com o problema da inferência das interações entre os SNPs, foram inicialmente selecionados os algoritmos GRLVQ e o SRNG. Como ponto de partida, na Seção 5.2.1 esses algoritmos são avaliados utilizando diferentes conjuntos de dados e um conjunto de parâmetros padrão, ou seja parâmetros comumente utilizados na literatura.

Em seguida, na Seção 5.2.2, são apresentados experimentos realizando uma busca de

parâmetros ideais. Para isso, são utilizados intervalos amplos de parâmetros[1], de forma a identificar, quais os parâmetros que causam uma maior influencia na acurácia dos algoritmos. Na Seção 5.2.3 são apresentados os tempos de processamento dos algoritmos ao lidar com os diferentes conjuntos de dados.

5.2.1 Experimentos Preliminares

Para os experimentos preliminares com o GRLVQ e SRNG, foi utilizado um conjunto de parâmetros padrão, apresentado na Tabela 5.4. Os valores selecionados dos parâmetros são baseados em experimentos anteriormente realizados em Hammer e Villmann (2002) e Araujo *et al.* (2013).

Foram utilizados os mesmos valores de parâmetros para ambos algoritmos, excetuando-se o número de protótipos que foi ajustado por tentativa e erro, de acordo com o tipo de conjunto de dados fornecido. Com isso, foi possível verificar se há diferenças entre a presença ou ausência do uso da vizinhança para inferência das interações entre o GRLVQ e SRNG.

Tabela 5.4 Parâmetros padrões utilizados no GRLVQ e SRNG.

Parâmetros		**Algoritmos**	
Sigla	**Descrição**	**GRLVQ**	**SRNG**
NNODES	Número de Protótipos	4^a e 20^b	4^a e 20^b
EPOCHS	Número de Interações	1000	1000
TAW	T.A. do Vetor de Relevâncias	0.01	0.01
TAP	T.A. Positiva	0.1	0.1
TAN	T.A. Negativa	0.05	0.05
TAU	Taxa de Decaimento	linear	linear
Gamma	Cooperação da Vizinhança	-	0.995

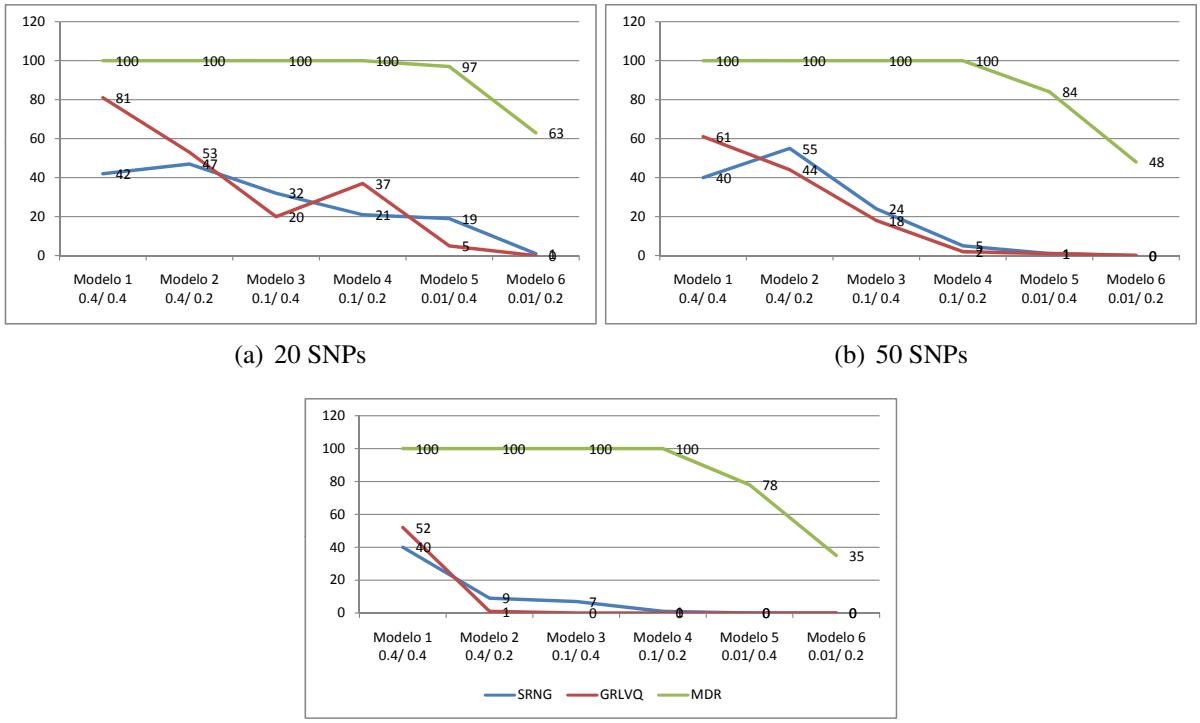
^{a} 4 protótipos utilizados em dados com duas interações e ^{b} 20 protótipos em dados com interações de alta ordem. T.A. significa Taxa de Aprendizagem.

Os mesmos conjuntos de dados utilizados para avaliar os filtros na Seção 5.1, também são utilizados para avaliar o GRLVQ e o SRNG. Os parâmetros utilizados para cada conjunto de dados diferenciam-se apenas, na quantidade de protótipos que são inseridos no mapa.

Como mostrado na Tabela 5.4, os experimentos com o conjunto com duas interações foram realizados utilizando 4 protótipos, enquanto que o conjunto com interações de alta ordem com 20 protótipos. Esses valores foram definidos a partir de testes de tentativa e erro, nos quais, se observou que conjuntos com mais SNPs interagindo necessitam de um maior número de protótipos.

[1]Os intervalos de parâmetros amplos se referem a seleção de valores mínimos e máximos para cada parâmetro utilizado no algoritmo.

Os resultados dos experimentos realizados com o GRLVQ e SRNG, utilizando o conjunto de dados com interações de ordem 2 (Velez *et al.*, 2007), são apresentados na Figura 5.1. Comparativamente, os mesmos conjuntos de dados é utilizando pela ferramenta MDR.



(a) 20 SNPs

(b) 50 SNPs

(c) 100 SNPs

Figura 5.1 Gráficos da acurácia do GRLVQ, SRNG e MDR ao analisar conjuntos de dados com 2 interações com 20, 50 e 100 SNPs

O MDR, por ser uma ferramenta exaustiva, investiga todas as interações de ordem 2 existente nos dados, e retorna como resultado, a interação que possui a maior relevância. Ao analisar os resultados obtidos, constatamos que o MDR obteve excelentes resultados, chegando a alcançar 100% de acerto nos modelos 1, 2, 3 e 4, representados por modelos com valor da herdabilidade 0.4 e 0.1.

O percentual máximo de acerto alcançado pelo GRLVQ foi de 81% no modelo 1 com herdabilidade 0.4 e MAF 0.4 e máximo alcançado pelo SRNG foi de 55% no modelo 2 com herdabilidade 0.4 e MAF 0.2. Os modelos 5 e 6, são os modelos com o mais baixo valor para a herdabilidade 0.01, sendo uma amostra difícil de identificar os SNPs relevantes, mesmo para uma ferramenta exaustiva como o MDR. Em Araujo e Guimaraes (2011), são apresentados resultados que mostram que esta deficiência ocorre em diversas outras ferramentas de inferência de epistasia, pois o valor da herdabilidade indica uma correlação muito baixa entre o coeficiente genético e o fenótipo.

Em um segundo experimento utilizando os mesmos parâmetros padrão, foram feitos experimentos com conjuntos de dados com interações de alta ordem. Os resultados obtidos são apresentados na Tabela 5.5 e mostram que a acurácia desses algoritmos é afetada pelo o número

de interações relevantes e pela quantidade de SNPs no conjunto de dados. A medida que estas variáveis aumentam, a acurácia é afetada.

Tabela 5.5 Acurácia do GRLVQ e SRNG com dados com interações de alta ordem.

Interações	Acurácia (%)					
	GRLVQ			SRNG		
	20 SNPs	50 SNPs	100 SNPs	20 SNPs	50 SNPs	100 SNPs
3	80	73	50	41	31	28
4	67	48	1	45	38	11
5	8	0	0	12	8	0

Após a obtenção desses resultados, levantou-se a hipótese de qual seria a real capacidade e limites desses algoritmos. Assim, fez-se necessário realizar um estudo sobre a influência da variação paramétrica em função da acurácia.

5.2.2 Calibração dos Parâmetros

Para investigar as reais capacidades e limitações dos algoritmos baseados em RLVQ. Foi proposta uma metodologia de busca paramétrica mais objetiva, sendo definidos intervalos de valores, ao invés de um valor fixo, para cada um dos parâmetros utilizado nos algoritmos.

Para a seleção dos intervalos de parâmetros, foram considerados os valores de parâmetros padrão, como ponto médio do intervalo, de forma a inclui-los dentro dos intervalos. O intervalo de parâmetros foi nomeado de **intervalo de parâmetros amplos** e os valores utilizados são descritos na Tabela 5.6.

Tabela 5.6 Intervalo dos Parâmetros Amplos utilizados no GRLVQ e SRNG

Parâmetros		Intervalo dos Parâmetros Amplos	
Sigla	**Descrição**	**GRLVQ**	**SRNG**
NNODES	Número de protótipos	[2 a 30]	[2 a 30]
EPOCHS	Número de Iterações	[500 a 10.000]	[500 a 10.000]
TAW	T.A. do Vetor de Relevâncias	[0.01 a 0.5]	[0.01 a 0.5]
TAP	T.A. Positiva	[0.01 a 0.5]	[0.01 a 0.5]
TAN	T.A. Negativa	[0.0025 a 0.25]	[0.0025 a 0.25]
TAU	Taxa de Decaimento	$[1\times10^{-5}$ a $1\times10^{-2}]$	$[1\times10^{-5}$ a $1\times10^{-2}]$
Gamma	Cooperação da Vizinhança	-	[0.001 a 1]

Apenas o SRNG possui o parâmetro da Cooperação da Vizinhança. T.A. significa Taxa de Aprendizagem.

A partir do intervalo de valores de cada parâmetro são selecionadas 100 amostras de valores. A seleção dessas amostras é realizada com o uso de um método estatístico denominado *Latin Hypercube Sampling* (LHS) (McKay *et al.*, 1979). O LHS garante a cobertura completa do intervalo de cada parâmetro.

Especificamente, o intervalo de cada parâmetro é dividido em 100 subintervalos de igual probabilidade e um único valor é selecionado aleatoriamente dentro de cada subintervalo. Os 100 diferentes valores obtidos para cada parâmetro são combinados entre si, e então armazenados em uma tabela para uso posterior no treinamento do algoritmo.

Utilizando a tabela de parâmetros, foram testados conjuntos de dados simulados com 100 diferentes amostras contendo 20 SNPs e três interações relevantes. Para cada uma das amostras, foram realizados 100 testes, utilizando os 100 diferentes valores de parâmetros armazenados. Executando os algoritmos com 100 diferentes amostras de parâmetros, foi possível avaliar a influência dos parâmetros sobre o desempenho dos algoritmos.

Métricas de Avaliação

Para avaliar a *performance* dos métodos, três diferentes métricas são comparadas e utilizadas nesta tese. Duas já descritas e utilizadas na literatura (Yang *et al.*, 2009; Shang *et al.*, 2011) e uma nova métrica proposta nesta tese.

A primeira métrica Power 1 é a forma mais comumente utilizada para avaliar o desempenho dos algoritmos. Ao realizar a inferência das interações, é apresentado um *ranking* das interações dos SNPs mais relevantes identificadas nos conjuntos de dados. Nos casos em que, os SNPs de fato mais relevantes são identificados na posição inicial do *ranking*, o contador da Power 1 é incrementada com $+1$, sendo considerado um acerto, e com zero caso não sejam identificados **todos** os SNPs relevantes. Essa é a métrica mais restritiva, pois considera acerto ou erro do algoritmo baseado na identificação de todos os SNPs nas primeiras posições do *ranking*. Assim, em estudos que realizam testes com 100 amostras diferentes, a pontuação máxima obtida será 100, caso todos as interações relevantes sejam identificadas.

Uma segunda métrica, Power 2, considera o número de SNPs pertencentes à interação que foram identificados. Por exemplo, caso um conjunto de dados possua três SNPs relevantes, mas no *ranking* surgiram apenas dois ou um nas primeiras posições, esse sucesso deve ser contabilizado com um incremento de 2 ou 1, respectivamente. Assim a Power 2 pode avaliar com um pouco mais de detalhe o desempenho do algoritmo mesmo quando não forem identificados todos os SNPs relevantes.

A terceira métrica *Average Ranking* (AR), proposta nesta tese é apresentada em Araujo e Guimaraes (2016), e considera o *ranking* médio da posição dos SNPs relevantes ao longo de todo o vetor de relevâncias. A métrica AR é calculada pela equação (5.1) e seu resultado varia de 0 a 1, onde 1 informa que todos os SNPs relevantes foram identificados nas primeiras posições do *ranking*, tendo seu valor gradualmente tendendo a zero a medida que os SNPs relevantes estão em posições mais baixas no *ranking*, sendo 0 quando estão nas últimas posições do *ranking*.

$$AR = 1 - \frac{\left(2 \times \sum_{i=0}^{R-1} rank[i]\right) - R \times (R-1)}{2DR - 2R^2} \tag{5.1}$$

onde, R é o número de interações relevantes no conjunto de dados; $rank[i]$ é o vetor de índices no qual os SNPs relevantes estão ordenados de acordo com sua relevância e D é o número de dimensões, ou seja, de SNPs no conjunto de dados.

A vantagem da métrica AR é que esta oferece uma maior riqueza de detalhes sobre a acurácia do algoritmo ao longo dos ajustes paramétricos, permitindo uma visualização gradual da acurácia em função dos diferentes valores de parâmetros utilizados.

Resultados com Varredura Ampla dos Parâmetros

Ao utilizar os parâmetros amplos, o GRLVQ apresentou uma média de acerto de 27%, o que significa dizer que dentre as 100 amostras testadas, o algoritmo identificou os três SNPs relevantes no topo da lista do *ranking*. Neste mesmo experimento, o número máximo de acertos foi de 49 amostras, o mínimo de 10 amostras e desvio padrão de 8,4.

O SRNG, assim como o GRLVQ, foi testado utilizando os parâmetros amplos, apresentando uma média de acerto de 44% (desvio de 13,2), sendo estatisticamente melhor em teste T-Student a um nível de significância de 5%, com rejeição da hipótese nula $H_0 : \mu_0 = \mu_1$, quando comparado com o GRLVQ. Neste mesmo experimento, o SRNG obteve um máximo de 75 acertos (GRLVQ: 49 acertos), mínimo de 14 (GRLVQ: 10 acertos) e desvio padrão de 13,2 (GRLVQ: 8,4).

Tabela 5.7 Avaliação dos Parâmetros Amplos do GRLVQ com conjunto de dados com interações entre 3 SNPs.

Acurácia	**Parâmetros**					
	NNODES	**TAP**	**TAN**	**TAW**	**TAU**	**EPOCHS**
100	15	0,45	0,06	0,26	0,0001	5008
99	13	0,30	0,02	0,37	0,0004	5536
95	7	0,48	0,09	0,44	0,0013	3272
90	25	0,23	0,06	0,42	0,0004	817
96	**15**	**0,37**	**0,06**	**0,37**	**0,0006**	**3658**
(4,5)	**(7,5)**	**(0,11)**	**(0,02)**	**(0,08)**	**(0,0005)**	**(2746)**
1	4	0,18	0,0035	0,04	0,0083	7441
3	20	0,30	0,067	0,03	0,0089	4907
3	29	0,08	0,015	0,29	0,0025	6122
3	28	0,09	0,015	0,12	0,0077	2856
2,5	**20**	**0,16**	**0,025**	**0,12**	**0,0073**	**5332**
(1)	**(8,2)**	**(0,08)**	**(0,02)**	**(0,08)**	**(0,002)**	**(1450)**

Médias dos parâmetros em negrito e desvio padrão entre parênteses. Em azul e em vermelho são as médias dos valores dos parâmetros mais relevantes, no melhor e pior caso, respectivamente.

Observando na Tabela 5.7, os resultados em função dos parâmetros, temos que, no GRLVQ, o melhor conjunto de parâmetros selecionado dentro do intervalo acertou todas as

amostras testadas. Isso, significa dizer que existe um único conjunto de parâmetros capaz de lidar com todos os modelos do conjunto de dados, e identificar corretamente todas as interações relevantes. Enquanto, o conjunto de parâmetros com o pior desempenho acertou apenas 1 amostra, apresentando uma média de acerto de 27 e desvio padrão de 25,2.

Na Tabela 5.7, também são apresentados os melhores ($\geq$90% de acerto exibido na parte superior da tabela) e piores ($\leq$3% de acerto exibido na parte inferior da tabela) parâmetros identificados para o GRLVQ. Considerando esses resultados, foi observado que três dos seis parâmetros utilizados no GRLVQ, o TAP, TAW e TAU, destacam-se por apresentar os valores mais divergentes dentre todos os parâmetros, com os melhores e piores resultados. Isto significa dizer que eles influenciam mais fortemente no desempenho dos modelos.

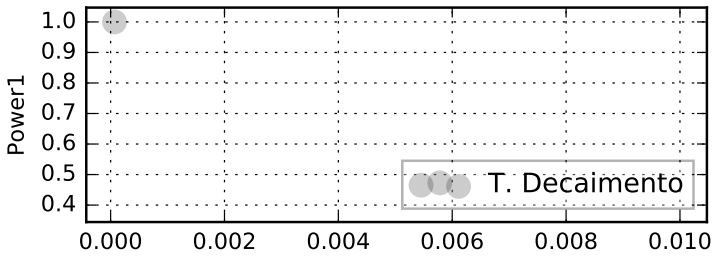
Considerando o parâmetro TAU (taxa de decaimento) aplicamos as três diferentes métricas: Power 1, Power 2 e *Average Ranking* (AR), já descritas na Seção 5.2.2, e as acurácias obtidas com diferentes representações são apresentadas nos gráficos da Figura 5.2. Através da métrica AR, é possível visualizar com mais clareza o comportamento da acurácia do GRLVQ em função da taxa de decaimento (parâmetro TAU). Nessa situação, observa-se que uma taxa de decaimento menor, favorece um melhor desempenho do algoritmo.

Utilizando a métrica AR, são apresentados os resultados do GRLVQ nos gráficos da Figura 5.3, exibindo o *ranking* médio dos SNPs relevantes em função dos seis parâmetros utilizados no GRLVQ. A posição dos círculos cinza mais ao topo do gráfico (no eixo y), indica os melhores resultados de acordo com o valor dos parâmetros (no eixo x). As regiões destacadas em vermelho indicam intervalos de interesse para cada um dos parâmetros, gráficos sem destaques não foram identificadas tendências por parte do parâmetro estudado. Com essas observações, esses parâmetros em destaque tornam-se os principais alvos de ajustes para adequar o algoritmo de inferência das interações.

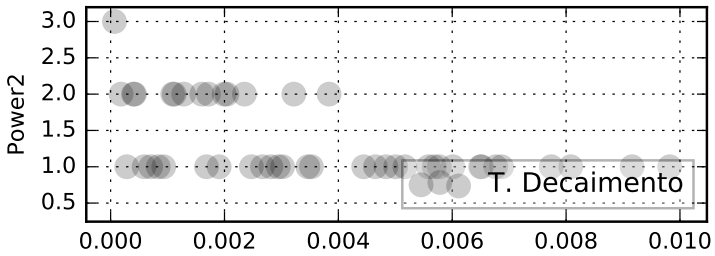
Na Tabela 5.8 são apresentados os melhores ($>$50% de acerto exibido na parte superior da tabela) e piores ($\leq$37% de acerto exibido na parte inferior da tabela) parâmetros identificados para o SRNG. O melhor resultado obtido apresentou um acerto de apenas 58% das amostras testadas (GRLVQ: 100% das amostras), enquanto o conjunto de parâmetros com o pior resultado apresentou um acerto de 31% (GRLVQ apenas 1%) com desvio padrão de 4,9 (GRLVQ: 25,2).

Com esses resultados, constata-se que dentre o intervalo de parâmetros utilizado, o GRLVQ foi capaz de obter o melhor resultado possível, sendo este resultado muito superior ao apresentado pelo SRNG. Um outro aspecto observado, está relacionado com os desvios padrões obtidos pelo GRLVQ e SRNG. O GRLVQ teve um desvio padrão de 25,2, enquanto o SRNG de apenas 4,9. Este resultado indica que a acurácia do GRLVQ é mais sensível ao ajuste dos parâmetros do que o SRNG, o que o torna mais difícil de ter seus parâmetros ajustados.

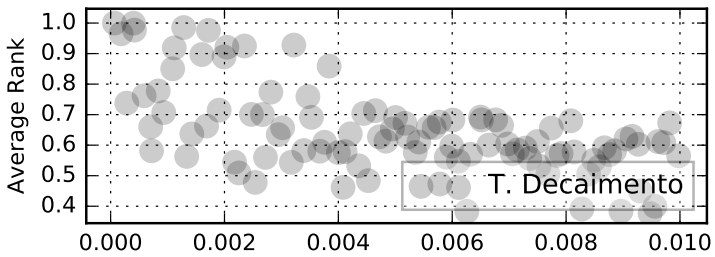
Considerando os resultados da Tabela 5.8, observamos que o SRNG mostrou uma dependência entre três parâmetros a **cooperação da vizinhança**, **número de protótipos** e a **taxa de aprendizagem positiva**. Os primeiros experimentos para ajuste dos parâmetros consideraram o ajuste desses parâmetros. No primeiro experimento foi ajustado o intervalo do número de



(a) Power 1



(b) Power 2



(c) *Average Ranking*

Figura 5.2 Acurácia do GRLVQ em função da taxa de decaimento (TAU) utilizando as métricas Power 1, Power 2 e *Average Ranking* (AR) em conjunto de dados com 20 SNPs e três interações. Valores mais altos indicam resultados melhores.

protótipos de 2 a 30, para 20 a 50; os resultados obtidos são exibidos na Figura 5.4 e mostram, como esperado, que desempenho do algoritmo melhora à medida que o parâmetro da vizinhança aumenta.

Assim, no segundo experimento, ajustou-se o intervalo do parâmetro da vizinhança. Subsequentemente, novos ajustes foram realizados em outros parâmetrosm tais como a taxa de aprendizagem positiva (TAP), obtendo com isso, um intervalo de parâmetros ajustado, apresentado na Tabela 5.9.

Resultados com Varredura Ajustada dos Parâmetros

Após ajustes dos intervalos de parâmetros, foram realizados experimentos com ambos os algoritmos utilizando **intervalos dos parâmetros ajustados** descritos na Tabela 5.9 e os mesmos conjuntos de dados utilizados no experimento com os intervalos de parâmetros amplos.

Com este ajuste dos parâmetros, ambos os algoritmos tiveram melhoras significativas em sua acurácia. O GRLVQ, por exemplo, enquanto utilizava o intervalo de parâmetros amplos teve

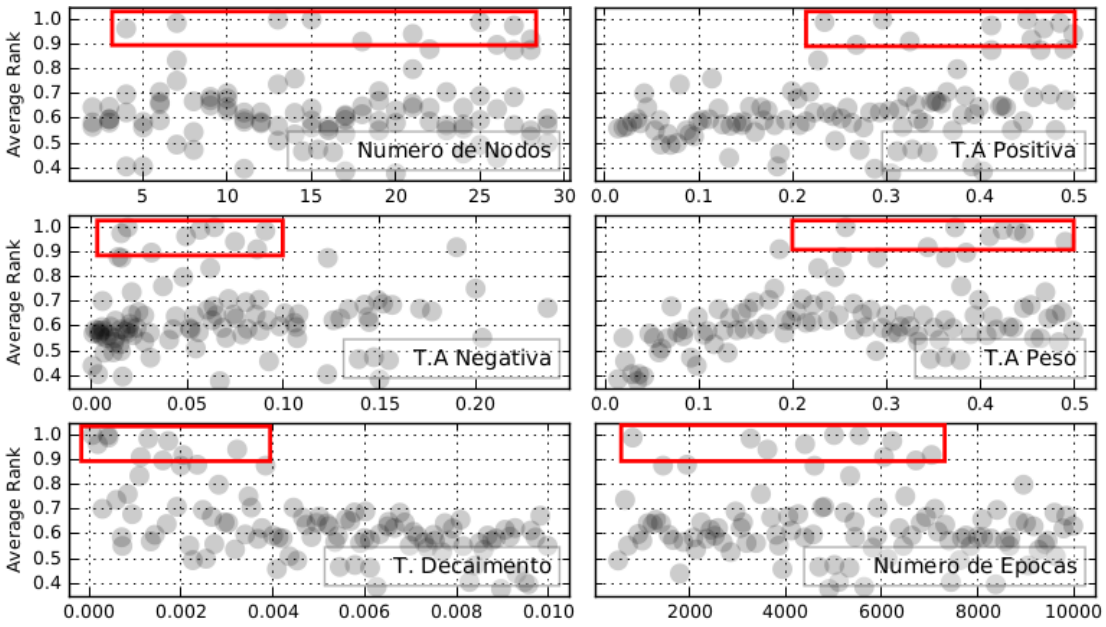


Figura 5.3 Gráfico comparativo da influência dos parâmetros com valores de intervalo amplos no desempenho do GRLVQ. As regiões destacadas em vermelho indicam regiões de interesse para novos intervalos paramétricos. T.A significa Taxa de Aprendizagem.

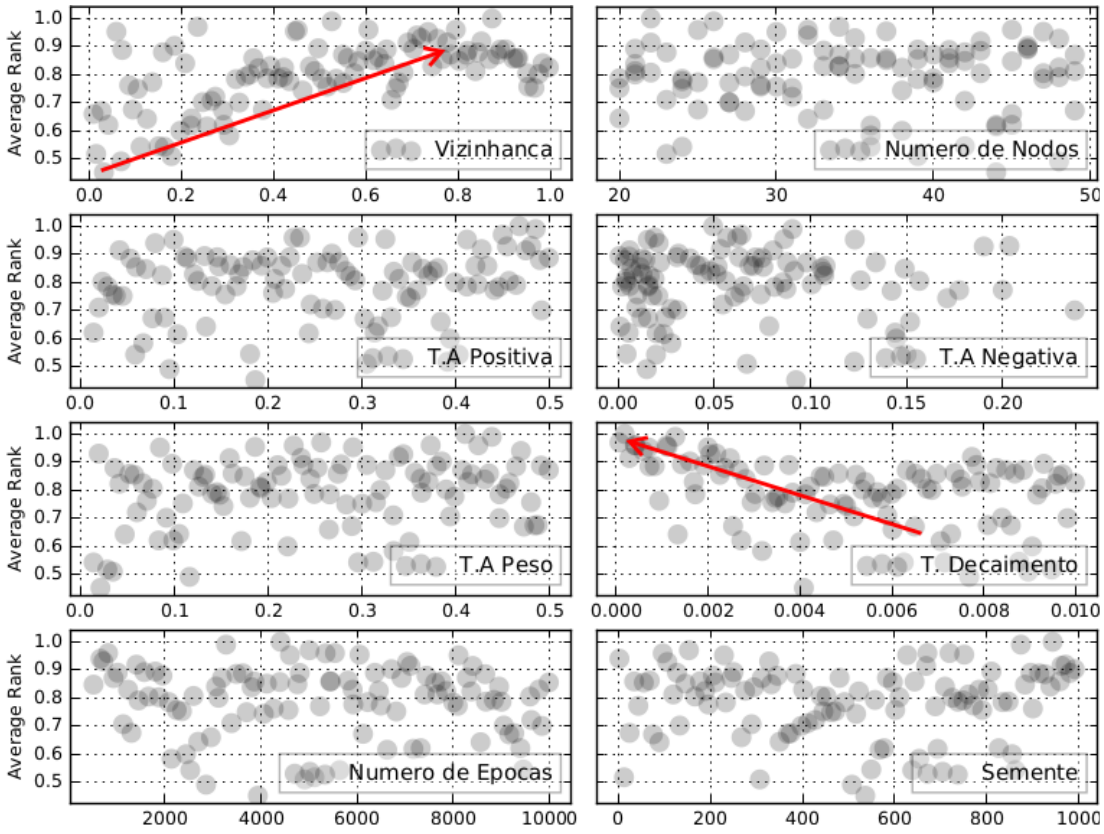


Figura 5.4 Gráfico comparativo da influência dos parâmetros no desempenho do SRNG. A seta em vermelho sobre os resultados, indica uma tendência dos parâmetros da Vizinhança e Taxa de Decaimento sobre a acurácia do SRNG.

Tabela 5.8 Avaliação dos Parâmetros Amplos do SRNG com conjunto de dados com interações entre 3 SNPs.

Acurácia	**Parâmetros**						
	GAMMA	**NNODES**	**TAP**	**TAN**	**TAW**	**TAU**	**EPOCHS**
58	0,95	24	0,38	0,14	0,40	0,0001	2648
53	0,75	16	0,21	0,01	0,28	0,0100	7191
53	0,41	6	0,15	0,06	0,14	0,0050	5957
52	0,35	24	0,10	0,02	0,17	0,0065	1862
52	0,91	29	0,06	0,02	0,18	0,0034	3026
51	0,54	27	0,01	0,04	0,16	0,0024	7472
53	**0,65**	**21**	**0,15**	**0,04**	**0,22**	**0,005**	**4693**
(2,5)	**(0,3)**	**(8,6)**	**(0,1)**	**(0,05)**	**(0,1)**	**(0,003)**	**(2471)**
37	0,28	7	0,20	0,01	0,25	0,0040	506
37	0,92	10	0,43	0,08	0,14	0,0099	2306
37	0,14	26	0,09	0,02	0,49	0,0054	2729
37	0,79	24	0,46	0,01	0,02	0,0050	7049
36	0,06	9	0,29	0,12	0,21	0,0047	6716
36	0,32	3	0,30	0,01	0,43	0,0049	5611
36	0,12	14	0,17	0,02	0,44	0,0039	3368
36	0,78	28	0,14	0,06	0,06	0,0075	5267
35	0,53	18	0,50	0,24	0,47	0,0056	5213
35	0,60	26	0,49	0,18	0,17	0,0055	8907
31	0,56	10	0,02	0,01	0,46	0,0090	4981
35,7	**0,46**	**15,9**	**0,3**	**0,07**	**0,3**	**0,006**	**4786**
(1,7)	**(0,3)**	**(8,9)**	**(0,2)**	**(0,08)**	**(0,2)**	**(0,002)**	**(2401)**

Médias dos parâmetros em negrito e desvio padrão entre parênteses. Em azul e vermelho são as diferenças significativas entre as médias do melhor e pior caso.

Tabela 5.9 Intervalo dos Parâmetros Ajustados utilizados no GRLVQ e SRNG

Parâmetros		**Intervalo dos Parâmetros Amplos**	
Sigla	**Descrição**	**GRLVQ**	**SRNG**
NNODES	Número de protótipos	[6 a 16]	[20 a 50]
EPOCHS	Número de Épocas	[500 a 10.000]	[500 a 10.000]
TAW	T. A. do Vetor de Relevâncias	[0.15 a 0.2]	[0.15 a 0.5]
TAP	T. A. Positiva	[0.4 a 0.5]	[0.2 a 0.5]
TAN	T. A. Negativa	[0.01 a 0.05]	[0.025 a 0.1]
TAU	Taxa de Decaimento	$[1\times10^{-5}$ a $9\times10^{-5}]$	$[1\times10^{-5}$ a $2\times10^{-3}]$
GAMMA	Cooperação da Vizinhança	-	[0.5 a 1]

Apenas o SRNG possui o parâmetro da Cooperação da Vizinhança. T.A. significa Taxa de Aprendizagem.

uma acurácia média de apenas 27% e desvio padrão 8,4 e o SRNG 58% e desvio padrão 13,2. Ao realizar a calibração dos parâmetros, o GRLVQ alcançou uma acurácia média de 99,4% (3,1) e o SRNG 94,7% (6,6).

Na Tabela 5.10, são apresentadas as médias dos resultados com o GRLVQ e SRNG, variando o número de interações e o número de SNPs no conjunto de dados. Nesses resultados, observamos que ambos os algoritmos, tiveram dificuldade em identificar as interações com o aumento no número de SNPs. É possível que, o ajuste de parâmetros realizado não seja adequado para os conjuntos com maiores quantidades de SNPs.

Tabela 5.10 Comparação das médias obtidas pelo GRLVQ e SRNG após o ajuste paramétrico com conjuntos de dados com interações de alta ordem.

		Acurácia (%)					
Int.	**Pop.**	**20 SNPs**		**100 SNPs**		**1000 SNPs**	
		GRLVQ	**SRNG**	**GRLVQ**	**SRNG**	**GRLVQ**	**SRNG**
3	800	99.4 (3.06)	94.7 (6.64)	99.9 (0.33)	96.6 (4.54)	5.6 (6.27)	0.1 (0.18)
4	800	91.7 (14.56)	95.0 (6.48)	91.5 (14.84)	88.8 (12.53)	-	-
5	800	42.9 (21.98)	74.7 (19.07)	39.9 (21.92)	50.0 (29.78)	-	-

Int. significa número de interações. Entre parênteses são apresentados os desvios padrões. Os resultados com: - para os dados com 1000 SNPs e 4 e 5 interações não foram realizados.

Os resultados do experimento realizado com o GRLVQ pode ser visualizados na Figura 5.5. Nesse experimento foi utilizado o mesmo conjunto de dados dos resultados exibidos na Figura 5.3, de forma que podemos visualizar comparativamente a melhora no desempenho do GRLVQ após os ajustes dos parâmetros.

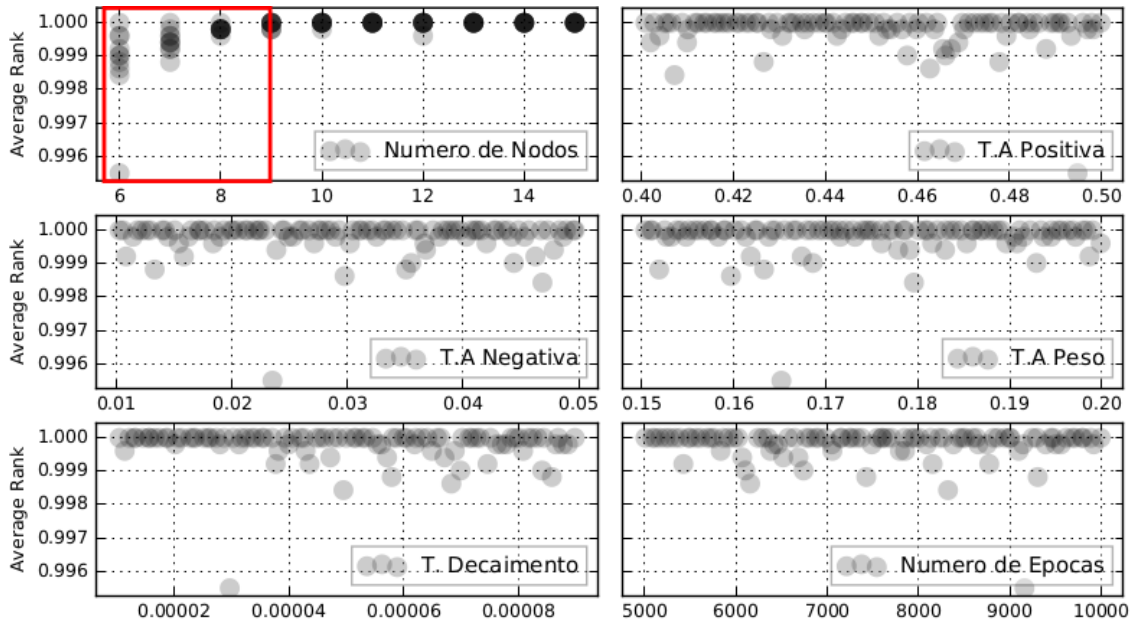
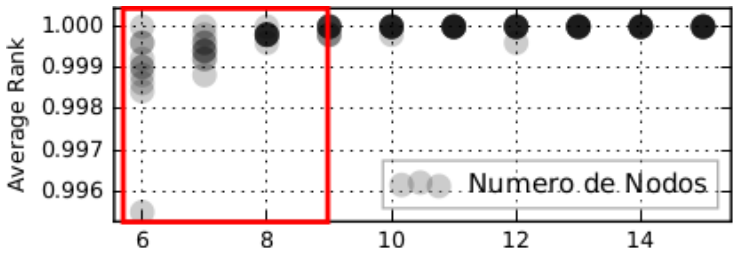


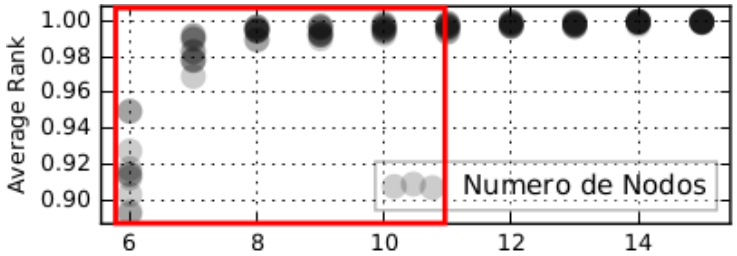
Figura 5.5 Gráfico comparativo da influência dos parâmetros ajustados no desempenho do GRLVQ.

Após verificar a melhora no desempenho do algoritmo, foi mantida a configuração paramétrica ajustada e foram realizados experimentos com conjuntos de dados com um maior número de interações (4 e 5 interações), no entanto, foi mantida fixa a quantidade de SNPs

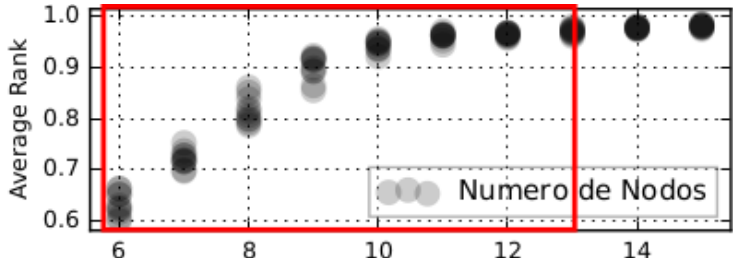
avaliados nos conjuntos de dados. Observando a Figura 5.6 são exibidos horizontalmente 3 conjuntos com 2 gráficos. Cada linha representa os resultados obtidos com conjuntos de dados com 3, 4 e 5 interações, sendo observado o parâmetro Número de protótipos destacado em vermelho.



(a) 3 interações



(b) 4 interações



(c) 5 interações

Figura 5.6 Gráfico comparativo do comportamento do parâmetro: Número de protótipos, em destaque, quando testados conjuntos de dados com 3, 4 e 5 interações pelo GRLVQ.

Considerando os resultados da Figura 5.6, nestes experimentos foram utilizados os mesmos parâmetros, sendo a diferença das acurácias correlacionada ao número de interações identificadas pelo algoritmo.

É observado um decréscimo no desempenho dos algoritmos, a medida que um maior número de interações estão presentes no conjunto de dados. Quando isto acontece, os gráficos mostram que se faz necessário utilizar um maior número de protótipos para detectar adequadamente todas as interações.

Esta observação, indica que uma inserção automática de protótipos poderia tornar o método menos sensível ao ajuste paramétrico, consequentemente gerando resultados melhores para uma gama maior de conjuntos de dados. Isto será tratado na Seção 5.3.

Ajustes Finos dos Intervalos de Parâmetros

Como mostrado na Tabela 5.10, vamos retomar nessa seção, o problema envolvendo o aumento do número de SNPs. Nesses experimentos, foram realizados ajustes finos nos parâmetros, ao utilizar conjuntos de dados com 1000 SNPs. O objetivo desses testes, é avaliar o desempenho dos algoritmos, ao lidar com o aumento do número de SNPs. Assim, na Tabela 5.11 é apresentado os resultados com e sem ajuste dos parâmetros.

Tabela 5.11 Comparação das médias obtidas pelo GRLVQ e SRNG antes e após o ajuste paramétrico fino com conjuntos de dados com 1000SNPs e interações de alta ordem.

Interações	**Acurácia (%)**			
	Antes do Ajuste		**Após o Ajuste**	
	GRLVQ	**SRNG**	**GRLVQ**	**SRNG**
3	5.6 (6.27)	0.1 (0.18)	94,9 (1,71)	12,5 (8,4)

O GRLVQ, diferentemente do SRNG, apresentou uma maior variabilidade diante dos diferentes intervalos de parâmetros, e por isso, foi possível verificar mais facilmente tendências em seus resultados, reveladas com o uso da métrica AR. Para a obtenção desses resultados, foram utilizados os intervalos de parâmetros descritos na Tabela 5.12.

Tabela 5.12 Parâmetros de Intervalos com ajustes para dados com 1000 SNPs.

Parâmetros		**Intervalo dos Parâmetros Amplos**	
Sigla	**Descrição**	**GRLVQ**	**SRNG**
NNODES	Número de protótipos	[20 a 50]	[15 a 30]
EPOCHS	Número de Épocas	[5000 a 10,000]	[5000 a 10,000]
TAW	T. A. do Vetor de Relevâncias	[0.15 a 0.5]	[0.15 a 0.2]
TAP	T. A. Positiva	[0.2 a 0.5]	[0.4 a 0.5]
TAN	T. A. Negativa	[0.025 a 0.1]	[0.01 a 0.05]
TAU	Taxa de Decaimento	$[5\times10^{-5}$ a $1\times10^{-5}]$	1×10^{-5} a 2×10^{-4}
GAMMA	Cooperação da Vizinhança	[0.6 a 0.9]	-

Apenas o SRNG possui o parâmetro da Cooperação da Vizinhança. T. A. significa Taxa de Aprendizagem

Para o GRLVQ, foi necessário o ajuste de três parâmetros: o número de épocas, sendo o valor mínimo do intervalo alterado de [500 a 10.000] para [5.000 a 10.000]; a taxa de decaimento de $[1\times10^{-5}$ a $9\times10^{-5}]$ para $[1\times10^{-5}$ a $2\times10^{-4}]$ e o número de protótipos também foi alterado de [6 a 16] para [15 a 30], no entanto, este último foi modificado apenas para adequar a complexidade do número de interações nos dados.

E para o SRNG, o resultado com melhor acurácia obtida, também teve modificações em três parâmetros, a cooperação da vizinhança alterado de [0.5 a 1] para [0.6 a 0.9], número de

épocas de [500 a 10.000] para [5.000 a 10.000] e a taxa de decaimento de [1×10^{-5} a 2×10^{-3}] para [5×10^{-5} a 1×10^{-5}].

Em Araujo e Guimaraes (2016), foram utilizados esses parâmetros para avaliar o GRLVQ com todo o conjunto de dados. Os resultados obtidos são apresentados na Tabela 5.13 utilizando duas métricas, média aritmética e o AR e mostram que mesmo em situações em que a média aritmética de acertos foi de 68,4 (dados com 1000 SNPs e 5 interações), o AR obteve um valor de 98,8. Isto revela que esses SNPs estão localizados em posições elevadas no vetor de relevâncias.

Tabela 5.13 Acurácia do GRLVQ em dados com 800 indivíduos e interações de alta ordem.

SNPs	**Métrica**	**Acurácia média das Interações**		
		3 int.	**4 int.**	**5 int.**
20	MA	99,7 (0,65)	99,6 (0,68)	93,3 (11,50)
	AR	99,9 (0,06)	99,9 (0,04)	99,6 (0,32)
100	MA	99,6 (0,68)	99,5 (3,45)	95,4 (10,27)
	AR	99,9 (0,02)	99,9 (0,04)	99,8 (0,25)
1000	MA	94,9 (1,71)	86,0 (3,36)	68,4 (5,22)
	AR	99,9 (0,49)	99,5 (1,49)	98,8 (2,52)

Em parênteses são apresentados os desvios padrão. Siglas MA: Média Aritmética; AR: *Average Ranking*.

5.2.3 Tempo de Processamento

Para avaliar o tempo de processamento dos algoritmos GRLVQ e SRNG, ambos foram executados em um mesmo ambiente computacional: Linux Ubuntu 14.04 64 bits, com processador Intel Xenon com 8 núcleos de 2.0 GHz e 16GB de memória RAM. Apesar do sistema multiprocessado, nenhum método de paralelização foi implementado nos algoritmos avaliados.

Tabela 5.14 Comparativo do Tempo de Processamento de uma amostra utilizando o GRLVQ, SRNG e MDR.

Int.	**Tempo de Processamento**								
	20 SNPs			**100 SNPs**			**1000 SNPs**		
	MDR	**GRLVQ**	**SRNG**	**MDR**	**GRLVQ**	**SRNG**	**MDR**	**GRLVQ**	**SRNG**
3	$\leq$1s	12s	30s	12,5s	27s	1min30s	6h	3min30s	10min30s
4	1,7s	12s	30s	5min45s	27s	1min30s	60d	3min30s	10min30s
5	4,5s	12s	30s	2h40min	27s	1min30s	12.000d	3min30s	10min30s

Int. significa número de interações.

Na Tabela 5.14 é exibido o tempo necessário para os métodos GRLVQ, SRNG e MDR realizarem a inferência dos SNPs com os diferentes conjuntos de dados testados. O GRLVQ ao ser executado com uma amostra com 20 SNPs e intervalo de parâmetros ajustado, demora cerca

de 12s, enquanto que o SRNG realiza o mesmo processo em 30s. Esta diferença de tempo é esperada, devido ao processamento extra necessário para o SRNG ajustar a vizinhança de cada protótipo vencedor.

Cada conjunto de dados possui 20, 100 e 1000 SNPs e cada um deles contém 100 amostras, onde cada amostra é analisada 1×100 vezes com os diferentes conjuntos de parâmetros gerados no LHS. Observando os resultados, verificamos que o GRLVQ é em média três vezes mais rápido que o SRNG. Adicionalmente, vale ressaltar que ambos os métodos possuem vantagens frente ao métodos combinatoriais de inferências de polimorfismos, como o MDR, pois seu custo computacional não depende diretamente da quantidade de interações presentes no conjunto de dados. No entanto, assim como os demais métodos, esses algoritmos são dependentes do número total de SNPs analisados e da quantidade de épocas utilizadas para avaliar os conjuntos de dados.

Considerando os resultados obtidos com os experimentos realizados com o GRLVQ e SRNG, exibidos na Figura 5.6, foi levantada a hipótese de que a inserção automática dos protótipos, pode ser uma alternativa interessante para adequar o número de protótipos ao número de interações presentes nos dados, fazendo com que a acurácia do GRLVQ não se deteriorando a medida que o número de interações aumenta.

A partir desta constatação, foi considerada a necessidade de aumentar o número de protótipos a medida que a complexidade dos dados cresce. Uma maneira que poderia ser adequada para tratar deste problema, é realizar a inserção automática dos protótipos. Propriedade já introduzida na variante do GRLVQ, proposta por Kietzmann *et al.* (2008), criando o algoritmo incremental GRLVQ (iGRLVQ). Experimentos utilizando os dados de polimorfismos foram realizados para avaliar essa metodologia e são apresentados Seção 5.3.

5.3 Experimentos com iGRLVQ

Nesta seção, será utilizado o algoritmo do iGRLVQ para verificar o desempenho de uma metodologia incremental na detecção das interações entre os SNPs. Neste experimento, o iGRLVQ foi avaliado utilizando o mesmo conjunto de parâmetros apresentado na Tabela 5.12 e ajustado para o GRLVQ. Foram avaliados os conjuntos de dados de alta ordem e o resultado obtido é exibido na Tabela 5.15. É observado que o desempenho do iGRLVQ, mesmo com a metodologia incremental deteriora-se com o número de interações, embora seja mais fortemente afetado pelo número de SNPs avaliados. Também foi observado que mesmo nessa situação o algoritmo convergia corretamente e a acurácia da classificação dos dados era alta em torno de 70-90% de acerto, ou seja, o algoritmo consegue separar adequadamente os pacientes casos e controles, no entanto os SNPs relevantes não foram adequadamente identificados.

Assim, avaliando esses resultados, foi levantada a hipótese de que a causa da não identificação dos SNPs relevantes esteja associada a posição de inserção dos protótipos no espaço dos dados. Foi verificado que a metodologia de inserção de novos protótipos utilizada no

Tabela 5.15 Acurácia dos métodos com diferentes modos de inserção de protótipos com dados de alta ordem.

Métodos	Acurácia(%)								
	20 SNPs			100 SNPs			1000 SNPs		
	3 int.	4 int.	5 int.	3 int.	4 int.	5 int.	3 int.	4 int.	5 int.
iGRLVQ^K	99	96	91	100	100	49	0	0	0

iGRLVQ^K proposto por Kietzmann *et al.* (2008). int. significa número de interações.

algoritmo do iGRLVQ é extremamente útil e eficaz para solucionar problemas de classificação dos dados, já que a acurácia da classificação apresentava-se $\leq 70\%$.

No entanto, o problema tratado não apenas está relacionado a classificação dos dados, mas também a seleção de características relevantes dos dados. Por isso, a inserção dos protótipos nas regiões de limites entre as classes, como proposto por Kietzmann *et al.* (2008), pode ser eficiente para separar os dados em diferentes classes, mas não é eficiente para identificar as interações de SNPs em dados com mais de 1000 SNPs, como mostra a Tabela 5.15.

Considerando avaliar a hipótese da inserção dos protótipos, serão apresentados na proxima seção, experimentos utilizando três diferentes propostas de inserção de novos protótipos.

5.3.1 Experimentos com a Inserção de Protótipos

Considerando que a hipótese de que forma de inserção dos protótipos no iGRLVQ, foi o principal problema na identificação das interações entre SNPs relevantes, foram então realizados experimentos utilizando duas diferentes propostas de inserção de novos protótipos. A primeira proposta realiza a inserção dos novos protótipos inserindo-os em posições aleatórias no espaço de dados, enquanto que na segunda proposta, a inserção é realizada sobre os protótipos de classes opostas, como feito no algoritmo do SGNG por Garcia e Forster (2012). Ambos os procedimentos de inserção são comparados com o proposto por Kietzmann *et al.* (2008), utilizando os mesmos conjuntos de parâmetros.

Os resultados obtidos são apresentados na Tabela 5.16, dentre as três formas de inserção avaliadas, a inserção dos novos protótipos de acordo com o algoritmo do SGNG, foi a que obteve os melhores resultados. E constata-se, de acordo com a hipótese levantada que a forma de inserção utilizada pelo o iGRLVQ^K não é adequada para o problema, pois mesmo uma forma de inserção de novos protótipos de forma aleatória (iGRLVQ^R) obteve melhores resultados.

5.3.2 Convergência

Além de considerar a forma como os protótipos são inseridos no mapa, também se mostrou necessário verificar a convergência do algoritmo em função das mudanças no vetor de relevâncias. Como vimos anteriormente, uma definição adequada do número de épocas é

Tabela 5.16 Acurácia dos métodos com diferentes modos de inserção de protótipos com dados de alta ordem.

Métodos	Acurácia(%)								
	20 SNPs			100 SNPs			1000 SNPs		
	3 int.	4 int.	5 int.	3 int.	4 int.	5 int.	3 int.	4 int.	5 int.
iGRLVQ^K	99	96	91	100	100	49	0	0	0
iGRLVQ^R	100	100	96	100	100	100	57	45	13
iGRLVQ^S	100	100	98	100	100	100	100	100	81

iGRLVQ^K proposto por Kietzmann *et al.* (2008), iGRLVQ^R inserção dos protótipos em posições aleatórias e iGRLVQ^S inserção dos protótipos sobre protótipos de classe oposta. int. signfica número de interações.

importante para identificar corretamente os SNPs relevantes, o que está fortemente correlacionado com a convergência do algoritmo. Assim no GRLVQ, como proposto por Hammer e Villmann (2002), o treinamento desse algoritmo se dá até que a sua convergência seja alcançada, dada pela redução das mudanças ocorridas nos protótipos ou até que o número limite de épocas seja alcançado.

Considerando a relevância do vetor de pesos para identificar os SNPs relevantes, verificou-se a sua convergência como forma de assegurar a identificação dos SNPs relevantes. Com isso, é levantada a hipótese de que quando a convergência do vetor do relevâncias não é alcançada, o algoritmo pode ainda assim classificar corretamente os dados em casos e controles, a partir da identificação de SNPs falso positivos. Isso ocorre especialmente quando a quantidade de dados é alta. Assim, para garantir o melhor resultado a ser alcançado, é preciso considerar também o decaimento do vetor de pesos em vez de apenas a convergência em função das variações dos protótipos no mapa.

Além disso, o uso do vetor de pesos pode garantir uma otimização no tempo de execução do algoritmo, ao verificar sua convergência ao final de cada época. Isto evita que o método seja finalizado antes de atingir a convergência e ao mesmo tempo evita que sejam desperdiçados recursos computacionais ao continuar o treinamento mesmo após a convergência.

O decaimento da atualização média realizada no vetor de pesos e dos protótipos são calculados a cada época de treinamento do algoritmo, e é dada pelo somatório do quadrado da soma das raízes de todas as mudanças ocorridas nesses vetores. Observando o gráfico da Figura 5.7, o experimento utilizando parâmetros ajustados mostra que o decaimento da atualização média no vetor de pesos é mais lento que o decaimento da atualização média nos protótipos a medida que o número de épocas cresce. Ao alcançar um limiar mínimo de mudanças no vetor de relevâncias, considera-se que o algoritmo convergiu tanto em função dos protótipos como também em função do vetor de relevâncias. Essa estratégia traz uma maior eficiência ao algoritmo na identificação dos SNPs relevantes, reduzindo o número de falsos positivos identificados na classificação dos dados.

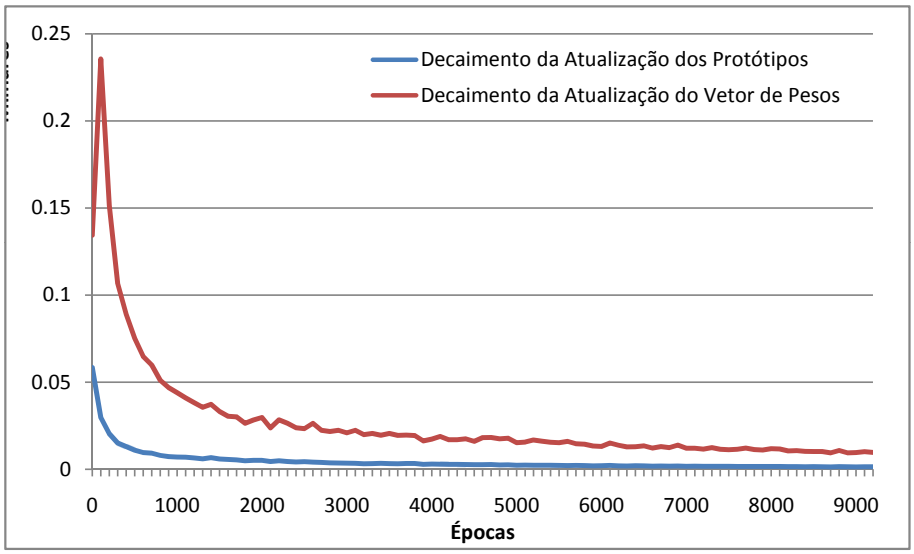


Figura 5.7 Taxa de decaimento do algoritmo em função da atualização média das variações nos protótipos e do vetor de pesos.

Utilizando os mesmos dados anteriores, foi observado com mais detalhes o comportamento do vetor de relevâncias durante o processo de aprendizagem. No gráfico apresentado na Figura 5.8 foi observado como ocorriam as mudanças dentro do vetor de relevâncias, em função das diferenças entre os SNPs que são relevantes e os demais. Na Figura 5.8 observa-se um efeito do distanciamento entre a relevância média atribuídas aos SNPs relevantes em comparação com a média atribuída aos SNPs não relevantes, acompanhada pela diminuição da variação da relevância para os SNPs relevantes no vetor de pesos.

Considerando que o conjunto de dados utilizado era composto por 1000 SNPs e 5 interações entre SNPs. Ao inicializar o vetor de relevâncias, a relevância máxima de cada SNP é de $\frac{1}{1000} = 0,001$. Sabendo-se que no conjunto de dados, utilizado neste experimento, o número de SNPs relevantes é igual a 5, a curva de convergência com as médias dos SNPs relevantes tende a convergir a um valor limite de $\frac{1}{5} = 0,2$ e os irrelevantes tendendo a zero, como ocorre na Figura 5.8.

Após a convergência, o treinamento do algoritmo é finalizado e um vetor de relevâncias é obtido. Utilizando o vetor de relevâncias como fonte de informação sobre quais SNPs são os mais relevantes no conjunto de dados analisados, avaliamos também a capacidade de diferentes métricas inferir as interações entre os SNPs.

5.3.3 Métrica de Avaliação da Interação

Para identificar os SNPs relevantes e suas interações, fazemos o uso das informações contidas no vetor de relevâncias. No entanto, apenas essa informação não fornece a quantidade de SNPs que participam da interação, mas apenas, a relevância deles para a classificação dos dados.

O objetivo aqui, é utilizar uma métrica capaz de identificar dentre os SNPs que estão no

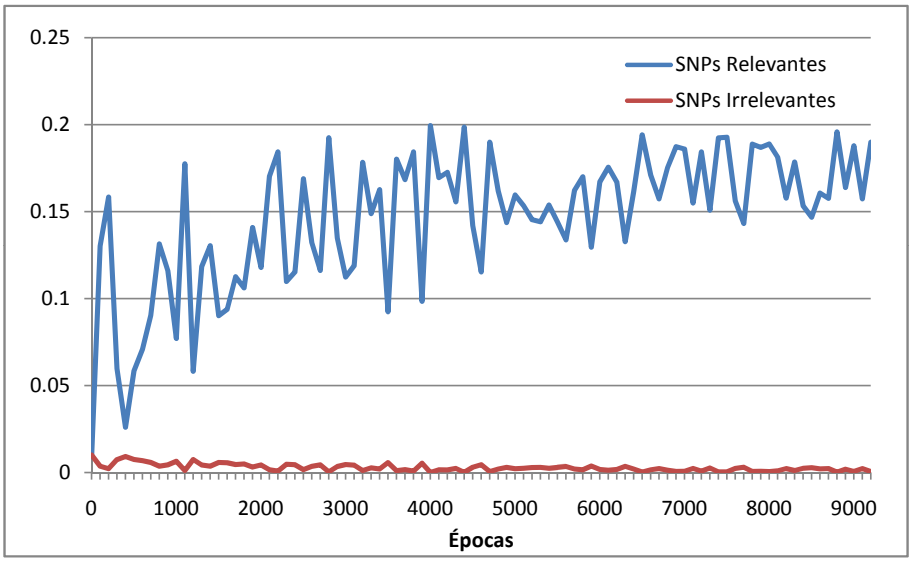


Figura 5.8 Comparação da variação média do vetor de pesos em função dos SNPs relevantes e irrelevantes.

topo do *ranking*, quais fazem parte de uma interação e quantos SNPs estão contidos nela. Além de verificar as interações, a métrica também será capaz de resgatar SNPs relevantes quando o algoritmo falha ao colocá-los no topo do *ranking*.

Após o treinamento do algoritmo, é obtido o vetor de relevâncias, que é ordenado decrescentemente, e os SNPs com as maiores relevâncias ficam no topo do *ranking* e são considerados os principais responsáveis por distinguir os pacientes casos de pacientes controles.

Para avaliar o k SNPs no topo do *ranking*, foram utilizadas diferentes métricas:

$$Acuracia = \frac{tP + tN}{tP + tN + fP + fN} \tag{5.2}$$

$$Precisao = \frac{tP}{tP + fP} \tag{5.3}$$

$$Exatidao = \frac{tP}{tP + fN} \tag{5.4}$$

$$MedidaF = \frac{2 \times tP}{2 \times tP + fP + fN} \tag{5.5}$$

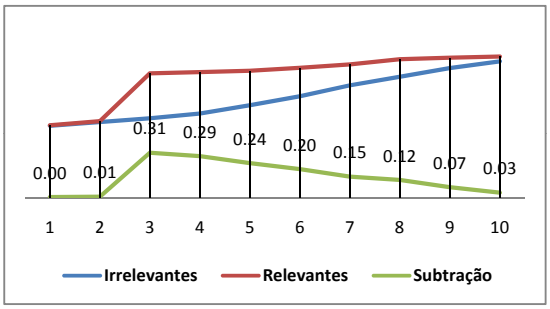
onde tP é o número de verdadeiros positivos, tN é o número verdadeiros negativos, fP é o número de falso positivos e fN é o número de falso negativos.

Para a avaliação das métricas, foram contruídos quatro conjuntos de dados contendo 10 SNPs, sendo então k = 10. Em três desses conjuntos, foram incluídos no primeiro 3 SNPs relevantes, no segundo 4 SNPs relevantes e no terceiro 5 SNPs relevantes e complementados com SNPs não relevantes até k = 10. No quarto conjunto, foram incluídos apenas SNPs não relevantes, selecionados aleatoriamente no conjunto de dados.

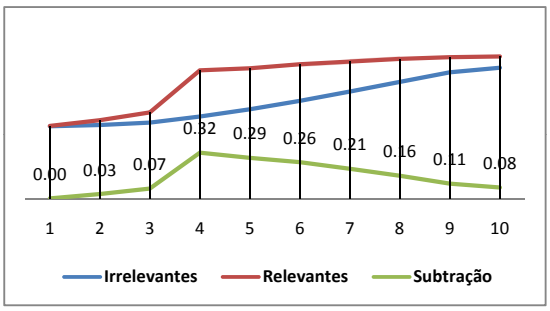
O objetivo desse experimento era verificar como as métricas se comportavam ao ana-

lisar esses diferentes conjuntos de dados. Em um primeiro passo, verificando a diferença de comportamento das métricas ao avaliar dados com e sem SNPs relevantes, e em um segundo passo, verificar se ocorriam distinções ao avaliar os dados com diferentes quantidade de SNPs relevantes.

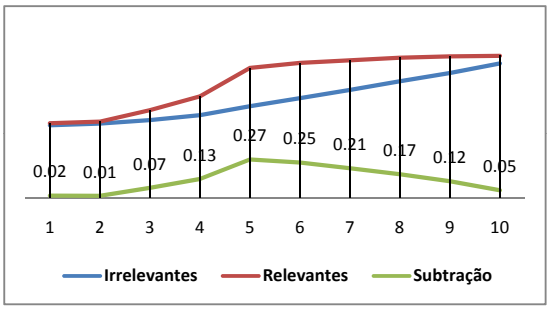
Neste experimento, todas as métricas foram testadas, mas a Acurácia foi a métrica que apresentou melhores resultados para a identificação das interações, como mostra na Figura 5.9. Neste gráfico, são apresentados três gráficos contendo três linhas, cada uma dessas linhas representam os valores da métrica da Acurácia para o grupo dos SNPs relevantes (vermelho), o grupo dos SNPs irrelevantes (azul) e a subtração do efeito do grupo dos irrelevantes pelos relevantes (verde).



(a) 3 Interações: Acurácia



(b) 4 Interações: Acurácia



(c) 5 Interações: Acurácia

Figura 5.9 Gráfico comparativo da métrica da Acurácia com SNPs Relevantes (vermelho), SNPs Irrelevantes (azul). A linha verde representa a substração das acurácias obtidas entre os SNPs relevantes e Irrelevantes.

Nessa subtração, observa-se claramente um pico que coincide com o número de interações de SNPs relevantes presente no conjunto analisado. Assim, para identificar a quantidade e os

SNPs incluídos na interação, será feito uso da métrica da Acurácia, no modelo proposto no Capítulo 6. Além disso, a Acurácia é uma métrica já bem estabelecida e utilizada em ferramentas de inferência de interações entre SNPs como MDR (Ritchie e Motsinger, 2005) e Polymorphism Interaction Analysis (PIA) (Goodman *et al.*, 2006).

5.4 Conclusão

O objetivo dos experimentos realizados com os filtros multivariados, era verificar se estes filtros eram capazes de reduzir a dimensionalidade dos dados, de forma a utilizá-los numa etapa de pré-processamento, para uma posterior aplicação em outras abordagens computacionais. No entanto, a partir dos resultados obtidos, foi verificado que os filtros multivariados aqui testados não foram robustos o suficiente para lidar com o problema da inferência das interações entre SNPs.

Nos experimentos com os algoritmos GRLVQ e SRNG foi observado que, mesmo utilizando conjuntos de parâmetros padrão, havia um potencial a ser explorado por parte desses algoritmos. Foram então propostas o uso da técnica de LHS e a métrica AR, para avaliar o comportamento dos algoritmos durante a busca paramétrica mais objetiva. Com esse estudo, pode-se observar a influência dos parâmetros ao lidar com o problema, e assim, propor hipóteses e realizar experimentos para lidar com o problema. Os experimentos para testar as hipóteses, avaliaram a inserção automática dos protótipos proposta no algoritmo do iGRLVQ, a convergência do algoritmo em função do vetor de relevância e a utilização de uma métrica para identificar as interações entre os SNPs.

A partir desses resultados, foi possível estudar o comportamento desses algoritmos frente ao problema da inferência da interações entre SNPs e sugerir adaptações no algoritmo do GRLVQ. Assim, no Capítulo 6 será apresentada uma nova metodologia baseada no GRLVQ, nomeada de *incremental GRLVQ for SNP interactions* (iGRLVQ-SNPi).

6

Modelo Computacional Proposto e Experimentos

Neste capítulo será apresentada uma nova metodologia inspirada em algoritmos baseados em RLVQ para realizar a inferência das interações entre os SNPs. O modelo computacional proposto nomeado como **incremental GRLVQ for SNP inference** (iGRLVQ-SNPi), tem como proposta realizar investigação de interações de dados genômicos com interações de ordem-n, ajustando a quantidade de protótipos através de métodos de inserção e remoção inspirados nos algoritmos do iGRLVQ (Kietzmann *et al.*, 2008) e SGNG (Garcia e Forster, 2012).

Em seguida são apresentados os experimentos utilizando diferentes conjuntos de dados, para comparar a acurácia e performance dessa nova metodologia com diferentes técnicas estudadas e descritas na literatura atual que realizam inferência das interações entre os SNPs.

6.1 Incremental GRLVQ for SNP Inference

O iGRLVQ-SNPi é um método inspirado nas metodologias de três algoritmos principais, o GRLVQ proposto por Hammer e Villmann (2002) que serve como base de todo o processo de treinamento e atualização dos protótipos e vetor de pesos, o iGRLVQ proposto por Kietzmann *et al.* (2008) como parte da metodologia utilizada para realizar o ajuste automático do número de protótipos e o SGNG proposto por Garcia e Forster (2012) sendo este fonte de inspiração para realizar a inserção de novos protótipos em um posicionamento mais estratégico no espaço de dados.

O processo de treinamento do iGRLVQ-SNPi é muito semelhante ao utilizado pelo GRLVQ. A função objetivo do iGRLVQ-SNPi é dada por:

$$C_{iGRLVQ-SNPi} = \sum_{i=1}^{m} sgd(\mu_\lambda(x_i)) \tag{6.1}$$

onde a função $sgd(x) = (1 + exp(-x))^{-1}$ é uma sigmóide logística e $\mu(x_i)$

$$\mu_\lambda(x_i) = \frac{d_\lambda^+(x_i) - d_\lambda^-(x_i)}{d_\lambda^+(x_i) + d_\lambda^-(x_i)} \tag{6.2}$$

onde, a $d_\lambda^-(x_i) = \lambda \times (x_i - w^+)^2$ e $d^-(x_i) = \lambda \times (x_i - w^-)^2$ são as distâncias ponderadas da amostra para os protótipos.

O algoritmo é inicializado com um protótipo para cada classe dos dados. A cada novo padrão x_i apresentado, são atualizados o protótipo mais próximo de mesma classe ω_{r^+}, de acordo com:

$$\Delta\omega_{r^+} = \varepsilon^+ \times sgd'_{\mu_\lambda(x_i)} \times \xi^- \times \frac{\partial d_\lambda^+(x_i)}{\partial \omega^{i+}}, \tag{6.3}$$

o protótipo mais próximo da classe oposta ω_{r^-}, de acordo com:

$$\Delta\omega_{r^-} = -\varepsilon^- \times sgd'_{\mu_\lambda(x_i)} \times \xi^+ \times \frac{\partial d_\lambda^-(x_i)}{\partial \omega^{i-}}, \tag{6.4}$$

e o vetor de relevâncias, de acordo com:

$$\Delta\lambda = -\varepsilon^\lambda \times sgd'_{\mu_\lambda(x_i)} \times \frac{\frac{2\times\partial d_\lambda^+(x_i)}{\partial\lambda\times d_\lambda^-(x_i)} - \frac{2\times\partial d_\lambda^+(x_i)\times\partial d_\lambda^-(x_i)}{\partial\lambda}}{(d_\lambda^+(x_i) + d_\lambda^-(x_i))^2} \tag{6.5}$$

onde ε^+, ε^- e ε^λ são respectivamente, as taxas de aprendizagem positiva (TAP), negativa (TAN) e do vetor de relevâncias (TAW). As distâncias derivadas com pesos são calculadas através das equações:

$$\xi^- = \frac{2 \times d_\lambda^-(x_i)}{(d_\lambda^+(x_i) + d_\lambda^-(x_i))^2} \quad e \quad \xi^+ = \frac{2 \times d_\lambda^+(x_i)}{(d_\lambda^+(x_i) + d_\lambda^-(x_i))^2} \tag{6.6}$$

Quando um novo padrão é apresentado o protótipo mais próximo de mesma classe é aproximado e o da classe oposta é afastado. Semelhantemente a como é feito em Kietzmann *et al.* (2008), sempre que uma amostra é apresentada, são contabilizados os erros de classificação considerando a distância da amostra para o protótipo de mesma classe. No entanto, em Kietzmann *et al.* (2008) o erro de classificação é contabilizado em uma única variável g para todas as classes de dados, e quando $g > g_{max}$ um novo protótipo é adicionado.

No iGRLVQ-SNPi, os erros de classificação são incrementados em cada um dos protótipos quando, a distância $d_\lambda^-(x_i)$ da amostra para o protótipo da classe oposta é menor que a distância $d_\lambda^+(x_i)$ para o protótipo de mesma classe. Quando isso ocorre, o protótipo da classe oposta, tem sua variável de erro $\omega_{r^-}Er$ incrementada.

Ao final de cada época de treinamento é verificado o erro de classificação $\omega_r Er$ de todos os protótipos. Quando esse erro ultrapassa o limiar de inserção $INSNode$, o protótipo com o maior erro é duplicado e representado com a classe oposta, da mesma forma como ocorre na inserção de novos protótipos em Garcia e Forster (2012). Ao final de cada interação, apenas um

protótipo é inserido por época. Quando isso ocorre, todos os protótipos tem seu parâmetro de erro setado para zero.

A remoção dos protótipos também ocorre a cada época do treinamento. Para que um protótipo seja mantido no mapa é necessário que ele represente um percentual mínimo de amostras *REMNode* no conjunto de dados. A representatividade de cada protótipo, é incrementada cada vez que ele é o vencedor para uma amostra da mesma classe. Ao final de cada época de treinamento, todos os protótipos tem sua representatividade verificada, e aqueles com a representatividade menor que o percentual mínimo de pacientes são removidos. A remoção do protótipo com baixa representatividade ainda leva em consideração se este não é o único protótipo de uma das classes. De forma que seja mantido ao menos um protótipo representante para cada uma das classes de dados.

Em comparação com o GRLVQ, o iGRLVQ-SNPi deixa de ter o parâmetro do número de nodos, pois a quantidade de nodos no mapa varia ao longo do treinamento. Em lugar disso, são considerados para fins de inserção e remoção dos protótipos dois limiares de inserção *INSNode* e remoção *REMNode* ajustados em função do número de amostras no conjunto de dados. No total a iGRLVQ-SNPi possui 7 diferentes parâmetros descritos na Tabela 6.1.

Tabela 6.1 Descrição dos Parâmetros do iGRLVQ-SNPi com os intervalos utilizados

Sigla	Parâmetro	Intervalo
ATP	Taxa de Aprendizagem Positiva	[0.8 a 0.9]
ATN	Taxa de Aprendizagem Negativa	[0.01 a 0.06]
ATW	Taxa de Aprendizagem do Vetor de Pesos	[0.2 a 0.3]
TAU	Taxa de Decaimento das Taxas de Aprendizagem	[5×10^{-6} a 1×10^{-5}]
EPOCHS	Número de Ciclos de Aprendizagem do Algoritmo	[10.000 a 30.000]
INSNode	Percentual de Inserção de Novos Protótipos	[0.01% a 0.1%]
REMNode	Percentual de Remoção dos Protótipos	[0.01% a 0.1%]

INSNode e REMNode são valores percentuais em função do número de amostras no conjunto de dados.

Um pseudo código do iGRLVQ-SNPi com um passo de treinamento é apresentado no Algoritmo 6, nele estão incluídos os processos de inserção e remoção dos protótipos.

Além da adição das funções de inserção e remoção de novos protótipos, um outro ponto considerado no iGRLVQ-SNPi diz respeito à forma como o algoritmo decide se a convergência foi alcançada. Originalmente o GRLVQ considera a convergência do algoritmo quando as atualizações ocorridas nos protótipos tendem a zero, de forma que já não ocorrem mudanças significativas das posições do protótipos no mapa, ou seja, à medida que as atualizações nos protótipos não alteram muito seu posicionamento, e a representatividade dos protótipos no mapa se torna estável, considera-se que o algoritmo convergiu.

No iGRLVQ-SNPi, sabendo-se da importância do vetor de pesos para a correta identificação dos SNPs relevantes, em vez de considerar a convergência do algoritmo em função da

Algoritmo 6: Treinamento do iGRLVQ-SNPi

Entrada: Conjunto de dados X

Saída: Vetor de relevância ordenado e as interações mais relevantes identificadas.

while $i \leq$ *Épocas* & *ChangeW* $\leq$ *Threshold* **do**
  for *all* (x_i, y_i) ε N **do**
    Encontra os protótipos mais próximo da mesma classe ω_{r^+} e da classe oposta ω_{r^-}
    Incrementa a representatividade do protótipo $\omega_{r^+}.Rep + 1$
    if $d_\lambda^+(x_i) > d_\lambda^-(x_i)$ **then**
      Incrementa o erro de classificação do protótipo $\omega_{r^-}.Er + 1$
    #Atualiza o protótipo da mesma classe da amostra de acordo com a equação (6.3)
    #Atualiza o protótipo da classe oposta da amostra de acordo com a equação (6.4)
    #Atualiza o vetor de relevâncias de acordo com a equação (6.5)
  #Remove Protótipos
  for *all Protótipos* ω_r **do**
    if $\omega_r.Rep < Data.rows() * 0.02$ **then**
      Remove ω_r
    else
      $\omega_r.Rep \leftarrow 0$
  #Insere Novo Protótipo
  for *all Protótipos* ω_r **do**
    #Encontra Protótipos com erro $\omega_r.Er$
    if $\omega_r.Er > Threshold$ **then**
      #Duplica e Insere Novo Protótipo na classe oposta
    $\omega_r.Er \leftarrow 0$
#Calcula a Acurácia das Interações

atualização dos protótipos, no iGRLVQ-SNPi a convergência do vetor de relevâncias é verificada como fator de decisão sobre a convergência do algoritmo.

Após os experimentos realizados no Capítulo 5, constatou-se a importância da convergência do vetor de relevâncias em detrimento da atualização dos protótipos. Para calcular tanto as mudanças nos protótipos ($change_{proto}$) quanto no vetor de relevâncias ($change_w$), a cada passo de treinamento, é calculado o quadrado dos valores de $\Delta\omega_{r^+}$, $\Delta\omega_{r^-}$ e de $\Delta\lambda$ para cada amostra acumulada e, ao final do treinamento, é obtido o $change_{proto}$ a partir da soma das raízes de $\Delta\omega_{r^+} + \Delta\omega_{r^-}$ e o $change_w$ a raiz dos valores acumulados de $\Delta\lambda$. Estes valores indicam de maneira global o quanto os protótipos e o vetor de relevâncias foram atualizados em cada época.

Ao final do processo de treinamento, o iGRLVQ-SNPi apresenta um vetor de relevâncias ordenado decrescentemente, estando os SNPs com as maiores relevâncias no topo do *ranking*, sendo os principais responsáveis por distinguir os pacientes casos dos controles.

Utilizando a métrica da Acurácia:

$$Acurcia = \frac{tP + tN}{tP + tN + fP + fN} \tag{6.7}$$

onde tP é o número de verdadeiros positivos, tN é o número verdadeiros negativos, fP é o número de falso positivos e fN é o número de falso negativos, são calculados os valores da acurácia dos k SNPs com as maiores relevâncias e dos k SNPs aleatoriamente selecionados.

Para os k SNPs selecionados no vetor de relevâncias, onde $k = 10$. A métrica da acurácia é utilizada para calcular a relevância de todas as n combinações possíveis, desde a relevância dos SNPs individualmente, onde é $n = 1$, até a interação com os 10 SNPs ($n = 10$). Para verificar qual a interação mais relevante, para cada tamanho de n, também é calcula a acurácia de SNPs aleatoriamente selecionados (assumidos como irrelevantes), e o valor obtido, é subtraído da acurácia dos SNPs selecionados a partir do vetor de relevâncias. As interações com os maiores valores, para cada n, teremos o conjunto de SNPs com as maiores interações identificadas.

Após a proposição do modelo, apresentado as modificações propostos para o algoritmo do GRLVQ, serão realizados experimentos com o iGRLVQ-SNPi, realizando comparações tanto de *performance* quanto da acurácia com outras ferramentas desenvolvidas para lidar com a inferência das interações entre os SNPs.

6.2 Experimentos com Outras Ferramentas

Esses experimentos têm como objetivo avaliar o desempenho do iGRLVQ-SNPi frente a diferentes ferramentas já bem estabelecidas em tratar a inferência entre SNPs. Para realizar esses experimentos, foram utilizados dois conjuntos de dados com características distintas. O primeiro foi o conjunto de dados (Shang *et al.*, 2016), com seis modelos epistáticos, sendo dois desses modelos também com efeitos marginais e interações entre dois SNPs. O segundo conjunto de dados (Himmelstein *et al.*, 2011) apresenta interações de alta ordem, com interações entre três, quatro e cinco SNPs. Ambos os conjuntos de dados são descritos no Capítulo 3 e os parâmetros utilizados no iGRLVQ-SNPi nos experimentos são apresentados na Tabela 6.1.

Os resultados obtidos com o conjunto de dados (Shang *et al.*, 2016) com duas interações utilizando as ferramentas SNPRuler, CINOEDV(P), CINOEDV(E), BOOST, iGRLVQ-SNPi são exibidos no gráfico da Figura 6.1. Sendo o CINOEDV(E), a ferramenta avaliada ao utilizar a busca do tipo exaustiva e o CINOEDV(P) utilizando a busca *Particle Swarm Optimization* (PSO), já descrita na Seção 4.2.3. Constata-se através do gráfico da Figura 6.1 que o iGRLVQ-SNPi tem um desempenho tão bom quanto as demais ferramentas avaliadas, ficando atrás apenas do CINOEDV(E), pois esta é uma ferramenta que realiza uma busca exaustiva para identificar os pares de interações entre os SNPs relevantes. Vale enfatizar que os resultados apresentados com o iGRLVQ-SNPi foram obtidos com um único conjunto de parâmetros para todos os modelos de dados.

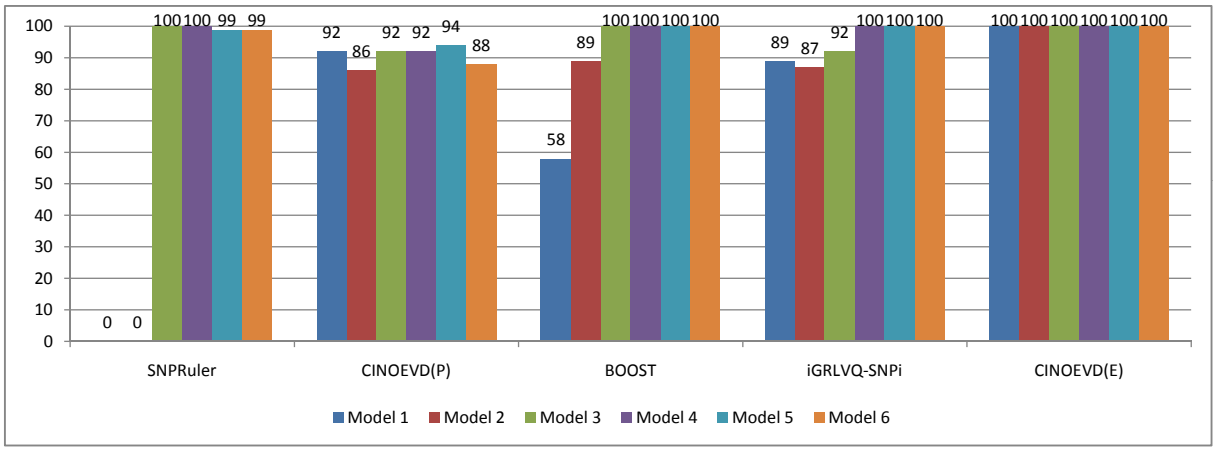


Figura 6.1 Gráfico exibindo a quantidade de SNPs relevantes identificados como mais relevantes nos 100 arquivos testados em dados, com duas interações (Shang *et al.*, 2016).

Considerando as características de cada um dos modelos epistáticos testados, observamos que principalmente os modelos 1 e 2, que são caracterizados por possuírem tanto efeito epistático quanto efeito marginal, são os mais difíceis de serem identificados pelo iGRLVQ-SNPi e demais ferramentas. Os demais modelos 3 a 6 são caracterizados apenas por apresentar efeitos epistáticos. A média dos resultados obtidos com o iGRLVQ-SNPi é exibida na Tabela 6.2, onde é apresentada a média dos acertos com o SNP no topo do *ranking*, o *Average Ranking* e os respectivos desvios dos diferentes modelos testados. Em todos os modelos, mesmo nos mais difíceis, a média aritmética ficou em 69,4 e utilizando o *Average Ranking* ficou acima de 95%, indicando que embora nos casos em que o iGRLVQ-SNPi não acertou todos SNPs relevantes no topo do *ranking*, estes se encontram pelo menos entre os 10 primeiros no *ranking*.

Tabela 6.2 Acurácia do iGRLVQ-SNPi em dados com 800 indivíduos e interações de alta ordem.

Métrica	**Acurácia média**					
	M1	**M2**	**M3**	**M4**	**M5**	**M6**
MA	72,6 (7,45)	69,4 (5,55)	77,5 (6,73)	97,0 (1,94)	95,8 (2,26)	99,9 (0,14)
AR	97,4 (0,46)	96,9 (0,58)	97,6 (0,46)	99,4 (0,25)	99,3 (0,29)	99,9 (0,02)

Em parêntesis são apresentados os desvios padrão. MA: Média Aritmética; AR: *Average Ranking*.

Considerando o problema da inferência das interações de alta ordem, foram testadas as ferramentas MDR, BEAM, CINOEDV(P), CINOEDV(E) e iGRLVQ-SNPi. Para o MDR foram mantidos os parâmetros padrão, modificando apenas o número de interações a serem avaliadas pela ferramenta. O BEAM teve o parâmetro $INITIALTRYS$ ajustado de acordo com o sugerido pelo autor (Zhang e Liu, 2007), $10 * L$, sendo L o número de interações que se deseja investigar. No SNPRuler os parâmetros utilizando foram $listSize$ igual a 1, este parâmetro se refere a uma lista com a quantidade de interações relevantes identificadas, $depth$ é a ordem das interações, sendo ajustado para 3, 4 e 5, de acordo com o conjunto de dados utilizado, e $updateRatio$ é

o passo de treinamento, sendo mantido 0.2 como sugerido pelo autor. Os parâmetros usados para o CINOEDV(P) foram os mesmos utilizados nos experimentos realizados em Shang *et al.* (2016), o número de partículas igual a 500, 10 iterações e o número de interações (MaxOrder) foi ajustado para cada conjunto de dados com 3, 4 e 5 interações. O iGRLVQ-SNPi utilizou os mesmos parâmetros utilizados no experimento anterior. Na Tabela 6.3 são apresentados os melhores resultados obtidos.

Tabela 6.3 Acurácia dos métodos com conjuntos de dados com 800 indivíduos e interações de alta ordem.

Int.	**Acurácia(%)**														
	20 SNPs					**100 SNPs**					**1000 SNPs**				
	M	**B**	**S**	**Cp**	**iG**	**M**	**B**	**S**	**Cp**	**iG**	**M**	**B**	**S**	**Cp**	**iG**
3 way	100	24	100	100	100	100	0	95	100	100	NA	0	0	100	100
4 way	100	0	50	90	100	100	0	7	66	100	NA	0	0	0	100
5 way	100	0	3	71	98	NA	0	0	34	100	NA	0	0	0	81

Abreviações M (MDR), B (BEAM), S (SNPRuler), Cp (CINOEDV(P)) e iG(iGRLVQ-SNPi). Nos resultados, NA refere-se a resultado não disponível, pois o MDR precisa de um longo tempo ou grande espaço em memória para processar os dados.

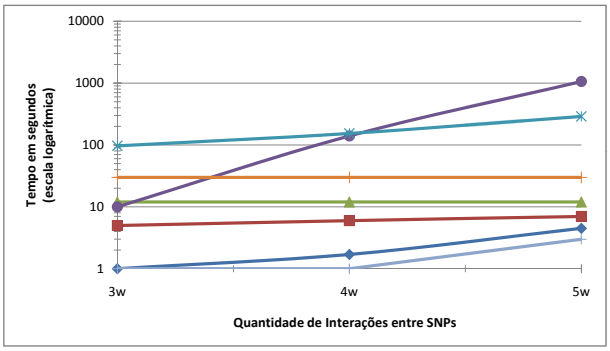
Os resultados obtidos com os dados de alta ordem mostram um excelente desempenho do MDR, iGRLVQ-SNPi e CINOEDV(P). No entanto, o MDR é o único método exaustivo incluído na comparação. O CINOEDV(P) obteve excelentes resultados avaliando dados com até 3 interações, porém, em dados com maior número de interações sua acurácia foi afetada.

6.2.1 Tempo de Processamento

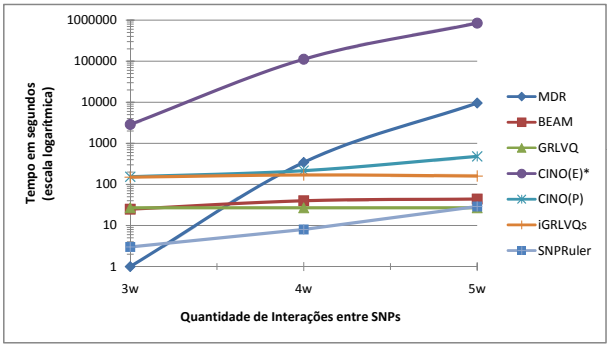
Para avaliar o tempo de processamento das diferentes ferramentas estudadas foram utilizadas diferentes configurações de conjunto de dados. Todos os experimentos foram executados em um mesmo ambiente computacional: Linux Ubuntu 14.04, 64 bits com processador Intel Xenon, com 8 núcleos de 2.0 GHz e 16GB de memória RAM. Apesar do sistema multiprocessado, nenhum método de paralelização foi implementado nas ferramentas avaliadas. Os resultados obtidos são apresentados na Figura 6.2.

O uso ideal de cada uma das ferramentas aqui estudadas está correlacionado com uma dependência entre a acurácia e o tempo de processamento. Dependendo do tamanho da amostra e número de interações a serem avaliadas, se for computacionalmente plausível, o uso de ferramentas exaustivas é o mais indicado. No entanto, os conjuntos de dados tendem a ficar cada vez maiores, assim como o número de interações a serem observadas, tornando o seu uso cada vez menos viável.

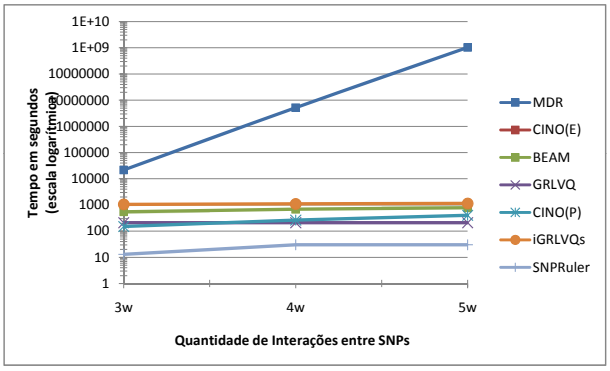
O MDR como mostra na Figura 6.2(a) mostrou-se a ferramenta mais eficiente tanto em acurácia quanto tempo de processamento quando realizados experimentos com dados com 20 SNPs. No entanto, devido ao seu caráter exaustivo a medida que o número de interações



(a) 20 SNPs



(b) 100 SNPs



(c) 1000 SNPs

Figura 6.2 Tempo de processamento em escala logarítmica das ferramentas MDR, BEAM, SNPRuler, CINOEDV(P), CINOEDV(E) e iGRLVQ-SNPi (iGRLVQ*s* abreviação para iGRLVQ-SNPi). ao analisar conjuntos de dados com 3, 4 e 5 interações com 20, 100 e 1000 SNPs. Nos gráficos, o tempo de processamento do CINOEDV(E) com 5 interações e 100 SNPs e do MDR com 1000 SNPs foram estimados.

e SNPs cresce, as demandas de memória e tempo crescem exponencialmente. Nos conjuntos com 1000 SNPs (Figura 6.2(c)) o MDR exibiu um tempo estimado de 6 horas para concluir a análise com 3 interações e seriam precisos mais de 60 dias para concluir com 5 interações. Esse mesmo comportamento foi observado ao realizar os testes com o CINOEDV(E), embora seja uma ferramenta um pouco mais lenta que o MDR, o CINOEDV(E) é tão eficiente quanto o MDR na inferência das interações quando se trata de dados com baixa ordem (≤ 2).

Dentre as ferramentas testadas, GRLVQ e o iGRLVQ-SNPi são os métodos que se

mantêm mais constantes no tempo de processamento com a variação do número de interações. O CINOEDV(P), por exemplo, embora seja mais rápido que o iGRLVQ-SNPi ao avaliar interações de baixa ordem, à medida que o número de interações cresce, o tempo de processamento também aumenta. Isso reflete o fato de que nesse método, assim como no SNPRuler e no BEAM, parâmetros devem ser ajustados para que a ferramenta realize inferências com diferentes números de SNPs nas interações.

6.3 Experimentos com Dados Reais

Para o estudo dos pacientes acometidos com o mal de Alzheimer, foi realizada uma seleção dos pacientes, e remoção de dados faltantes. Em seguida é apresentado o resultado obtido com a inferência dos SNPs relevantes ao utilizar o iGRLVQ-SNPi, validando os resultados obtidos com as informações já coletadas e descritas na Seção 3.3 e também em publicações científicas relacionadas com a AD.

6.3.1 Pré-processamento dos Dados

Os procedimentos a serem realizados incluem a análise do conjunto de dados, verificação do número de SNPs, de casos e controles, sexo dos pacientes. Também são verificadas as quantidades de dados faltantes, realizando a remoção da informação quando necessário, utilizando para isso, a ferramenta PLINK, proposto por Purcell *et al.* (2007) para remover as informações incorretas.

Nas análises de dados de GWAS surge um grande número de variantes genéticas relacionadas com doenças complexas. Muitas destas variações são detectadas e possuem um efeito pequeno sobre a doença e mesmo quando consideradas em conjunto atuam em uma pequena porção da variância genética total. Devido a este pequeno poder estatístico sobre a variância, para que estas variações genéticas sejam detectadas e validadas se faz necessário o uso de uma maior quantidade de pacientes.

No entanto, esta grande gama de dados traz consigo inconvenientes, pois simples erros aleatórios nos dados podem influenciar na detecção destas variantes, escondendo ou reduzindo os efeitos reais. Diante disso, se faz necessário ter uma atenção especial na qualidade dos dados que serão analisados, surgindo assim a necessidade de realizar um pré-processamento. Nesta Tese, o pré-processamento dos dados foi dividido nas etapas ilustradas na Figura 6.3 e descritas em detalhes nas seções seguintes.

6.3.2 Seleção de Pacientes

Como já dito na Seção 3.3, o conjunto de dados reais foi obtido da *Alzheimer's Disease Neuroimaging Initiative (ADNI) database*, sendo composto por 757 pacientes (449 homens e 308 mulheres) com um total de 620.901 SNPs. Dos 757 pacientes, 214 são do tipo controle, ou

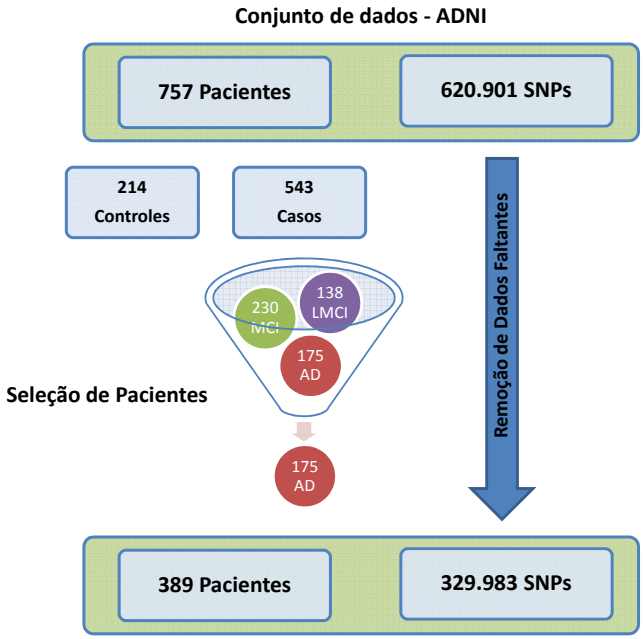


Figura 6.3 Fluxograma exibindo o passo a passo do procedimento realizado para pré-processamento dos dados.

seja pacientes sadios e 543 do tipo caso, sendo 175 pacientes com a Doença de Alzheimer e 368 pacientes com um comprometimento cognitivo leve.

Considerando nosso interesse em apenas analisar os dados dos pacientes com AD, foi realizada uma seleção dos pacientes, como pode ser observada na segunda etapa da Figura 6.3. Assim nosso conjunto de dados final é composto por 389 pacientes, sendo 214 pacientes controle (99 mulheres e 115 homens) e 175 pacientes caso com AD (82 mulheres e 93 homens) e 620.901 SNPs.

6.3.3 Remoção de Dados Faltantes

Após a seleção dos pacientes, foi observado com a utilização da ferramenta **PLINK** (Purcell *et al.*, 2007), que alguns SNPs só continham informações para os pacientes dos grupos LMCI e MCI. Assim ao remover esses pacientes, constatamos que 28.369 SNPs ficaram com 100% de dados faltantes, sendo removidos pois não traziam nenhuma informação para nosso estudo ficando um total de 592.532 SNPs. Desses 592.532 SNPs, 329.983 SNPs não tinham nenhuma informação faltante e esse conjunto foi utilizado nos experimentos para inferir os SNPs relevantes.

6.3.4 Inferência com Dados Reais

Após o pré-processamento dos dados, utilizando o iGRLVQ-SNPi, foi realizado um experimento preliminar utilizando um subconjunto do conjunto de dados real. Nesse subconjunto, estavam contidos 76 SNPs identificados nos dados reais e referenciados na literatura como

relevantes fatores de risco a Doença de Alzheimer. Utilizando esses dados, os SNPs foram avaliados utilizando o mesmo intervalo de parâmetros já ajustado para o iGRLVQ-SNPi, e utilizado nos experimentos das seções anteriores (Tabela 6.1). Dos resultados obtidos, sete conjuntos de interações entre SNPs obtiveram os maiores valores de acurácia 0,83. Nesses sete conjuntos, estavam incluídos com maior frequência (em ≤ 3 ocorrências de 7) os SNPs e seus respectivos genes: rs405509 (APOE), rs4935775 (SORL1), rs8106922 (TOMM40), rs1468503 (NPC2), rs11264359 (FDPS), rs1871045 (BCAM) e rs8070723 (MAPT).

Tabela 6.4 Relação dos 25 SNPs mais relevantes identificados pelo iGRLVQ-SNPi utilizando os dados do ADNI.

	Índice	**Cr**	**SNPrs**	**Gene**	**Relevância (%)**
1	43327	2	rs267992	CYTIP	0,30%
2	290654	17	rs8078624	-	0,29%
3	328557	X	rs2235306	APLN	0,24%
4	74429	3	rs710493	-	0,23%
5	3304	1	rs415314	-	0,23%
6	3834	1	rs6697778	-	0,22%
7	329017	X	rs384230	-	0,21%
8	294632	18	rs9966184	-	0,21%
9	123795	6	rs720988	LOC101929770	0,20%
10	50436	2	rs888524	-	0,19%
11	294814	18	rs1681045	MIR924HG	0,19%
12	294823	18	rs650998	MIR924HG	0,19%
13	329571	X	rs6540428	-	0,18%
14	83754	4	rs9884248	NAA11	0,17%
15	73385	3	rs1805610	SOX2-OT	0,17%
16	41185	2	rs17740707	THSD7B	0,16%
17	74433	3	rs710512	TPRG1	0,14%
18	329783	X	rs5970148	-	0,14%
19	61687	3	rs28705165	-	0,14%
20	183037	9	rs10121793	-	0,14%
21	306228	20	rs6113720	TGM3	0,14%
22	151192	7	rs1721447	-	0,14%
23	200132	10	rs3763679	PPP3CB	0,14%
24	73401	3	rs2176612	SOX2-OT	0,13%
25	287435	17	rs1113953	-	0,13%

Abreviações Cr para cromossomo, SNPrs é o número de referência utilizado nos bancos de dados.

Considerando esta informação, o conjunto de dados com 389 pacientes e 329.983 SNPs foi avaliado e os 25 SNPs mais relevantes identificados são apresentados na Tabela 6.4. Dentre os 25 SNPs mais relevantes, nenhum deles está contido no subconjunto com os 76 SNPs relevantes indicados na literatura. No entanto, avaliando os 25 SNPs identificados, foi observado que o SNP mais relevante identificado rs267992 localizado no gene CYTIP do cromossomo 2, foi apontado por Hohman *et al.* (2016) como um candidato em potencial a fator de risco da AD,

a ser explorado. Outros dois SNPs dessa lista, o rs1805610 e o rs2176612, localizados no gene SOX2-OT do cromossomo 3, seu gene, também denominado SOX2, está correlacionado a manutenção das células neuronais, sendo apresentado em Sarlak *et al.* (2016) que a baixa expressão desse gene leva a um processo neurodegenerativo em ratos e seu decréscimo tem sido observado em pacientes com AD.

Utilizando a base de dados KEGG (Kanehisa *et al.*, 2014), o SNP rs3763679, localizado no gene PPP3CB do cromossomo 10, foi identificado na via metabólica da doença de Alzheimer e Esclerose Lateral Amiotrófica (outra doença neurodegenerativa). Por fim, quatro SNPs dessa lista, SNPs rs2235306, rs384230, rs6540428 e rs5970148, estão localizados no cromossomo X, esta maior incidência de SNPs relevantes identificados neste cromossomo pode levantar a hipótese de estarem correlacionados aos estudos epidemiológicos apresentados por Grimm *et al.* (2016), que relata que as mulheres são mais acometidas pela doença de Alzheimer do que os homens, representando $\frac{2}{3}$ dos casos de AD. Esse maior número de casos no sexo feminino, pode estar relacionado a fatores de risco presentes no cromossomo sexual X. Para os demais SNPs não relatados, não foram encontradas fontes científicas ou informações sobre seus genes ou sua associação com a doença de Alzheimer.

Um terceiro experimento, foi realizado criando um conjunto de dados com 1000 SNPs a partir do conjunto de dados reais. Para esse experimento, foram inseridos os 76 SNPs relevantes relatados na literatura e o conjunto foi complementado com os SNPs mais relevantes identificados no experimento anterior (sem repetição dos 76 relevantes). No resultado obtido com o iGRLVQ-SNPi, foi selecionado o resultado com a interação com maior acurácia obtida (0,85). Essa interação é composta por 5 SNPs, sendo eles e seus respectivos genes: rs405509 (APOE), rs4935775 (SORL1), rs1468503 (NPC2), rs11264359 (FDPS) e rs1050565 (BLMH). Desses cinco SNPs, quatro deles já haviam sido identificados anteriormente, como alguns dos mais relevantes no experimento com o conjunto com 76 SNPs.

6.4 Conclusão

Neste capítulo, foram apresentadas as modificações propostas para o algoritmo do GRLVQ se tornar adaptável ao problema da inferência das interações entre SNPs. Para isso, foram utilizadas diferentes algoritmos como fonte de inspiração para o seu desenvolvimento. Sendo também apresentados os experimentos realizados com o iGRLVQ-SNPi, realizando comparações de performance e acurácia com outras ferramentas desenvolvidas para lidar com a inferência das interações entre SNPs.

7

Considerações Finais

No Capítulo 1 desta tese, foram apresentados os desafios envolvidos em realizar a inferências da interações entre SNPs, a motivação da pesquisa, que envolve a identificação de interações entre SNPs relevantes associados a diferentes doenças e os objetivos propostos. Um primeiro objetivo alcançado foi o estudo de diferentes técnicas de seleção de características. Um segundo objetivo alcançado foi o estudo de algoritmos baseados em LVQ, observando as vantagens e desvantagens desses algoritmos. Com a análise desses resultados, foi possível propor um novo modelo, sendo este o terceiro e mais importante objetivo alcançado nesta tese.

No Capítulo 5 foram apresentados os experimentos e resultados obtidos com a avaliação das diferentes abordagens computacionais descritas nesta tese. Inicialmente, foram realizados experimentos com os filtros multivariados, foi verificado que embora essas técnicas, tenham sido propostas para detectar correlação entre variáveis, elas não foram capazes de identificar as interações em conjuntos de dados com 1000 SNPs. O desempenho dessas técnicas para o problema em mãos se degrada significativamente com o aumento no número de interações e da dimensionalidade dos dados, tornando-as praticamente inviáveis para aplicação em filtragem de SNPs relevantes em conjuntos de dados de GWAS.

Os algoritmos de aprendizagem de máquina baseados em RLVQ, como o GRLVQ e o SRNG, apresentaram grande potencial na detecção das interações em conjuntos de dados de polimorfismo. Estudos iniciais mostraram que, mesmo utilizando parâmetros padrão, o uso do vetor de relevância na inferência dos SNPs relevantes possibilitava a identificação das interações entre SNPs relevantes. Resultados ainda melhores foram mostrados após o ajustes finos dos parâmetros utilizando a técnica LHS, conjuntamente com o uso da métrica *Average Ranking* (AR), proposta nesta tese, para avaliar o comportamento dos resultados ao utilizar diferentes conjuntos de parâmetros.

Em experimentos com o GRLVQ, utilizando conjuntos de dados com diferentes quantidade de interações entre SNPs e avaliando os resultados com o uso da métrica AR, foi possível observar que a quantidade de SNPs envolvidos nas interações influenciava na acurácia do algoritmo, pois um maior número de interações exigia um maior número de protótipos para representa-las. Essa observação apontou a necessidade de utilizar uma quantidade maior de

protótipos, para que o GRLVQ fosse capaz de identificar corretamente as interações entre os SNPs.

Uma alternativa estudada foi realizar experimentos com o iGRLVQ, o que revelou não só a importância da adaptação automática do número dos protótipos com a complexidade dos dados, como também revelou que, a posição na qual os protótipos são inseridos no mapa é relevante para a convergência e identificação dos SNPs relevantes.

Outros experimentos realizados envolveram a análise e proposição do uso do decaimento da atualização média do vetor de pesos para verificar a convergência do algoritmo. Nesse experimento, foi observado que o vetor de pesos precisava de um pouco mais de épocas para convergir corretamente em comparação com as atualizações dos protótipos, para que assim pudesse assegurar a identificação dos SNPs relevantes.

Além disso, foi avaliado o uso de quatro métricas estatísticas, Acurácia, Precisão, Exatidão e Medida F, para serem utilizadas na identificação das interações entre os SNPs mais relevantes. A métrica da Acurácia foi escolhida pois, além de ser utilizada por outras ferramentas de inferência de interações epistáticas, apresentou resultados mais estáveis na verificação das interações dos SNPs relevantes em comparação com as demais métricas.

Inspirado nos resultados obtidos com esses experimentos e tendo como base o algoritmo do GRLVQ, no Capítulo 6 é proposto o **incremental GRLVQ for SNP inference** (iGRLVQ-SNPi). Este método possui o número de protótipos variável, aumentando de forma dinâmica conforme surge a necessidade, essa necessidade tem sido associada a complexidade dos dados em função da quantidade de SNPs interagindo. O método utiliza o decaimento da atualização do vetor de pesos como parâmetro para verificar sua convergência e utiliza a métrica da Acurácia como forma de verificar e resgatar as interações entre os SNPs mais relevantes.

No Capítulo 6, são apresentados experimentos avaliando as capacidades do iGRLVQ-SNPi, ao realizar análises comparativas com outras ferramentas. O iGRLVQ-SNPi mostrou ter um desempenho tão bom quanto ferramentas de estado arte, com a vantagem de que é um método que não tem sua performance afetada pela quantidade de interações presentes no conjunto de dados. Diferentemente do BOOST, que só realiza inferência com duas interações, ou mesmo o CINOEDV, que permite a avaliação de até 5 interações, o iGRLVQ-SNPi utilizando a inserção e remoção de protótipos, representa os pacientes e identifica os SNPs relevantes, considerando a sua relevância de forma conjunta através do uso do vetor de relevâncias.

Vale considerar que mesmo em situações nas quais o iGRLVQ-SNPi não consegue identificar corretamente os SNPs relevantes no topo do *ranking*, como em situações em que os parâmetros não possam ser adequadamente ajustados, ainda assim, os SNPs relevantes ficam em altas posições no *ranking*. Sendo então recuperados com o uso da métrica da acurácia para identificar a interação mais relevante. Este fato também revela uma possibilidade de se utilizar este método como uma técnica de filtragem para a redução no número de dimensões.

Nos experimentos com dados reais, foram inicialmente avaliadas as relevâncias dos 76 SNPs descritos na literatura como fatores de risco da AD, anotando os SNPs mais relevantes e as

suas interações. Ao analisar o conjunto de dados com mais de 300 mil SNPs esses SNPs não foram identificados, no entanto, outros SNPs associados com a AD foram identificados dentre os 25 mais relevantes. Um terceiro experimento envolvendo esses dados mostrou que ao utilizar uma dimensionalidade menor (1000 SNPs) no conjuntos de dados, o iGRLVQ-SNPi foi capaz de identificar uma interação mais alta acurácia (0,85) envolvendo cinco SNPs, a presença de quatro SNPs apontados como fatores de risco da AD na literatura.

7.1 Contribuições

Uma das contribuições mais interessantes deste trabalho foi demonstrar através de extensos experimentos, o uso eficiente de um algoritmo utilizado para classificação de dados para resolver um problema ainda não tratado com esses algoritmos.

Para avaliar os algoritmos, foi proposta uma metodologia baseada em LHS para avaliar de forma mais eficiente intervalos de parâmetros e a partir dos resultados obtidos, foi proposta uma nova métrica nomeada de *Average Ranking* para visualizar de forma mais eficiente o posicionamento dos SNPs relevantes no ranking.

Com estes experimentos, foi possível verificar o comportamento dos algoritmos baseados em LVQ e propor modificações em sua metodologia de forma a torna-los mais adaptáveis ao problema. Sendo a contribuição mais relevante, a proposição de um novo modelo para tratar o problema da inferência das interações entre SNPs, comparando a técnica proposta com métodos do estado da arte e demonstrando um excelente desempenho do novo modelo na recuperação dos SNPs envolvidos.

7.2 Trabalhos Futuros

Apesar dos esforços realizados na proposição do modelo, diversos pontos ainda não foram tratados adequadamente e precisam ser abordados futuramente para que seja possível lidar com o problema da inferência das interações entre SNPs de forma mais robusta.

Como trabalho futuro, pode ser verificada a limitação do modelo em função da dimensionalidade dos dados, embora tenha sido empregados esforços para lidar com este problema, ainda se faz necessário uma etapa anterior a inferência das interações que possam reduzir a dimensionalidade dos dados e assim, tornar mais fácil o processo de aprendizado para o modelo.

Uma melhoria ao modelo, seria verificar sua capacidade ao lidar com alguns problemas que podem ser encontrados nos dados reais, tais como, dados faltantes, dados desbalanceados, erros de genotipagem e fenocópias (quando um fenótipo é expresso é diferente do genótipo).

Avaliação dos grupos formados pelos protótipos casos e controles, as informações contidas nesses grupos podem revelar subgrupos de SNPs associados com a doença.

Realizar mais experimentos com o modelo utilizando outras bases de dados reais, de forma a avaliar as reais capacidades e limitações do modelo.

Aplicar o modelo para outros tipos de problemas, como por exemplo, dados de microarray.

7.3 Publicações

Esta Tese foi desenvolvida na Universidade Federal de Pernambuco, no Centro de Informática. Durante o período de Março de 2012 a Fevereiro de 2017, com um período de licença maternidade. Este trabalho foi financiado pela FACEPE sobre os processos: IBPG-0761-1.03/11 e IBPG-0761-1.03/15.

Publicações relacionadas e produzidas durante esta Tese:

F. R. B. Araujo; K. S. Guimarães. Inference of High-Order Epistatic Interactions Using Generalized Relevance Learning Vector Quantization with Parametric Adjustment. The 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI). DOI: 10.1109/ICTAI.2016.0104

F. R. B. Araujo; K. S. Guimarães. A Novel Incremental GRLVQ Applied for Inference of High-Order SNPs Interactions. A ser revisado e submetido.

Referências

Abdi, H., 2007. Bonferroni and šidák corrections for multiple comparisons. Encyclopedia of measurement and statistics 3, 103–107.

Alzheimer's-Association, Março 2016. Inside the brain: Alzheimer's brain tour. Acessado em Março de 2016, disponivel em: http://www.alz.org/.

Araujo, F. R. B., Bassani, H. F., Araujo, A. F. R., 2013. Learning vector quantization with local adaptive weighting for relevance determination in genome-wide association studies. Em: The 2013 International Joint Conference on Neural Networks, IJCNN. IEEE Computer Society, pp. 1–8.

Araujo, F. R. B., Guimaraes, K. S., 2011. Computational Tools For SNP Interactions - How Good Are They? Em: 11th IEEE International Conference in Bioinformatics and Bioengineering (BIBE). pp. 295–298.

Araujo, F. R. B., Guimaraes, K. S., Nov 2016. Inference of high-order epistatic interactions using generalized relevance learning vector quantization with parametric adjustment. Em: The 2016 International Conference Tools with Artificial Intelligence ICTAI. Institute of Electrical and Electronics Engineers (IEEE).

Araujo, F. R. B., Gusmao, E. G., Guimaraes, K. S., Nov 2011. A case-control study of non-parametric approaches for detecting snp-snp interactions. 2011 30th International Conference of the Chilean Computer Science Society.

Ardlie, K. G., Kruglyak, L., Seielstad, M., Apr 2002. Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3 (4), 299–309.

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., Tanzi, R. E., Jan 2007. Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database. Nature Genetics 39, 17–23, 10.1038/ng1934.

Bojer, T., Hammer, B., Schunk, D., von Toschanowitz, K., 2001. Relevance determination in learning vector quantization. Em: side publications, D. (Ed.), European Symposium on Artificial Neural Networks. p. 271–276.

Botstein, D., Risch, N., Mar 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genetics 33 Suppl, 228–237.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H. M., Jul 2007. Forecasting the global burden of Alzheimer's disease. Alzheimers & Dementia 3 (3), 186–191.

Bush, W. S., Dudek, S. M., Ritchie, M. D., Sep 2006. Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. Bioinformatics 22 (17), 2173–2174.

Cancer-Society, A., 2010. Breast cancer facts & figures. Acessado em Março de 2016, disponivel em: http://www.cancer.org.

Chen, L., Yu, G., Miller, D. J., Song, L., Langefeld, C., Herrington, D., Liu, Y., Wang, Y., 2009. A ground truth based comparative study on detecting epistatic SNPs. Em: Proceedings IEEE Int. Conf. Bioinformatics and Biomedicine Workshop BIBMW. pp. 26–31.

Clancy, S., 2008. Rna functions. Nature Education 1 (1), 102.

Consortium, I. H. G. S., Feb 2001. Initial sequencing and analysis of the human genome. Nature 409 (6822), 860–921.

Consortium, I. H. G. S., Oct 2004. Finishing the euchromatic sequence of the human genome. Nature 431 (7011), 931–945.

Cordell, H. J., Oct 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human Molecular Genetics 11 (20), 2463–2468.

Devall, M., Roubroeks, J., Mill, J., Weedon, M., Lunnon, K., Feb 2016. Epigenetic regulation of mitochondrial function in neurodegenerative disease: New insights from advances in genomic technologies. Neuroscience Letters.

Fang, G., Haznadar, M., Wang, W., Yu, H., Steinbach, M., Church, T. R., Oetting, W. S., Van Ness, B., Kumar, V., 2012. High-order SNP combinations associated with complex diseases: efficient discovery, statistical power and functional interactions. PLoS ONE 7 (4), e33531.

Feuk, L., Carson, A. R., Scherer, S. W., Feb 2006. Structural variation in the human genome. Nature Reviews Genetics 7 (2), 85–97.

Fritzke, B., 1995. A growing neural gas network learns topologies. Em: Advances in Neural Information Processing Systems 7. MIT Press, pp. 625–632.

Garcia, K., Forster, C. H. Q., 2012. Supervised growing neural gas. Lecture Notes in Computer Science, 502–507.

Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S., Fiske, A., Pedersen, N. L., Feb 2006. Role of genes and environments for explaining Alzheimer disease. Archives of General Psychiatry Journal 63 (2), 168–174.

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., Barabasi, A. L., May 2007. The human disease network. PNAS, Proceedings of the National Academy of Sciences 104 (21), 8685–8690.

Goodman, J. E., Mechanic, L. E., Luke, B. T., Ambs, S., Chanock, S., Harris, C. C., April 2006. Exploring SNP-SNP interactions and colon cancer risk using polymorphism interaction analysis. International Journal of Cancer 118(7) (1), 1790–1797.

Gorre, M., Mohammed, M., Ellwood, K., Hsu, N., Paquette, R., Rao, P., Sawyers, C., 2001. Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. Science 293(5531), 876–880.

Grimm, A., Mensah-Nyagan, A. G., Eckert, A., Apr 2016. Alzheimer, mitochondria and gender. Neuroscience & Biobehavioral Reviews.

Hahn, L. W., Ritchie, M. D., Moore, J. H., Feb 2003. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19 (3), 376–382.

Hall, M., 1999. Correlation-based feature subset selection for machine learning. Tese de Doutorado, University of Waikato.

Hammer, B., Strickert, M., Villmann, T., 2005. Supervised neural gas with general similarity measure. Neural Process Lett 21 (1), 21–44.

Hammer, B., Villmann, T., 2002. Generalized relevance learning vector quantization. Neural Networks 15 (8-9), 1059–1068.

He, H., Oetting, W. S., Brott, M. J., Basu, S., 2009. Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study. BMC Medical Genetics 10, 127.

Himmelstein, D. S., Greene, C. S., Moore, J. H., 2011. Evolving hard problems: Generating human genetics datasets with a complex etiology. BioData Mining 4 (1), 21.

Hohman, T. J., Cooke-Bailey, J. N., Reitz, C., Jun, G., Naj, A., Beecham, G. W., Liu, Z., Carney, R. M., Vance, J. M., Cuccaro, M. L., et al., Mar 2016. Global and local ancestry in African-Americans: Implications for alzheimer's disease risk. Alzheimer's & Dementia 12 (3), 233–243.

Iughetti, P., Suzuki, O., Godoi, P., Alves, V., Sertie, A., Zorick, T., Soares, F., Camargo, A., Moreira, E., di Loreto, C., Moreira-Filho, C., Simpson, A., Oliva, G., Passos-Bueno, M., 2001. A polymorphism in endostatin, an angiogenesis inhibitor, predisposes for the development of prostatic adenocarcinoma. Cancer Research 61(20), 7375–7378.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., Jan 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research 42 (Database issue), 199–205.

Keogh, E., Mueen, A., 2010. Curse of Dimensionality. Springer US, Boston, MA, pp. 257–258.

Kietzmann, T. C., Lange, S., Riedmiller, M., 2008. Incremental grlvq: Learning relevant features for 3d object recognition. Neurocomputing, 2868–2879.

Kira, K., Rendell, L. A., 1992. The feature selection problem: Traditional methods and a new algorithm. Em: Proceedings of the Tenth National Conference on Artificial Intelligence. AAAI'92. AAAI Press, pp. 129–134.

Kohonen, T., 1997. Self-organizing Maps. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Kononenko, I., 1994. Estimating attributes: analysis and extensions of relief. Em: European conference on machine learning. p. 171–182.

Martinetz, T. M., Berkovich, S. G., Schulten, K. J., Julho 1993. 'neural-gas' network for vector quantization and its application to time-series prediction. IEEE Transactions on Neural Networks 4 (4), 558–569.

McKay, M. D., Beckman, R. J., Conover, W. J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 (2), 239–245.

Miko, I., LeJeune, L., 2009. Essentials of genetics. Nature, Cambridge, MA.

Minelli, C., Grandi, A. D., Weichenberger, C. X., Gögele, M., Modenese, M., Attia, J., Barrett, J. H., Boehnke, M., Borsani, G., Casari, G., Fox, C. S., Freina, T., Hicks, A. A., Marroni, F., Parmigiani, G., Pastore, A., Pattaro, C., Pfeufer, A., Ruggeri, F., Schwienbacher, C., Taliun, D., Pramstaller, P. P., Domingues, F. S., Thompson, J. R., Feb 2013. Importance of different types of prior knowledge in selecting genome-wide findings for follow-up. Genetic Epidemiology 37 (2), 205–213.

Moore, J. H., Asselbergs, F. W., Williams, S. M., Feb 2010. Bioinformatics challenges for genome-wide association studies. Bioinformatics 26 (4), 445–455.

Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W., Ritchie, M. D., May 2008. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genetic Epidemiology 32 (4), 325–340.

Muzzey, D., Evans, E. A., Lieber, C., 2015. Understanding the Basics of NGS: From Mechanism to Variant Calling. Current Genetic Medicine Reports 3 (4), 158–165.

Myers, A., Holmans, P., Marshall, H., Kwon, J., Meyer, D., Ramic, D., Shears, S., Booth, J., DeVrieze, F. W., Crook, R., Hamshere, M., Abraham, R., Tunstall, N., Rice, F., Carty, S., Lillystone, S., Kehoe, P., Rudrasingham, V., Jones, L., Lovestone, S., Perez-Tur, J., Williams, J., Owen, M. J., Hardy, J., Goate, A. M., Dec 2000. Susceptibility locus for Alzheimer's disease on chromosome 10. Science 290 (5500), 2304–2305.

Nature-Education, 2010. A brief history of genetics: Defining experiments in genetics. Cambridge, MA.

NIH, Abril 2010. The new genetics. Publication N10, 662.

NLM, N. L. M., Janeiro 2017. Genetics home reference.
URL `https://ghr.nlm.nih.gov/`.

Onay, V. U., Briollais, L., Knight, J. A., Shi, E., Wang, Y., Wells, S., Li, H., Rajendram, I., Andrulis, I. L., Ozcelik, H., 2006. SNP-SNP interactions in breast cancer susceptibility. BMC Cancer 6, 114.

Pinto, L. F. R., Maio-Junho 2007. Suscetibilidade ao câncer ligada a genes de baixa penetrância: Desafios na busca de prevenção individualizada. Prática Hospitalar 51, 86–90.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., Sham, P. C., Sep 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81 (3), 559–575.

Ritchie, M., Motsinger, A. A., 2005. Multifactor dimensionality reduction for detecting gene-gene and gene-environment interactions in pharmacogenomics studies. Phamacogenomics 6(8), 823–34.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., Moore, J. H., Jul 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. American Journal of Human Genetics 69 (1), 138–147.

Ritchie, M. D., White, B. C., Parker, J. S., Hahn, L. W., Moore, J. H., Jul 2003. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. BMC Bioinformatics 4, 28.

Saeys, Y., Inza, I., Larrañaga, P., Oct 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23 (19), 2507–2517.

Sarlak, G., Htoo, H., Hernandez, J.-F., Iizasa, H., Checler, F., Konietzko, U., Song, W., Vincent, B., Jan 2016. Sox2 functionally interacts with Betaapp, the Betaapp intracellular domain and adam10 at a transcriptional level in human cells. Neuroscience 312, 153–164.

Sato, A., Yamada, K., 1996. Generalized learning vector quantization. Advances in Neural Information Processing Systems 8, 423–429, touretzky DS, Mozer MC, Hasselmo ME (eds).

Saykin, A. J., Shen, L., Foroud, T. M., Potkin, S. G., Swaminathan, S., Kim, S., Risacher, S. L., Nho, K., Huentelman, M. J., Craig, D. W., Thompson, P. M., Stein, J. L., Moore, J. H., Farrer, L. A., Green, R. C., Bertram, L., Jack, C. R., Weiner, M. W., May 2010. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. Alzheimers Dement 6 (3), 265–273.

Schneider, P., Biehl, M., Schleif, F.-M., Hammer, B., 2007. Advanced metric adaptation in Generalized LVQ for classification of mass spectrometry data. Em: Proceedings of 6th International Workshop on Self-Organizing Maps.

Schwender, H., Ickstadt, K., Jan 2008. Identification of SNP interactions using logic regression. Biostatistics 9, 187–198.

Shang, J., Sun, Y., Liu, J.-X., Xia, J., Zhang, J., Zheng, C.-H., 2016. Cinoedv: a co-information based method for detecting and visualizing n-order epistatic interactions. BMC Bioinformatics 17 (1), 214.

Shang, J., Zhang, J., Sun, Y., Liu, D., Ye, D., Yin, Y., 2011. Performance analysis of novel methods for detecting epistasis. BMC Bioinformatics 12 (1), 475.

Shen, L., Kim, S., Risacher, S. L., Nho, K., cols, Nov 2010. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. Neuroimage 53 (3), 1051–1063.

Stankiewicz, P., Lupski, J. R., 2010. Structural variation in the human genome and its role in disease. Annual Review of Medicine 61, 437–455.

Strickert, M., Bojer, T., Hammer, B., 2001. Generalized Relevance LVQ for Time Series. Em: Dorffner, G., Bischof, H., Hornik, K. (Eds.), Artificial Neural Networks. International Conference. Proceedings. Lecture Notes in Computer Science, 2130. Springer, pp. 677–683.

Strickert, M., Seiffert, U., Sreenivasulu, N., Weschke, W., Villmann, T., Hammer, B., 2006. Generalized relevance LVQ (GRLVQ) with correlation measures for gene expression analysis. Neurocomputing 69 (7-9).

Teng, S., Madej, T., Panchenko, A., Alexov, E., Mar 2009. Modeling effects of human single nucleotide polymorphisms on protein-protein interactions. Biophysical Journal 96 (6), 2178–2188.

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., de Andrade, M., Doheny, K. F., Haines, J. L., Hayes, G., Jarvik, G., Jiang, L., Kullo, I. J., Li, R., Ling, H., Manolio, T. A., Matsumoto, M., McCarty, C. A., McDavid, A. N., Mirel, D. B., Paschall, J. E., Pugh, E. W., Rasmussen, L. V., Wilke, R. A., Zuvich, R. L., Ritchie, M. D., Jan 2011. Quality control procedures for genome-wide association studies. Current Protocols in Human Genetics Chapter 1, Unit1.19.

Velez, D. R., White, B. C., Motsinger, A. A., S, B. W., MD, R., Williams, S. M., JH, M., 2007. A balanced accuracy metric for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genetic Epidemiology 31, 306–315.

Venter, J. C., Adams, M. D., Myers, E. W., et al., Feb 2001. The sequence of the human genome. Science 291 (5507), 1304–1351.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. S., Yu, W., Sep 2010a. Boost: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. American Journal of Human Genetics 87 (3), 325–340.

Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N. L. S., Yu, W., Jan 2010b. Predictive rule inference for epistatic interaction detection in genome-wide association studies. Bioinformatics 26 (1), 30–37.

Wang, J., Joshi, T., Valliyodan, B., Shi, H., Liang, Y., Nguyen, H. T., Zhang, J., Xu, D., 2015. A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies. BMC Genomics 16, 1011.

Wang, Y., Makedon, F. S., Ford, J. C., Pearlman, J., Dec 2004. Hykgene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics 21 (8), 1530–1537.

Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F., Mewes, H. W., Feb 2005. Gene selection from microarray data for cancer classification—a machine learning approach. Computational Biology and Chemistry 29 (1), 37–46.

Wei, W. H., Hemani, G., Haley, C. S., Nov 2014. Detecting epistasis in human complex traits. Nature Reviews Genetics 15 (11), 722–733.

Williamson, J., Goldman, J., Marder, K. S., Mar 2009. Genetic aspects of Alzheimer disease. Neurologist 15 (2), 80–86.

Yang, C., He, Z., Wan, X., Yang, Q., Xue, H., Yu, W., Feb 2009. SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. Bioinformatics 25 (4), 504–511.

Yang, C., Wan, X., Yang, Q., Xue, H., Yu, W., 2010. Identifying main effects and epistatic interactions from large-scale snp data via adaptive group lasso. BMC Bioinformatics 11 Suppl 1, S18.

Yu, L., Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research 5, 1205–1224.

Zhang, Y., Jan 2012. A novel bayesian graphical model for genome-wide multi-SNP association mapping. Genetic Epidemiology 36 (1), 36–47.

Zhang, Y., Liu, J. S., Sep 2007. Bayesian inference of epistatic interactions in case-control studies. Nature Genetics 39 (9), 1167–1173.

Zhao, Y., Clark, W. T., Mort, M., Cooper, D. N., Radivojac, P., Mooney, S. D., Sep 2011. Prediction of functional regulatory snps in monogenic and complex disease. Human Mutation 32 (10), 1183–1190.

Zhao, Z., Liu, H., 1991. Searching for interacting features. Em: Proceedings of the international joint conference on artificial intelligence. p. 1156–1167.

Apêndice

Tabela 1 Relação de SNPs e Genes identificados nos dados do ADNI

	SNP	Gene		SNP	Gene
1	rs2075650	TOMM40	39	rs2008691	CYP19A1
2	rs439401	APOE	40	rs11136000	CLU
3	rs12989701	BIN1	41	rs597668	EXOC3L2
4	rs405509	APOE	42	rs11264359	FDPS
5	rs689021	SORL1	43	rs1799986	LRP1
6	rs11767557	EPHA1	44	rs1800896	IL10
7	rs4935775	SORL1	45	rs868750	CHAT
8	rs8106922	TOMM40	46	rs1871045	BCAM
9	rs6859	PVRL2	47	rs767199	CYP19A1
10	rs669	A2M	48	rs2230806	ABCA1
11	rs4646957	IDE	49	rs4935774	SORL1
12	rs3865444	CD33	50	rs677909	PICALM
13	rs668156	HMGCS2	51	rs671	ALDH2
14	rs532208	HMGCS2	52	rs1554338	MAPK8IP1
15	rs3818361	CR1	53	rs7100623	IDE
16	rs6701713	CR1	54	rs1805087	MTR
17	rs4878104	DAPK1	55	rs732765	DLST
18	rs1143634	IL1B	56	rs989692	MME
19	rs3764650	ABCA7	57	rs1871047	PVRL2
20	rs7012010	CLU	58	rs1695	GSTP1
21	rs2276346	SORL1	59	rs2074877	MYH13
22	rs1378577	ABCG1	60	rs1049296	TF
23	rs213045	ECE1	61	rs1050565	BLMH
24	rs5167	APOC4	62	rs3827225	ABCG1
25	rs3800324	PGBD1	63	rs8070723	MAPT
26	rs13022344	TRAK2	64	rs2149632	IDE
27	rs2070045	SORL1	65	rs11030104	BDNF
28	rs7070570	VR22	66	rs3810950	CHAT
29	rs3851179	PICALM	67	rs242557	MAPT
30	rs1468503	NPC2	68	rs1880676	CHAT
31	rs700518	CYP19A1	69	rs662	PON1
32	rs541458	PICALM	70	rs1441008	HMGCS2
33	rs556349	SORL1	71	rs4948288	ARID5B
34	rs570113	ACAD8	72	rs2227564	PLAU
35	rs1800562	HFE	73	rs2066718	ABCA1
36	rs3761740	HMGCR	74	rs6265	BDNF
37	rs708272	CETP	75	rs688	LDLR
38	rs1699102	SORL1	76	rs4934	SERPINA3

Tabela 2 Relação 1: dos 674 Genes identificados nos dados do ADNI

	Cr	ID	Gene		Cr	ID	Gene		Cr	ID	Gene
1	1	635	CR1	46	1	626	NGF	91	3	566	SST
2	1	107	IL10	47	1	320	PCSK9	92	3	738	TP63
3	1	246	HSD11B1	48	1	509	PMVK	93	3	144	AHSG
4	1	244	GSTM3	49	1	149	POU2F1	94	3	446	OGG1
5	1	256	NTRK1	50	1	242	PRDX6	95	3	77	CCR2
6	1	448	PARP1	51	1	510	PRKAB2	96	3	73	MME
7	1	188	TP73	52	1	252	REN	97	3	72	TF
8	1	578	ECE1	53	1	335	RGS4	98	3	524	CAV3
9	1	108	CRP	54	1	511	RXRG	99	3	78	CCR5
10	1	515	DHCR24	55	1	434	SORT1	100	3	662	DRD3
11	1	114	SOAT1	56	1	663	TARDBP	101	3	504	RFTN1
12	1	243	AGT	57	1	255	TNFRSF14	102	3	557	SLC2A2
13	1	567	CFH	58	1	253	TNFRSF1B	103	3	160	TGM4
14	1	450	CTSS	59	1	254	TNFRSF4	104	4	580	COL25A1
15	1	251	HSPG2	60	1	158	TNFRSF8	105	4	101	APBB2
16	1	321	PTGS2	61	1	396	USF1	106	4	381	LRPAP1
17	1	116	CHRNB2	62	2	708	BIN1	107	4	600	LRAT
18	1	507	FDPS	63	2	723	EPC2	108	4	375	ALB
19	1	126	GSTM1	64	2	725	EPHA4	109	4	162	ART3
20	1	508	HMGCS2	65	2	731	HPCAL1	110	4	263	CASP3
21	1	71	NCSTN	66	2	587	LHCGR	111	4	264	CASP6
22	1	451	FAM63A	67	2	680	LRP2	112	4	526	CPE
23	1	197	MTR	68	2	453	TRAK2	113	4	161	CSN1S1
24	1	452	LMNA	69	2	469	EIF2AK2	114	4	127	CXCL1
25	1	55	PSEN2	70	2	53	IL1B	115	4	386	CXCL10
26	1	216	APH1A	71	2	191	IL1RN	116	4	528	FABP2
27	1	505	APOA1BP	72	2	49	IL1A	117	4	553	GC
28	1	506	APOA2	73	2	190	ABCA12	118	4	120	IL8
29	1	249	CAMK1G	74	2	189	ABCG5	119	4	552	PPARGC1A
30	1	178	CLCNKB	75	2	712	ADAM17	120	4	575	RFC1
31	1	609	COG2	76	2	239	APOB	121	4	79	SNCA
32	1	150	COL11A1	77	2	257	CASP8	122	4	435	SORCS2
33	1	422	CSF1	78	2	159	EFEMP1	123	4	340	SRP72
34	1	241	DVL1	79	2	538	GRB14	124	4	380	UCHL1
35	1	573	F11R	80	2	543	HK2	125	5	640	ARSB
36	1	152	FCER1G	81	2	545	IRS1	126	5	405	HMGCR
37	1	419	GBA	82	2	258	NDUFS1	127	5	637	EFNA5
38	1	151	GBP2	83	2	547	NEUROD1	128	5	564	WWC1 (KIBRA)
39	1	240	GMEB1	84	2	97	SLC11A1	129	5	270	PIK3R1
40	1	245	GSTM4	85	2	558	SOS1	130	5	579	ADRB2
41	1	69	HTR6	86	2	259	TANK	131	5	265	CAST
42	1	250	IRF6	87	3	74	BCHE	132	5	121	CD14
43	1	179	LCK	88	3	262	GSK3B	133	5	266	CRHBP
44	1	90	LRP8	89	3	261	APOD	134	5	80	FGF1
45	1	70	MTHFR	90	3	96	PPARG	135	5	503	HMGCS1

Tabela 3 Relação 2: dos 674 Genes identificados nos dados do ADNI

	Cr	ID	Gene		Cr	ID	Gene		Cr	ID	Gene
136	5	163	HMMR	181	6	645	MICB	226	8	87	LPL
137	5	267	IL4	182	6	202	NOTCH4	227	8	208	NAT2
138	5	268	NDUFS4	183	6	393	PLG	228	8	206	PLAT
139	5	269	NRC31	184	6	406	PPARD	229	8	611	ADAM9
140	5	586	PPP2R2B	185	6	625	PPP1R10	230	8	275	ADRA1A
141	5	502	PRKAA1	186	6	322	PSMB9	231	8	522	ADRB3
142	5	192	SLC6A3	187	6	556	RPS6KA2	232	8	204	CHRNA2
143	5	416	SNCAIP	188	6	497	RXRB	233	8	205	CHRNA6
144	6	715	AGPAT1	189	6	363	SOD2	234	8	276	CRH
145	6	593	BAT1	190	6	378	TAP1	235	8	154	DPYS
146	6	442	C4A	191	6	648	TBP	236	8	560	ENPP2
147	6	443	C4B	192	6	415	TREM2	237	8	529	FABP4
148	6	447	F13A1	193	6	574	TUBB	238	8	209	NAT1
149	6	596	MICA	194	7	724	EPHA1	239	8	277	NCOA2
150	6	739	ZNF292	195	7	718	ATP6V0A4	240	8	198	NRG1
151	6	271	AGER	196	7	722	DGKB	241	8	207	WRN
152	6	379	TAP2	197	7	675	GWA-7q31.1	242	9	401	DFNB31
153	6	81	TNF	198	7	50	IL6	243	9	402	POMT1
154	6	720	CD2AP	199	7	735	NXPH1	244	9	89	ABCA2
155	6	182	ESR1	200	7	334	PON3	245	9	615	IL33
156	6	456	LOC651924	201	7	583	RELN	246	9	238	TLR4
157	6	589	NEDD9	202	7	403	DLD	247	9	638	PRUNE2
158	6	193	VEGF	203	7	146	CDK5	248	9	581	RXRA
159	6	382	HLA	204	7	639	MAGI2	249	9	88	ABCA1
160	6	604	ATXN1	205	7	141	PON1	250	9	103	APBA1
161	6	86	HFE	206	7	274	PPP1R3A	251	9	572	GWA-9p24.3
162	6	455	UBD	207	7	223	SERPINE1	252	9	279	TRAF2
163	6	454	PGBD1	208	7	457	GWA-7p15.2	253	9	400	DAPK1
164	6	501	ACAT2	209	7	115	NOS3	254	9	644	CDKN2BAS
165	6	500	APOBEC2	210	7	408	ABCB1	255	9	563	VCP
166	6	194	APOM	211	7	140	ACHE	256	9	237	VLDLR
167	6	445	C2	212	7	521	ADCYAP1R1	257	9	570	GOLM1
168	6	444	CFB	213	7	184	AHR	258	9	102	UBQLN1
169	6	499	CYP39A1	214	7	582	CAV1	259	9	667	CDKN2A
170	6	527	ENPP1	215	7	153	CD36	260	9	337	DBH
171	6	498	FLOT1	216	7	165	FGL2	261	9	530	FBP1
172	6	531	FOXO3A	217	7	534	GCK	262	9	652	GRIN3A
173	6	272	FYN	218	7	537	GRB10	263	9	599	HSPA5
174	6	362	GLO1	219	7	164	LAMB1	264	9	412	IFT74
175	6	535	GLP1R	220	7	631	miRNA-29a/b	265	9	278	NDUFA8
176	6	129	HSPA1A	221	7	145	NPY	266	9	222	NTRK2
177	6	128	HSPA1B	222	7	273	NUDT1	267	9	333	OPRS1
178	6	441	HSPA1L	223	7	548	PAX4	268	9	166	PSMB7
179	6	195	LPA	224	7	142	PON2	269	9	352	PTENP1
180	6	82	LTA	225		323	CLU	270	9	664	SEMA4D

Tabela 4 Relação 3: dos 674 Genes identificados nos dados do ADNI

	Cr	ID	Gene		Cr	ID	Gene		Cr	ID	Gene
271	10	716	ARID5B	316	10	365	CYP2C8	361	10	437	SORCS3
272	10	717	ARL5B	317	10	474	DNAJC12	362	10	482	SUPV3L1
273	10	649	CACNB2	318	10	342	ECHS1	363	10	629	TACR2
274	10	357	LOC439999	319	10	169	EGR2	364	10	598	TCF7L2
275	10	655	TET1	320	10	383	ENTPD7	365	10	236	TFAM
276	10	666	C10orf112	321	10	122	FLJ20445	366	10	366	TLL2
277	10	281	CDC2	322	10	355	GOT1	367	10	483	TSPAN15
278	10	473	DKK1	323	10	57	GSTO1	368	10	484	UBE2D1
279	10	356	PCGF5	324	10	58	GSTO2	369	10	392	VCL
280	10	358	ALDH18A1	325	10	654	HECTD2	370	10	387	VPS26A
281	10	470	ALOX5	326	10	542	HK1	371	10	372	WNT8B
282	10	585	CALHM1	327	10	370	HPSE2	372	10	485	ZWINT
283	10	385	DNMBP	328	10	343	KCNMA1	373	11	733	MS4A4E
284	10	360	hCG2039140	329	10	123	KIF11	374	11	734	MS4A6A
285	10	248	LRRTM3	330	10	344	LIPF	375	11	562	GAB2
286	10	650	PTPLA	331	10	345	MINPP1	376	11	199	APOA4
287	10	480	SGPL1	332	10	346	MYST4	377	11	732	MS4A4A
288	10	359	SORCS1	333	10	391	NDST2	378	11	438	SORL1
289	10	201	CHAT	334	10	282	NDUFB8	379	11	423	APOA5
290	10	56	PLAU	335	10	373	NEURL	380	11	424	HBG2
291	10	354	ANK3	336	10	475	NEUROG3	381	11	449	IL18
292	10	458	EBF3	337	10	347	OAT	382	11	135	MMP1
293	10	512	ADAM12	338	10	348	OPTN	383	11	636	PICALM
294	10	124	HHEX	339	10	168	PDCD11	384	11	44	FE65
295	10	45	VR22	340	10	368	PGAM1	385	11	134	GSTP1
296	10	200	FAS	341	10	661	PITRM1	386	11	132	MMP3
297	10	185	LIPA	342	10	476	PLCE1	387	11	289	APOA1
298	10	186	CH25H	343	10	349	PNLIPRP1	388	11	287	MAPK8IP1
299	10	61	IDE	344	10	364	PPP1R3C	389	11	493	ACAD8
300	10	384	ABCC2	345	10	167	PPP3CB	390	11	51	BACE1
301	10	496	ACF	346	10	388	PRG1	391	11	42	CTSD
302	10	338	ACTA2	347	10	59	PRSS11	392	11	109	BDNF
303	10	280	ADRA2A	348	10	350	PSAP	393	11	516	ABCC8
304	10	130	ADRB1	349	10	351	PTEN	394	11	494	ABCG4
305	10	471	ANXA8	350	10	477	RASSF4	395	11	131	APOC3
306	10	495	AP3M1	351	10	478	SAR1A	396	11	283	BIRC3
307	10	472	APBB1IP	352	10	353	SCD	397	11	284	CASP4
308	10	339	BAG3	353	10	479	SEC24C	398	11	226	CAT
309	10	187	BICC1	354	10	369	SFRP5	399	11	285	CHRM1
310	10	390	CAMK2G	355	10	513	SH3PXD2A	400	11	440	CNTF
311	10	341	CHST3	356	10	481	SIRT1	401	11	590	DRD4
312	10	371	COX15	357	10	224	SLC18A3	402	11	286	FADD
313	10	594	CXCL12	358	10	367	SLIT1	403	11	532	GAL
314	10	228	CYP17A1	359	10	374	SLK	404	11	544	INPPL1
315	10	324	CYP2C19	360	10	76	SNCG	405	11	110	INS

Tabela 5 Relação 4: dos 674 Genes identificados nos dados do ADNI

	Cr	ID	Gene		Cr	ID	Gene		Cr	ID	Gene
406	11	546	KCNJ11	451	12	610	PZP	496	15	404	IREB2
407	11	549	PDE3B	452	12	221	SCARB1	497	15	217	APH1B
408	11	642	RTN3	453	12	98	SLC11A2	498	15	421	LIPC
409	11	75	TCN1	454	12	293	TNFRSF1A	499	15	148	ADAM10
410	11	133	TPH1	455	13	669	PPARA	500	15	211	CHRNA3
411	11	288	UCP2	456	13	713	ALOX5AP	501	15	213	CHRNB4
412	11	492	ZNF202	457	13	418	ATXN8OS	502	15	229	CYP1A1
413	12	48	OLR1	458	13	525	CDX2	503	15	407	IGF1R
414	12	47	TFCP2	459	13	218	HTR2A	504	15	595	SMAD3
415	12	328	GAPDH	460	14	93	KNS2	505	16	414	MEFV
416	12	326	NCAPD2	461	14	623	NGB	506	16	646	UBE2I
417	12	117	NOS1	462	14	439	NP	507	16	119	CETP
418	12	327	PPM1H	463	14	336	SEL1L	508	16	298	ERCC4
419	12	399	ATF7	464	14	297	SERPINA1	509	16	614	HMOX2
420	12	656	CHD4	465	14	147	ESR2	510	16	299	HSD11B2
421	12	601	FAM113B	466	14	459	GWA-14q32.13	511	16	413	NQO1
422	12	657	LPAR5	467	14	105	CYP46	512	16	180	PHKG2
423	12	514	LRP6	468	14	519	SOS2	513	16	410	TMC5
424	12	660	TAPBPL	469	14	622	TMEM63C	514	16	436	VPS35
425	12	737	TMEM132C	470	14	230	SERPINA3	515	17	613	CDK5R1
426	12	290	ALDH2	471	14	84	PSEN1	516	17	665	COX10
427	12	100	GNB3	472	14	490	NPC2	517	17	726	GLOD4
428	12	220	VDR	473	14	231	DLST	518	17	95	SREBF1
429	12	641	CAND1	474	14	736	RGS6	519	17	301	PNMT
430	12	668	IGF1	475	14	411	ARID4A	520	17	94	MPO
431	12	132	M6PR	476	14	294	CTSG	521	17	394	SERPINF2
432	12	325	PKP2P1	477	14	295	FOS	522	17	584	GRN
433	12	46	LRP1	478	14	617	GSTZ1	523	17	302	TP53
434	12	52	A2M	479	14	603	GWA-14q31.2	524	17	92	CCL2
435	12	491	APOBEC1	480	14	618	HIF1A	525	17	461	TNK1
436	12	608	A2MP	481	14	568	MTHFD1	526	17	462	MYH13
437	12	520	ABCC9	482	14	296	NFKBIA	527	17	329	SLC6A4
438	12	170	AVPR1A	483	14	619	NUMB	528	17	104	BLMH
439	12	397	C12orf41	484	14	172	PCK2	529	17	125	ACE
440	12	196	C1R	485	14	620	POMT2	530	17	232	MAPT (TAU)
441	12	398	CCNT1	486	14	621	SGPP1	531	17	174	ADORA2B
442	12	536	GPD1	487	14	624	TMED8	532	17	576	CCL3
443	12	292	GRIN2B	488	14	171	ZAP128	533	17	91	CCL5
444	12	561	GYS2	489	15	210	CHRFAM7A	534	17	627	CCL8
445	12	658	IFFO1	490	15	420	MEF2A	535	17	377	FTSJ3
446	12	291	IFNG	491	15	571	GWA-15p21.2	536	17	539	GRB2
447	12	647	LRRK2	492	15	183	CYP19A1	537	17	540	GRB7
448	12	550	PFKM	493	15	460	AGC1	538	17	673	KIF18B
449	12	137	PLA2G1B	494	15	214	CHRNA7	539	17	155	MYH8
450	12	553	PPP1CC	495	15	670	CSK	540	17	300	NGFR

Tabela 6 Relação 5: dos 674 Genes identificados nos dados do ADNI

	Cr	ID	Gene		Cr	ID	Gene		Cr	ID	Gene
541	17	118	NOS2A	586	19	307	NDUFA3	631	22	672	CECR2
542	17	555	PYY	587	19	308	NDUFB7	632	22	331	CYP2D6
543	17	376	TANC2	588	19	309	NDUFS7	633	22	111	MIF
544	17	559	TCF2	589	19	260	NR1H2	634	22	176	MTP18
545	17	628	THRA	590	19	215	PEN2	635	22	315	NDUFA6
546	17	173	TM4SF5	591	19	554	PPP2R1A	636	22	643	PPIL2
547	18	409	DSC1	592	19	175	RPS15	637	22	203	TCN2
548	18	651	LOC388458	593	19	395	USF2	638	X	316	AR
549	18	304	MC2R	594	19	468	XRCC1	639	X	318	MAOA
550	18	303	BCL2	595	20	212	CHRNA4	640	X	565	OTC
551	18	632	TTR	596	20	465	PCK1	641	X	602	PCDH11X
552	19	85	APOE	597	20	68	PRND	642	X	177	ACSL4
553	19	463	TOMM40	598	20	67	PRNP	643	X	488	EBP
554	19	714	ABCA7	599	20	66	CST3	644	X	317	HTR2C
555	19	247	APOC1	600	20	313	CD40	645	X	319	MAOB
556	19	605	CD33	601	20	312	GMEB2	646	X	332	TIMP1
557	19	634	APOC4	602	20	517	HNF4A	647	X	633	TNMD
558	19	606	PVRL2	603	20	551	PLCG1	648	MT	683	MT-COI
559	19	709	EXOC3L2	604	21	467	S100B	649	MT	693	MT-ND4
560	19	305	APOC2	605	21	235	APP	650	MT	695	MT-ND4L
561	19	677	BCAM	606	21	427	NCAM2	651	MT	696	MT-ND6
562	19	612	CARD8	607	21	225	CBS	652	MT	697	MT-RNR1
563	19	679	CLPTM1	608	21	486	ABCG1	653	MT	686	MT-CYB
564	19	464	GALP	609	21	54	BACE2	654	MT	681	MT-ATP6
565	19	113	ICAM1	610	21	432	C21ORF55	655	MT	682	MT-ATP8
566	19	330	GAPDHS	611	21	431	C21ORF63	656	MT	684	MT-CO2
567	19	99	TGFB1	612	21	432	DOPEY2	657	MT	685	MT-CO3
568	19	106	LDLR	613	21	429	DYRK1A	658	MT	687	MT-DLOOP
569	19	219	PIN1	614	21	430	KCNJ6	659	MT	707	MT-TH
570	19	181	AKAP8	615	21	487	LSS	660	MT	706	MT-TS2
571	19	523	AKT2	616	21	157	MCM3AP	661	MT	702	MT-L2
572	19	678	APOC1P1	617	21	630	OLIG2	662	MT	692	MT-ND3
573	19	310	BCL3	618	21	426	PRSS7	663	MT	705	MT-TR
574	19	676	CBLC	619	21	428	RUNX1	664	MT	704	MT-TT
575	19	306	CYP4F3	620	21	425	SAMSN1	665	MT	694	MT-ND5
576	19	569	DNM2	621	22	719	CCDC134	666	MT	690	MT-ND1
577	19	389	ERCC2	622	22	234	SEPT3	667	MT	699	MT-TG
578	19	156	GNA11	623	22	671	SYN3	668	MT	688	MT-haplo
579	19	653	GRIN3B	624	22	143	GSTT1	669	MT	691	MT-ND2
580	19	541	GYS1	625	22	227	COMT	670	MT	703	MT-TQ
581	19	518	INSR	626	22	417	HMOX1	671	MT	689	MT-NC7
582	19	361	KLK1	627	22	466	BCR	672	MT	698	MT-RNR2
583	19	588	LHB	628	22	112	PPARA	673	MT	700	MT-TI
584	19	311	LIPE	629	22	314	ACO2	674	MT	701	MT-TK
585	19	489	LRP3	630	22	139	ADORA2A				