UNIVERSIDADE FEDERAL DE JUIZ DE FORA
PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

UMA METODOLOGIA PARA DETECÇÃO DE INTERAÇÕES EPISTÁTICAS EM ESTUDOS DE ASSOCIAÇÃO

Igor Magalhães Ribeiro

Juiz de Fora
Junho de 2019

Igor Magalhães Ribeiro

Uma metodologia para detecção de interações epistáticas em estudos de associação

Tese apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Doutor em Modelagem Computacional.

Orientador: Prof. D.Sc. Carlos Cristiano Hasenclever Borges

Coorientador: Prof. D.Sc. Wagner Antonio Arbex

Juiz de Fora

2019

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Ribeiro, Igor Magalhães .
Uma metodologia para detecção de interações epistáticas em estudos de associação / Igor Magalhães Ribeiro. -- 2019.
190 f.

Orientador: Carlos Cristiano Hasenclever Borges
Coorientador: Wagner Antonio Arbex
Tese (doutorado) - Universidade Federal de Juiz de Fora, ICE/Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2019.

1. Bioinformática. 2. GWAS. 3. Inteligência Computacional. 4. Aprendizagem de Máquina. 5. Programação Genética. I. Borges, Carlos Cristiano Hasenclever, orient. II. Arbex, Wagner Antonio, coorient. III. Título.

Igor Magalhães Ribeiro

Uma metodologia para detecção de interações epistáticas em estudos de associação

Tese apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Doutor em Modelagem Computacional.

Aprovada em 19 de Junho de 2019.

BANCA EXAMINADORA

Prof. D.Sc. Carlos Cristiano Hasenclever Borges - Orientador
Universidade Federal de Juiz de Fora

Prof. D.Sc. Wagner Antonio Arbex - Coorientador
Universidade Federal de Juiz de Fora

Prof. D.Sc. Gloria Regina Franco
Universidade Federal de Minas Gerais

Prof. D.Sc. Fabrízzio Condé de Oliveira
Universidade Federal Fluminense

Prof. D.Sc. Raul Fonseca Neto
Universidade Federal de Juiz de Fora

Prof. D.Sc Heder Soares Bernardino
Universidade Federal de Juiz de Fora

AGRADECIMENTOS

À Universidade Federal de Juiz de Fora, a CAPES e ao Programa de Pós-graduação em Modelagem Computacional, por permitirem a realização deste trabalho de doutorado e, por me proporcionarem as condições para o andamento do mesmo.

Aos meus orientadores Prof. Dr .Carlos Cristiano Hasenclever Borges e Prof. Dr. Wagner Antonio Arbex, pelo ensinamento e pela orientação.

Aos membros da banca, Prof. Dra. Glória Regina Franco, Prof. Dr. Fabrízzio Condé de Oliveira, Prof Dr. Héder Soares Bernardino, ao Prof Dr. Raul Fonseca Neto, Prof. Dr. Fabyano Fonseca e Silva, Prof. Dr. Moysés Nascimento, Prof. Dr. Itamar Leite de Oliveira e Prof. Dr. Saulo Moraes Villela pelo interesse demonstrado em participarem da minha banca e, assim, me concederem o privilégio de poder contar com suas avaliações.

Aos meus colegas, professores e funcionários da pós graduação, em especial a Bruno Zonovelli, Rafael Veiga, Marcelo Ribeiro, Aldemon Bonifácio, Andréz Valdez, William Yamashita, Alexandre Cançado, Grasiele Duarte, Patricia Fonseca, Rafaelle Finotti, Erick Mario, Érica Carvalho, Jesuliana Ulysses, Luiz Aquino, Lucas Meneses e Dênis Vargas pelas conversas, café e simpatia durante esses últimos anos.

A toda minha família - aos quais agradeço na pessoa de minha mãe, Maria de Fátima Magalhães, meu pai Dinei Ribeiro, meu irmão Hugo Magalhães e meus tios Rubens Magalhães e Sérgio Magalhães pelo apoio e incentivo em todos os momentos.

Aos amigos, em especial Leandro de Deus, Ana Beatriz Vilhena, Gian Stopa, Rafael Ragazzi, Thiago Batista, Rafael Marques, Thiago Roberto, Marcus Vargas, Ênio Oliveira, Abel Leitão, Fabiano Macedo, Jéssica Farineli, Juliana Matos, Rafael Bouças, Gláucio Paiva, Lílian Mazetti, Izaias Vilarino, Ana Maria Caula e Fabricio Brandi.

'A natureza é nossa casa e na natureza estamos em casa. Este mundo estranho, diversificado e assombroso que exploramos, onde o espaço se debulha, o tempo não existe e as coisas podem não estar em lugar algum, não é algo que nos afasta de nós: é somente aquilo que nossa natural curiosidade nos mostra da nossa casa. Da trama da qual somos feitos nós mesmos. Somos feitos da mesma poeira de estrelas de que são feitas as coisas, e quer quando estamos imersos na dor, quer quando rimos e a alegria resplandece, não fazemos mais do que ser aquilo que não podemos deixar de ser: uma parte do nosso mundo'

(Carlo Rovelli)

RESUMO

Estudos de associação genômica ampla (GWAS) buscam identificar marcadores moleculares do tipo SNP que influenciam um determinado fenótipo de interesse, como por exemplo características específicas ou doenças. Os SNPs são responsáveis pela formação de alelos sendo esse tipo de marcador utilizado para identificar um *lócus* que pode representar uma correlação próxima a um gene ou a própria mutação. Para determinar os mecanismos genéticos que influenciam o fenótipo são utilizados milhares ou até centenas de milhares de SNPs que são genotipados à partir de dois grupos de indivíduos: os que expressam e os que não expressam o fenótipo respectivamente, estudo conhecido como caso-controle. A causa de algumas doenças complexas como câncer cervical, leucemia, diabetes do tipo I e II envolvem múltiplos genes que co-atuam para expressar esse fenótipo, essa interação é denominada de epistasia. Nesse tipo de interação um gene pode inibir ou potencializar o efeitos dos demais. O entendimento adequado desse processo passa pela determinação do mapeamento não-linear entre o genótipo e o fenótipo. Além da complexidade de encontrar os SNPs causais envolvidos em interações epistáticas, a maior parte das doenças complexas apresentam baixa a herdabilidade. Desta forma, quanto menor a herdabilidade, maior a influência de fatores ambientais e menor a explicação genotípica do fenótipo. A maior parte dos estudos de GWAS utilizam testes estatísticos de hipóteses com valor-p de cada SNP presente no conjunto de dados, sendo estes estudos frequentemente baseados em modelos de regressão. Porém, esse tipo de modelo é eficiente para capturar apenas efeitos marginais ou efeitos de ações gênicas aditivas, ou seja, casos onde a relação genótipo-fenótipo é linear. Neste trabalho é proposta uma metodologia capaz de encontrar relações não lineares entre genótipo-fenótipo em bases de dados com grandes quantidades de SNPs. A metodologia também foi desenvolvida para lidar com diferentes níveis de herdabilidade. O modelo proposto é composto de três etapas distintas. A primeira etapa é responsável pela identificação e seleção de subgrupos de SNPs de interesse. Assim, o conjunto de dados é particionado em pequenos grupos de tamanho fixo e todas as possíveis permutações intra-grupos são classificadas através de validação cruzada aplicada em um método de comitê de classificadores do tipo *boosting*. Os grupos que apresentarem marcadores com maior potencial explicativo são selecionados para a fase posterior. Na segunda etapa, um processo de ranqueamento dos marcadores

SNPs selecionados a partir dos subgrupos definidos na primeira etapa, um modelo de floresta randômica com potencial para capturar a relevância dos SNPs avaliados é aplicado nesta fase do processo. Na terceira etapa, o ranqueamento dos marcadores serve como referência para a geração de uma população de marcadores, que servirá de base para a aplicação de um método evolucionista de programação genética que tem como objetivo determinar possíveis associações entre os SNPs ranqueados. Assim, ao final do processo, são apresentadas as relações genotípicas que expressam o fenótipo de interesse a partir de interações epistáticas, baseando-se na interpretabilidade das regras geradas. A metodologia proposta foi comparada com outros modelos existentes na literatura, inclusive com o método referência conhecido como MDR, uma variação do MDR com um método de inicialização conhecido como ReliefF e o GPAS. Foram realizados diversos experimentos com bases de dados simuladas, dentre eles, conjuntos de dados compostos de 100, 1000 e 10000 marcadores, com diferentes níveis de herdabilidade variando de 0.4 e 0.1 e MAF de 0.4 e 0.2. O método foi analisado em dados que apresentam epistasia sem efeito principal em 70 modelos conhecidos na literatura com herdabilidade variando de 0.4 até 0.01. Também foram executados experimentos com interações entre mais de dois SNPs e experimentos com conjuntos de dados com um número expressivo de SNPs. Os resultados indicam que a utilização da metodologia é promissora se comparada com outros modelos na literatura de GWAS.

Palavras-chave: Bioinformática. GWAS. Inteligência Computacional. Aprendizagem de Máquina. Programação Genética.

ABSTRACT

Genomic Wide Association Studies (GWAS) aims to identify SNPs that influence a particular phenotype, such as specific characteristics or diseases. SNPs are responsible for allele formation and this markers are used to identify a *lócus* that may represent a close correlation to a gene or the mutation itself. To determine the genetic mechanisms that influence the phenotype are used thousands or even hundreds of thousands of SNPs that are genotyped from two groups of individuals: case and control. The reason behind some complex diseases such as cervical cancer, leukemia, type I and II diabetes involve multiple genes combining to express this phenotype. This interaction is known as epistasis. In epistasis a gene can inhibit or potentiate the effects of the other. From the statistical point of view, the objective is to find a non-linear mapping between the genotype and the phenotype. In addition to the complexity of finding causal SNPs involved in epistatic interactions, most complex diseases have low heritability. Thus, the lower the heritability, the greater the influence of environmental factors and the less the genotype explanation of the phenotype. Most of GWAS use statistical tests of p-value hypotheses of each SNP present in the data set. These studies are often based on regression models. However, this type of model is efficient to capture only marginal effects or effects of additive gene actions. Cases where the genotype-phenotype relationship is linear. This work proposes a methodology capable of finding non-linear relationships between genotype-phenotype in data sets with large amounts of SNPs. The approach was also developed to deal with different levels of heritability. The proposed model is composed of three distinct steps. The first step is responsible for identifying and selecting subgroups of significant SNPs. The dataset is partitioned into small fixed-size groups and all possible permutations of each group are sorted by cross-validation by a ensemble method of boosting classifiers. The best groups are selected for the later stage. In the second step a process of ranking the selected SNPs from the subgroups defined in the first step is performed. A random forest model with potential to capture the relevance of the SNPs evaluated is applied at this stage of the process. The ranking of markers serves as a reference for the generation of a population markers, which will serve as the basis for the application of an evolutionary method of genetic programming that aims to determine possible associations between the SNPs ranked. At the end of the process, the genotypic relations that express

the phenotype of interest from epistatic interactions are presented, based on the interpretability of the generated rules. The proposed methodology was compared with other models in the literature, including the reference method known as MDR, a variation of the MDR with an initialization method known as ReliefF and GPAS. Several experiments were carried out with simulated datasets, including data sets composed of 100, 1000 and 10000 SNPs with different levels of heritability varying from 0.4 to 0.1 and MAF of 0.4 and 0.2. The method was analyzed in data presenting epistasis without main effect in 70 models known in the literature with heritability ranging from 0.4 to 0.01. We also performed experiments with interactions between more than two SNPs and experiments with data sets that present an expressive number of SNPs. The results denote the use of the methodology is promising compared to other models in GWAS literature.

Keywords: Bioinformatics. GWAS. Computational Intelligence. Machine Learning. Genetic Programming.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 Introdução

Durante a última década, estudos de associação ampla do genoma (do inglês, *Genome-Wide Association Studies*-GWAS) auxiliaram na descoberta e entendimento de doenças genéticas. Esse tipo de estudo visa a identificação de genes[1] ou marcadores associados à genes que podem contribuir para o aumento ou diminuição do risco de indivíduos desenvolverem uma determinada doença. O interesse nesse tipo de pesquisa cresceu devido ao barateamento no custo de genotipagem e o surgimento da tecnologia baseada em *chips*[2] de densidade que permitiram as medições de centenas de milhares de variações de sequências de DNA no genoma humano (MOORE; WHITE, 2007; BUSH; MOORE, 2012) . A forma mais comum de variação genômica ou marcador é conhecida como polimorfismo de base única (do inglês, *single nucleotide polymorphisms*-SNP). Essas variações são correspondentes a alternância (substituição, deleção ou inserção) dos nucleotídeos A, T, C e G em uma única posição do genoma. SNPs não são causadores diretos de doenças, entretanto a utilização desse tipo de marcador ajuda na localização no genoma dos reais fatores genéticos responsáveis pelo fenótipo. Através desses marcadores, regiões do genoma podem ser mapeadas e identificadas (ARBEX, 2009).

Em GWAS, onde os SNPs são utilizados como marcadores genéticos, o objetivo é identificar *loci*[3] associados com um fenótipo (doença ou característica de interesse). A hipótese é que a variação do fenótipo possa ser explicada a partir de um ou uma combinação desse tipo de marcador (GONDRO; WERF; HAYES, 2013), com a identificação desses marcadores podendo auxiliar na associação direta ou indireta dos mecanismos causais de uma determinada doença. De forma direta pode-se encontrar o marcador, não sendo assim necessário encontrar o verdadeiro *locus*[4] (SNPs podem apresentar alto desequilíbrio de ligação[5], dessa forma SNPs podem estar relacionados e podem conter parcialmente a

[1] Unidade fundamental da hereditariedade (PEARSON, 2006).

[2] Os *chips* de densidade foram desenvolvidos com o objetivo de se obter amostras de genotipagem com centenas de milhares ou milhões de marcadores de diversos indivíduos à partir de um único ensaio (CAETANO, 2009).

[3] Plural de *locus*.

[4] A palavra *locus* significa "lugar"em latim, representa um local fixo em um determinado cromossomo onde um gene ou marcador está localizado. A lista organizada de *loci* conhecidos para um cromossomo é chamada de mapa genético.

[5] Indica uma associação não aleatória entre SNPs, dessa forma, mesmo não sendo o marcador responsável pelo fenótipo, pode estar altamente correlacionado com o verdadeiro SNPs causal (PIERCE, 2011).

informação do verdadeiro *locus* da variante causal) e indiretamente indicando o próprio gene, via metabólica entre outras características biológicas. Um dos grandes desafios de GWAS é interpretar e entender a grande quantidade de informação obtida através do processo de sequenciamento, quando os marcadores moleculares são identificados (BUSH; MOORE, 2012; MANOLIO, 2010).

A Figura 1.1 mostra o número de publicações sobre GWAS entre o ano de 2005 e 2018, indicando que esta área têm se tornado alvo de interesse de diversos grupos de pesquisa. Pode-se observar que até o final de 2018, aproximadamente 3400 estudos de GWAS haviam sido publicados em periódicos científicos.

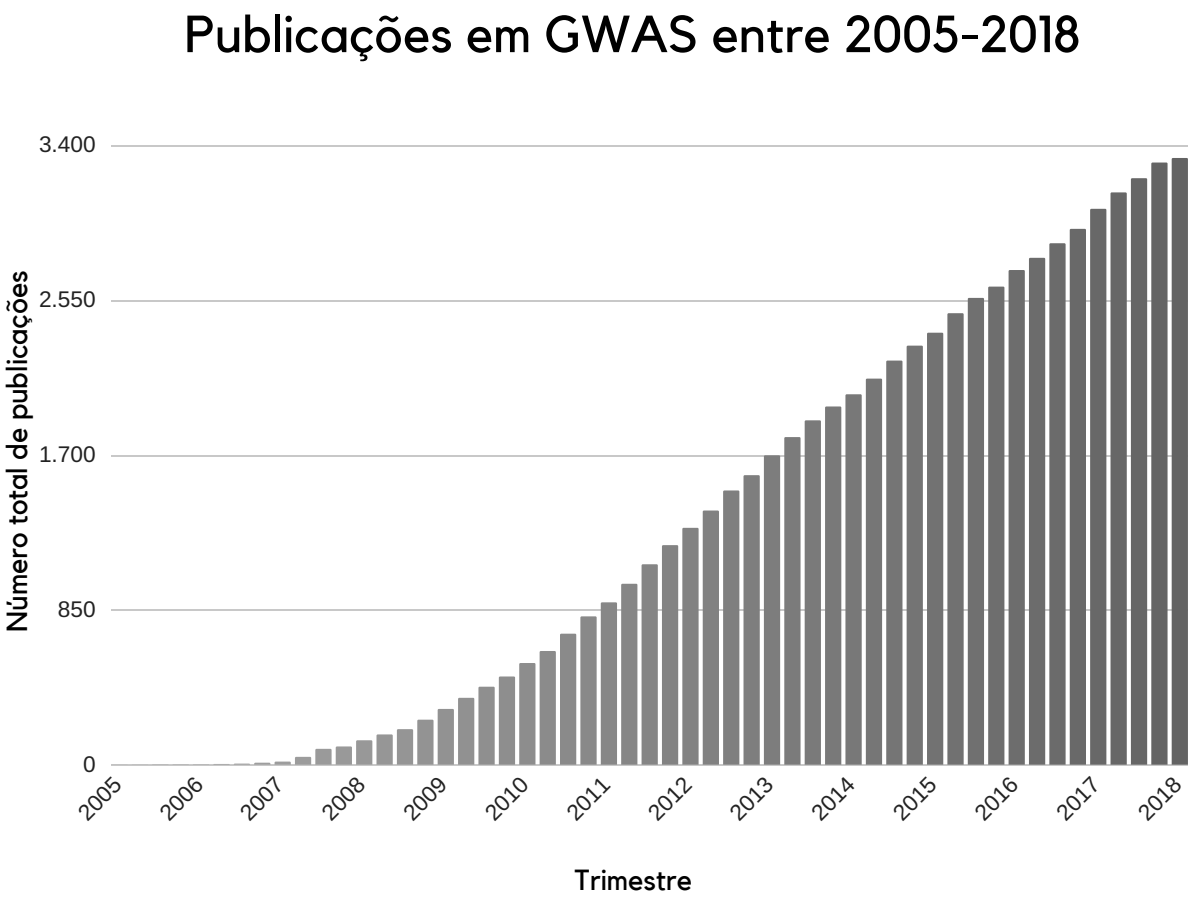


Figura 1.1: Número de publicações acumuladas sobre GWAS entre 2005 e 2018. Figura adaptada de (WELTER et al., 2014; MACARTHUR et al., 2016).

O National Human Genome Research Institute (NHGRI) e o the European Bioinformatics Institute (EMBL-EBI) mantém um catálogo sobre GWAS com diagramas sobre todas as associações encontradas no genoma humano. Cada associação apresenta um valor-p $\leq 5 \times 10^{-8}$ mapeados por localização cromossômica. Este catálogo serve como controle de qualidade sendo manualmente atualizado. Até 2014 aproximadamente 100 mil

marcadores foram inseridos (WELTER et al., 2014).

Como em GWAS o objetivo é associar marcadores a determinados fenótipos, frequentemente trabalha-se com fenótipos relacionados a doenças. Dessa forma, as doenças podem ser classificadas em: (i) apresentam traços mendelianos, em que estritamente seguem o padrão de herança mendeliana, onde tipicamente o causador do fenótipo vêm da mutação de um único gene que apresenta alta penetrância [6]. (ii) apresentam traços complexos (doenças complexas) que em contraste com as doenças mendelianas, são geralmente afetadas por mutações de uma combinação de múltiplos *loci* ao longo do genoma, onde cada *locus* contribui tipicamente com uma penetrância relativamente baixa para o fenótipo. Apesar das descobertas obtidas em grande parte dos estudos catalogados em (WELTER et al., 2014) serem promissoras, em muitos casos elas não são explicativas para o entendimento de doenças complexas. Diversos outros fatores devem ser levados em consideração em doenças desse tipo, como: efeitos epigenéticos, fatores ambientais, interações gene-ambiente e epistasia (PIERCE, 2011). [7]

1.1 Motivação

Métodos estatísticos utilizados tradicionalmente em estudos de GWAS são geralmente univariados ou monoatributo (associando diretamente um SNP com o fenótipo de interesse). Associações uni-variadas detectam marcadores fortemente associados ao fenótipo com alto valor marginal (um único marcador é responsável por parte ou toda a explicação da característica de interesse), sendo mais adequadas em estudos que buscam por interações gênicas aditivas[8], como no caso do trabalho apresentado em (EASTON; EELES, 2008), que foi o ponto de partida para a identificação de novos *loci* para doenças complexas como: câncer de próstata, câncer de mama, câncer de pulmão e melanoma. Entretanto, ao considerar a abordagem monoatributo, negligência-se interações gene-gene

[6] A penetrância corresponde a proporção de indivíduos que apresentam um alelo particular de um gene que expressa um fenótipo associado. Alelos que apresentam alta penetrância são mais facilmente demonstrados porque devido a sua presença, o fenótipo pode ser expressado (GRIFFITHS, 2008).

[7] Epistasia (do grego, *epi*,sobre,*stasis*, inibição) representa uma interação gênica. Dá-se o nome de epistasia o efeito gerado pela associação de um gene com outro ou um grupo de genes, onde o seu efeito é modificado ou mascarado pela associação.

[8] Em um ação gênica aditiva, cada par de genes possui um efeito singular e independente dos demais que estão presentes no genótipo. Dessa forma, a ação total do genótipo sobre o fenótipo é dada pela soma dos efeitos de cada par de genes, isso implica que uma simples substituição de alelos em um gene pode interferir no resultado do efeito gênico total (VALENTE, 2001).

(ou SNP-SNP), desconsiderando ações gênicas como epistasia, além de interações gene-ambiente.

De acordo com (MITTAG et al., 2012), estudos que adotam a metodologia baseada em apenas um único *lócus* apresentam um pequeno poder preditivo em relação à doenças complexas, mesmo em casos onde SNPs indicam alta associação com o fenótipo. Em estudos de GWAS, *loci* que dão origem a fenótipos frequentemente apresentam muitos efeitos de pequenas magnitude, além de inibição ou potencialização de seus efeitos quando combinados com outros marcadores.

Para tentar contornar os problemas da metodologia monoatributo que são discutidas em (MOORE; WILLIAMS, 2009; MOORE; ANDREWS, 2014), técnicas de aprendizagem de máquina que consideram múltiplos marcadores em problemas de classificação e regressão têm sido utilizados e vêm apresentando resultados consideráveis (MITTAG et al., 2012). Uma revisão mais detalhada das metodologias baseadas em aprendizagem de máquina para problemas de GWAS pode ser encontrada no Capítulo 2 deste trabalho.

Em (MOORE, 2004) são discutidos os três principais desafios encontrados na identificação de marcadores associados à doenças em abordagens que levam em consideração a utilização do genoma inteiro. O primeiro grande desafio é em relação aos métodos de aprendizagem de máquina e mineração de dados que precisam ser desenvolvidos para capturar interações entre diversas combinações de SNPs, bem como variáveis ambientais que possam gerar influência na determinação do fenótipo. Isso se deve a limitações das abordagens estatísticas paramétricas convencionais como regressão linear, capaz de lidar somente com modelagem monoatributo ou regressão logística, que apesar de capturar interações não lineares, apresentam problemas com a quantidade de marcadores presentes no conjunto de dados e na ordem das interações (MOORE; WILLIAMS, 2009).

O segundo desafio se refere à seleção de SNPs que devem estar incluídos na análise. O estudo de interações epistáticas se torna um problema de busca com um número muito grande de marcadores p e um pequeno número de indivíduos n. Além disso, SNPs envolvidos em interações epistáticas podem apresentar uma menor frequência alélica (do inglês, *Minor Allele Frequency*-MAF[9]) pequena. Dessa forma, os conjuntos de dados frequentemente são esparsos e de dimensionalidade alta, o que demanda um outro problema, alto custo computacional. Nesse tipo de problema, embora a complexidade computacional

[9]frequência em que o segundo alelo mais comum ocorre em uma dada população.

seja linear em relação ao número de indivíduos na população, ela se torna exponencial à medida em que a ordem das interações a serem consideradas aumenta (em interações entre dois SNPs causais em um conjunto de dados com um milhão de marcadores, são necessários 5×10^{11} testes de combinações, para interações entre três SNPs causais,$4,2 \times 10^{22}$, para quatro SNPs causais, $8,3 \times 10^{27}$, se tornando o problema mais intratável a medida em que a ordem de interação aumenta). A maioria dos métodos presentes na literatura trabalha com interações de ordem dois e três. Frequentemente essa tarefa é realizada por um algoritmo de filtro ou de busca exaustiva ou estocástica (MOORE; ASSELBERGS; WILLIAMS, 2010).

O último desafio é relacionado à interpretabilidade dos resultados gerados, considerando que diversos métodos estatísticos não são interpretáveis do ponto de vista biológico. Ainda segundo (MOORE, 2004), modelos de programação genética e árvores de decisão apresentam potencial para fornecer um melhor entendimento dos mecanismos e etiologias das doenças estudadas.

Além da epistasia, outra dificuldade na busca pelo entendimento de doenças complexas se deve a herdabilidade. A herdabilidade pode ser estimada pela razão entre as variâncias do genótipo e fenótipo (herdabilidade em sentido amplo (H^2)). Esta razão mede a proporcionalidade de quanto o fator genético influencia o fenótipo (GRIFFITHS, 2008). A herdabilidade interfere diretamente na capacidade de seleção dos SNPs causais. Quanto menor a herdabilidade, menor é a explicação obtida através do genótipo, e de forma inversa maior a influência dos fatores ambientais.

Diversas condições médicas ou doenças apresentam herdabilidade baixa ($h^2 < 0.3$) como por exemplo: Câncer de bexiga (0.07-0.31) (GU; WU, 2011); Asma (0.3) (TAN et al., 2005); diabetes do tipo-2 (0.26) (POULSEN et al., 1999); câncer cervical (0.22) (CZENE; LICHTENSTEIN; HEMMINKI, 2002); leucemia (0,01) (CZEN; LICHTENSTEIN; HEMMINKI, 2002), etc. Portanto, é necessário desenvolver algoritmos capazes de identificar fatores de risco em diferentes níveis de herdabilidade.

Dessa forma, neste trabalho, tem-se como motivação as dificuldades relativas aos desafios de lidar com grandes quantidades de dados, a seleção de SNPs envolvidos em interações epistáticas em diferentes níveis de herdabilidade e finalmente o último desafio, referente a interpretação dos resultados gerados.

Como consequência direta da identificação dos marcadores envolvidos em interações

epistáticas e que podem dar origem a doenças complexas, surge à possibilidade de melhorias para etapas posteriores de estudos de GWAS. Assim, o desenvolvimento de métodos mais eficientes em GWAS podem gerar contribuições significativas na descoberta dos mecanismos genéticos de diversas doenças complexas, podendo gerar melhorias em diferentes áreas da sociedade.

1.2 Objetivo geral

Este trabalho têm como objetivo principal propor um modelo computacional para a identificação de interações epistáticas e que seja capaz de trabalhar com conjuntos de dados de GWAS compostos por milhares ou centenas de milhares de marcadores genéticos do tipo SNP que apresentam baixa herdabilidade, geralmente característica recorrente em doenças complexas.

1.2.1 Objetivos específicos

Os objetivos específicos desse trabalho são os seguintes:

- comparar o método proposto com abordagens computacionais distintas e principalmente com a abordagem referência utilizada em estudos de GWAS para determinar marcadores do tipo SNP envolvidos em interações epistáticas de segunda ordem. As comparações foram realizadas em uma série de conjuntos de dados com diferentes tamanhos, diferentes níveis de herdabilidade e MAF de 0.4 e 0.2. Além da comparação com poder de detecção de cada abordagem, as regras de associação que explicam as interações epistáticas encontradas por cada método são discutidas;

- comparar o método proposto com outras abordagens para determinar marcadores do tipo SNP envolvidos em interações epistáticas diferentes ordens;

- avaliar o método proposto em conjuntos de dados simulados que apresentam interações epistáticas sem efeito principal que foram definidos em outros estudos relevantes na literatura;

- demonstrar a eficiência do método em conjuntos de dados simulados com grandes números de marcadores e que repliquem a capacidade amostral de SNPs apresentadas por *chips* de genotipagem.

1.3 Estrutura do Texto

Devido a multidisciplinaridade envolvida no problema abordado neste trabalho, é importante a apresentação e discussão de diversos conceitos que abrangem conhecimentos específicos de diferentes áreas do conhecimento, como: biologia, genética, estatística, matemática e ciência da computação. Dessa forma, apresenta-se à seguir, a estrutura do texto adotada para os seguintes capítulos:

O Capítulo 1, Introdução, apresenta de forma breve uma introdução sobre o problema tratado neste trabalho, suas principais dificuldades e desafios, introduzindo conceitos biólogicos, motivação e objetivos a serem tratados sobre o problema em questão.

O Capítulo 2, Conceitos biológicos, apresenta e discute os principais conceitos referentes ao conteúdo genético e biológico a cerca do problema, apresentando informações necessárias para o seu entendimento, além de todo o conteúdo relacionado a GWAS, seus principais trabalhos e uma revisão bibliográfica dos métodos computacionais mais utilizados para a descoberta de associações.

O Capítulo 3, Métodos de inteligência computacional, apresenta um breve referencial teórico sobre todas as técnicas estatísticas e computacionais necessárias para o entendimento do modelo desenvolvido neste trabalho. São abordados temas como classificação, aprendizagem de máquina, comitê de classificadores e programação genética.

O Capítulo 4, Metodologia proposta, apresenta em detalhes todas as etapas do desenvolvimento do modelo computacional apresentado neste trabalho, de forma a cumprir os objetivos estabelecidos e justificar o uso de cada técnica.

O Capítulo 5, Experimentos preliminares, discute os experimentos realizados para o desenvolvimento do modelo proposto e suas justificativas. O capítulo destina-se a apresentar as principais dificuldades encontradas em cada etapa de seu desenvolvimento.

O Capítulo 6, Experimentos Computacionais, apresenta e discute os resultados obtidos à partir dos experimentos computacionais realizados. Dentre eles, uma série de experimentos com tamanhos de base de dados variadas, herdabilidade, MAF e comparação com outros algoritmos. Também foram realizados experimentos visando avaliar o modelo proposto futuramente em bases de dados reais.

O Capítulo 7, Conclusão, apresenta os comentários finais que levaram a conclusão do trabalho, observações, justificativas e trabalhos futuros.

2 Conceitos biológicos

Este capítulo têm como objetivo introduzir os conceitos biológicos que permeiam os estudos deste trabalho. Dentre eles, se concentram informações sobre genética básica e conceitos relacionados a GWAS, bem como as principais etapas desse tipo de estudo, medidas mais utilizadas para o cálculo do desequilíbrio de ligação e uma revisão da literatura sobre trabalhos correlatos.

Nas primeiras seções serão introduzidos conceitos básicos de genética humana. A finalidade é oferecer um embasamento teórico para um melhor entendimento de conceitos biológicos que envolvem estudos de GWAS que serão posteriormente apresentados neste capítulo.

2.1 Cromossomos

A informação genética de um indivíduo humano está armazenada em 23 pares de cromossomos, localizados no núcleo de todas as células somáticas[1] (com exceção de óvulos e espermatozoides). Desses, 44 são denominados autossômicos e 2 de sexuais (em mulheres, são representados por dois cromossomos X, e em homens por um cromossomo X e um cromossomo Y). A Figura 2.1 apresenta uma imagem de microscópio do cariograma[2] dos cromossomos haplóides[3] de indivíduo do sexo masculino (ALMGREN et al., 2003).

A estrutura dos cromossomos é linear e composta de ácido desoxirribonucleico (do inglês, *deoxyribonucleic acid*-DNA). Uma fita de DNA consiste de uma sequência composta de diversos nucleotídeos (bases nitrogenadas) que podem ser de quatro tipos diferentes: (A) adenina, (T) timina, (C) citosina e (G) guanina (BROOKES, 1999). Apesar da estrutura da molécula de DNA ser linear e extensa, em organismos eucariontes[4] elas são altamente dobradas e consequentemente condensadas. Para armazenar o DNA em um pequeno volume, cada molécula é enrolada diversas vezes até que seja armazenada em torno de proteínas denominadas histonas, formando uma haste que determina um cromossomo

[1] células somáticas são responsáveis pela formação de tecidos e órgãos em organismos multicelulares

[2] o cariograma representa a imagem dos cromossomos

[3] haplóide significa uma única representação de cada par

[4] Organismos eucariontes possuem células com estruturas mais complexas do que as células procarióticas, elas possuem organelas e estruturas endomembranares. Constituiem organismos eucariontes as plantas, animais e alguns tipos de fungos.

(PIERCE, 2012).

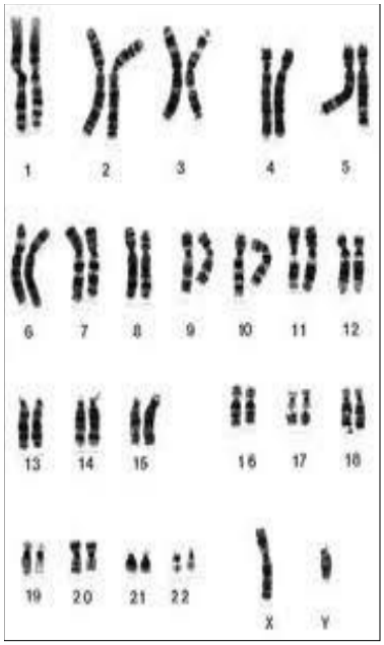


Figura 2.1: Representação do cariograma dos cromossomos presentes em indivíduos do sexo masculino. Figura extraída de (ALMGREN et al., 2003).

De acordo com (PIERCE, 2012), um cromossomo é constituído de três elementos essenciais: um centrômero, um par de telômeros e mecanismos que dão origem ao processo de replicação. Os centrômeros são filamentos que dividem o cromossomo em 2 braços e têm influencia sobre o seu movimento durante a divisão celular. Em humanos, existe um único centrômero por cromossomo. De acordo com a localização da base do centrômero, os cromossomos podem ser classificados em quatro tipos: metacêntrico (possuem o centrômero no meio, formando dois braços do mesmo tamanho), acrocêntrico (centrômero está posicionado em uma das extremidades do cromossomo), telocêntrico (apresentam o centrômero na extremidade do cromossomo, em sua região terminal, dando origem a um único braço) e submetacêntrico (assemelham-se a forma da letra J e, desta maneira, formam dois braços de tamanhos desiguais). A Figura 2.2 ilustra os quatro tipos de cromossomos existentes em organismos eucariontes.

Os telômeros são as extremidades dos cromossomos, ou seja, as pontas da estrutura linear onde o DNA encontra-se empacotado, sendo responsáveis por proteger e estabilizar as extremidades cromossômicas. O último elemento constituinte do cromossomo são os locais de origem da replicação. Esses sítios definem a região onde a síntese de DNA têm o seu inicio. Na preparação para a divisão celular, cada réplica cromossômica faz uma cópia de si mesma, sendo estas duas cópias inicialmente idênticas, chamadas cromátides irmãs

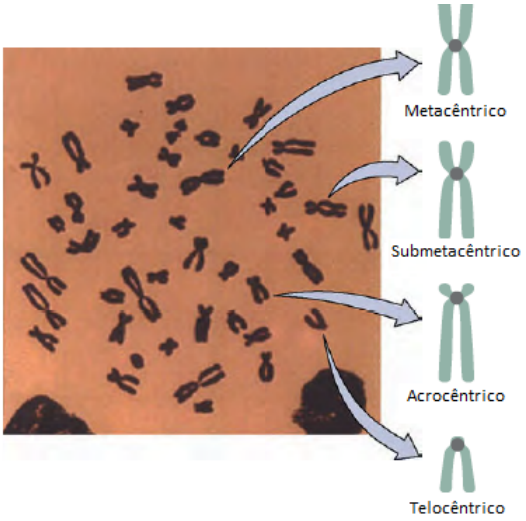


Figura 2.2: Representação da classificação dos tipos de cromossomos de acordo com a posição do centrômero. Figura extraída de (PIERCE, 2012).

e são mantidas juntas no centrômero dando origem ao cromossomo (PIERCE, 2012).

Em 1953, Watson e Crick propuseram a hipótese sobre a estrutura de dupla hélice dos cromossomos, indicando que duas fitas de DNA são anexadas juntas. Portanto, cada nucleotídeo pode formar um par complementar através de pontes de hidrogênio com somente uma outra base da outra fita de DNA, gerando os seguintes pares de base complementares: A-T, G-C, T-A e C-G. Dessa forma cada cromossomo pode ser descrito por sua sequência de DNA. Ambas as fitas carregam a mesma informação genética e o tamanho das sequências se difere em cada cromossomo. O número total de pares de base em todos os 23 pares de cromossomos no genoma é de aproximadamente 3×10^9 pares.

Como os cromossomos autossômicos são formados por pares, são denominados de pares homólogos (ou seja, similares em comprimento, pares de bases e estrutura). Porém, os pares de cromossomos autossômicos não são idênticos entre si. Como cada um dos pares é herdado pela mãe e outro pelo pai, eles diferenciam-se através da existência de polimorfismos genéticos (ALMGREN et al., 2003).

Estima-se que mais de 99% do código genético de indivíduos da mesma espécie sejam idênticos. Apenas pequenas variações em determinadas posições do DNA são responsáveis pela variação genética em uma população. Essa variações podem ser constituídas de substituições, deleções ou inserções de nucleotídeos, sendo denominadas de polimorfismos genéticos. Esses polimorfismos nas sequências de DNA determinam posições ou *loci* no genoma que podem ser utilizados como marcadores genéticos. É importante ressaltar que

as alterações provocadas por polimorfismos genéticos podem gerar alterações na sequência de uma proteína ou na sua expressão ou podem ser silenciosos e não contribuírem para alteração fenotípica (BROOKES, 1999).

2.2 Marcadores genéticos

Um marcador genético consiste de um segmento de DNA com uma localização conhecida (*lócus*) em um cromossomo em que a herança genética pode ser facilmente identificada. Marcadores genéticos devem ser determinados por métodos de análise e genotipagem. Sendo identificados, torna-se possível detectar diferenças genéticas entre indivíduos (BROWN, 2006).

Para a identificação de marcadores genéticos, pesquisadores buscam determinar marcos nos cromossomos. Foram descobertos mais de 2000 marcadores em mais de 1000 trabalhos de GWAS. A Figura 2.3 demonstra marcadores genéticos encontrados em estudos de GWAS que foram associados a diversas patologias.

As variações de sequências distintas que podem ocorrer em um determinado *lócus* são denominadas de alelos. Para exemplificar, considere a Figura 2.4, onde cada indivíduo é representado por pares de cromossomos. Pode-se observar que somente as duas últimas bases são distintas das demais. Considerando somente a coluna identificada como SNP1, nota-se que nos três indivíduos existe uma substituição de T por C. Dessa forma, defini-se dois alelos: o primeiro corresponde a uma sequência de DNA contendo uma base C, e a segunda que apresenta uma sequência de DNA contendo uma base T. Os dois alelos são representados respectivamente por A e a, onde o A representa a variante homozigoto dominante e a representa a variação homozigoto recessiva.

Como os cromossomos são apresentados em pares homólogos, em um único indivíduo um *lócus* está presente em ambos os pares. Assim, a estrutura genética é especificada por dois alelos, um em cada cromossomo do par. A combinação dos dois alelos em ambos cromossomos é conhecida como o genótipo do indivíduo nesse *lócus*. Retornando ao exemplo da Figura 2.4, pode-se observar que o indivíduo 1 apresenta uma combinação de T/C, o indivíduo 2 de T/T e o indivíduo 3 de C/C. Logo 3 genótipos são possíveis para esse *lócus*: AA, Aa e aa, onde $T/T = AA$, $C/C = aa$ e $T/C = Aa$. Deste modo os indivíduos 2 e 3 são denominados de homozigotos (dominante e recessivo respectivamente)

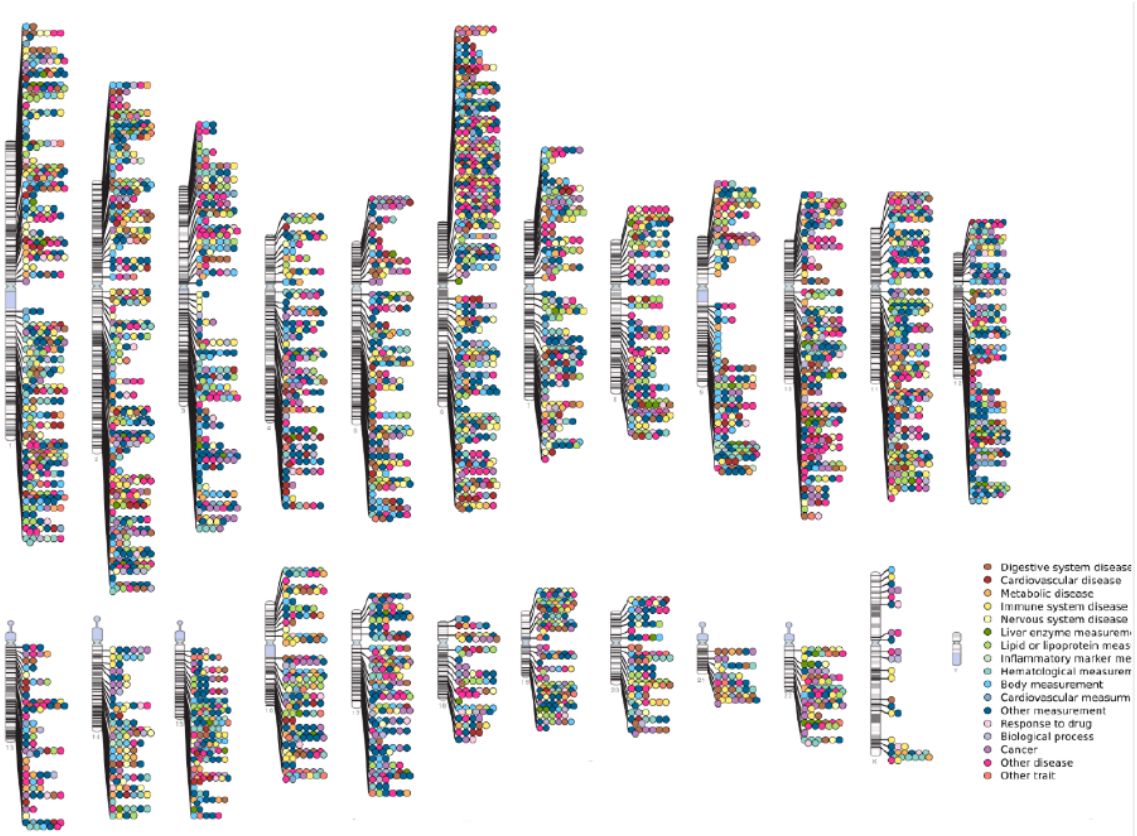


Figura 2.3: Catálogo de marcadores genéticos do tipo SNP associados a doenças por meio de estudos de GWAS. Pode-se observar os SNPs encontrados e separados por cromossomo e por cores distintas. Onde cada cor representa a uma doença de um sistema distinto. Figura extraída de e pode ser encontrada atualizada em: http://www.genome.gov/GWAStudies/

e o indivíduo 1 de heterozigoto. A mesma análise se aplica a segunda coluna denominada no exemplo de $SNP2$.

Com o avanço das tecnologias de biologia molecular, foi possível detectar polimorfismos em nível de DNA. Assim, a caracterização genética de diversos organismos se tornou além de viável, mais hábil e ágil.

É importante salientar diversos marcadores genéticos além dos SNPs, alguns dos mais utilizados são:

- Microssatélites (do inglês, *Simple Sequence Repeats Polymorphisms*-SSRP): um marcador genético utilizado principalmente em estudos de ancestralidade, são altamente polimórficos e podem ser caracterizados por nucleotídeos em sequências repetitivas curtas. Os marcadores são obtidos via amplificação das sequências próximas a essas regiões.

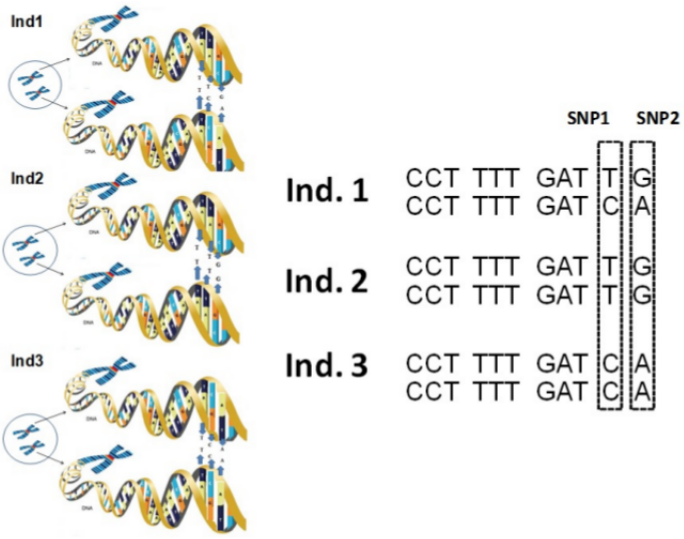


Figura 2.4: Exemplo de uma amostra de três indivíduos que tiveram o seu genótipo mapeado por SNPs. Figura adaptada de (SILVA, 2013).

- Polimorfismo no comprimento do fragmento de restrição (do inglês, *Restriction fragment lenght polymorphism*-RFLP): enzimas de restrição cortam o DNA em posições pré-definidas para a geração de fragmentos de DNA de diferentes tamanhos. Essa técnica permite a codominância de marcadores e alta reprodutibilidade. Comparada com as demais técnicas, RFLP necessita de uma grande quantidade de material genético para a obtenção dos fragmentos.

- DNA polimórfico amplificado aleatoriamente (do inglês, *Random amplified polymorphic DNA*-RAPD): é uma variação da PCR[5], se distinguindo em duas principais características: utilização de um único *primer*[6] no lugar de um par e esse *primer* apresenta uma sequência aleatória, sendo a sua sequência alvo desconhecida. É uma técnica simples e apresenta um baixo nível de reprodutibilidade.

- CAPS (do inglês, *Cleaved amplified polymorphic sequence*): os fragmentos do DNA são obtidos por uma reação de PCR. Em seguida é realizada uma digestão através de enzimas de restrição, conhecidas como PCR-RFLP. A técnica é vantajosa, especialmente em casos onde a sequência de DNA é conhecida previamente, devido a necessidade da utilização de *primers* específicos.

- AFLP (do inglês, *Amplified fragment lenght polymorphism*): É uma combinação

[5]Reação em cadeia da polimerase (do inglês, *Polymerase chain reaction*-PCR) consiste em um método de amplificação de DNA sem o uso de um organismo vivo.

[6]Sequência de nucleotídeos que atua como ponto de inicialização do processo de síntese de DNA.

das técnicas RFLP e RAPD descritas anteriormente, entretanto apresenta alguns passos adicionais. São gerados fragmentos entre 80 e 500 pares de bases através de digestão com enzimas de restrição do DNA. Por meio desta técnica, os fragmentos são conectados a nucleotídeos que são amplificados em uma reação de PCR.

- ISSR (do inglês, *Inter Simple Sequence Repeats*): são gerados a partir de marcadores microssatélites consistindo de fragmentos de DNA amplificados também por PCR. Neste caso, os microssatélites conhecidos são utilizados como *primers* de uma sequência. Através desta tecnologia é possível realizar a detecção de muitos polimorfismos.

2.3 Polimorfismos de base única

Um polimorfismo de base única, ou SNP, é uma variação em uma única posição na sequência de DNA entre indivíduos de uma mesma espécie. Em humanos, estima-se que existam aproximadamente 12 milhões destas variações, que ocorrem em média, a cada 600bp[7], podendo ser detectados em todo o código genético. Essas variações correspondem a alternância dos nucleotídeos A, T, C e G em uma frequência alélica mínima de 1% em uma determinada população (BROOKES, 1999).

SNPs também são responsáveis pela formação de alelos, onde as formas podem ser bi, tri ou tetra-alélicas. Os bi-alélicos apresentam duas formas distintas e são os mais comuns, as formas tri e tetra-alélicas são atípicas e raramente são encontradas em humanos. Por esse motivo, SNPs são frequentemente denominados de "marcadores bi-alélicos"(BROOKES, 1999). A Figura 2.5 mostra um exemplo de um SNP bi-alélico, de um tri-alélico e de um tetra-alélico respectivamente.

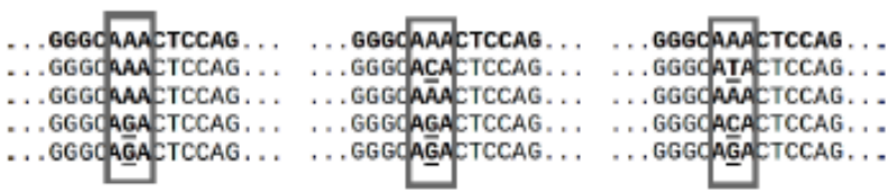


Figura 2.5: Representação de um SNP hipotético bi-alélico, tri-alélico e tetra-alélico. A primeira linha representa a sequência consenso e as bases sublinhadas dentro da janela representam os polimorfismos. Figura extraída de (ARBEX, 2009).

[7]pares de bases (do inglês, *base pair*-bp)

A frequência de um SNP é determinada em termos de menor frequência alélica (*i.e*, um SNP com menor frequência alélica de 0.40 implica que 40% da população apresenta o alelo contra o alelo mais comum ou maior alelo, que é encontrado em 60% da população) (BUSH; MOORE, 2012).

As substituições mais frequentes que ocorrem no DNA são as que envolvem bases nitrogenadas de mesma característica estrutural. Esse tipo de substituição ocorre entre duas purinas (A/G ou G/A) ou entre duas pirimidinas (C/T ou T/C), sendo denominada de transição. As trocas também podem ocorrer entre bases nitrogenadas com características estruturais diferentes entre si, ou seja, há uma troca entre uma purina e uma pirimidina ou vice-versa. Esse tipo de substituição é conhecida como transversão (GUIMARÃES; COSTA, 2002). A Figura 2.6 mostra um exemplo de transição e transversão.

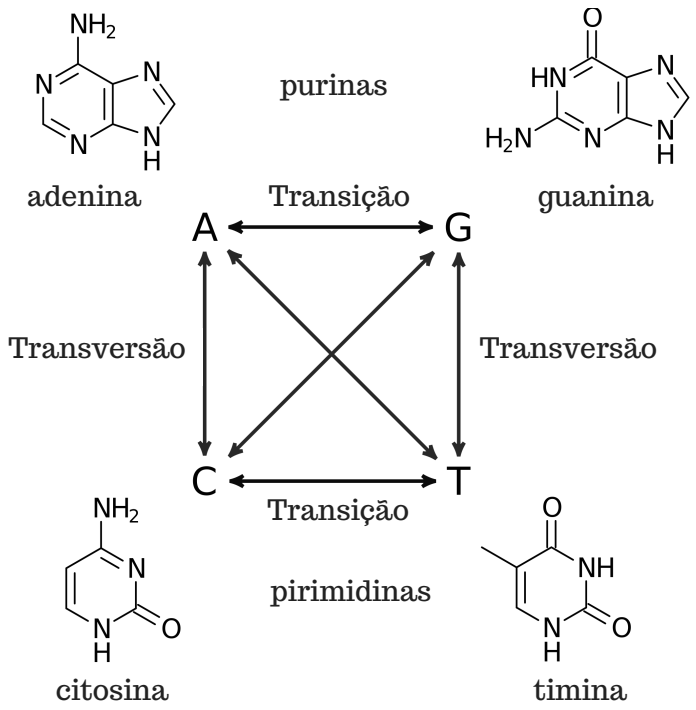


Figura 2.6: Exemplo de substituições entre bases nitrogenadas e do processo de transição e transversão. Figura baseada em (GUIMARÃES; COSTA, 2002).

A maioria dos SNPs não determinam efeito algum para os indivíduos. Entretanto, alguns desses marcadores podem ser muito importantes no estudo da genética humana, podendo ser encontrados em regiões codificadoras e alterarem o processo de formação de proteínas. SNPs também podem auxiliar na predição da resposta de indivíduos a determinados tipos de drogas, suscetibilidade a fatores ambientais e o desenvolvimento

de determinadas doenças (BROOKES, 1999). A relevância na descoberta de SNPs levou a criação de um projeto internacional para catalogá-los em larga escala. O projeto é conhecido como HapMap[8] e envolve cientistas e agências de pesquisa de países como Canadá, China, Japão, Nigéria, Reino Unido e Estados Unidos.

2.4 Genotipagem

A genotipagem é o processo de identificação de marcadores que variam entre indivíduos. Para cada marcador em um determinado indivíduo, existem dois alelos iguais ou distintos (como mencionado anteriormente, cada uma proveniente de um dos seus progenitores). O propósito desse processo é encontrar uma determinada característica de interesse. Através da comparação destas sequências, é possível determinar qual traço um indivíduo possui para cada gene, ou seja, se ele possui um genótipo homozigoto dominante ou recessivo ou se é heterozigoto para um determinado gene ou marcador.

Durante um longo período de tempo, a utilização de marcadores SNPs em estudos de associação, bem como em outros estudos como rastreabilidade genética, confirmação de paternidade, detecção de doenças sofreram com limitações tecnológicas. O cenário mudou drasticamente nos últimos anos com a geração de novas tecnologias de genotipagem em massa de SNPs (CAETANO, 2009).

Até meados da metade da última década o método referência para prospecção de SNPs era baseado no método de sequenciamento Sanger (SANGER; NICKLEN; COULSON, 1977). Nesse caso, a detecção de SNPs ao longo do genoma é dada pelo alinhamento de uma sequência de um fragmento aleatório do genoma com uma sequência consenso (CAETANO, 2009). Todavia, essa técnica enfrenta problemas devido a limitação da capacidade de geração e de análise de sequências de um laboratório, bem como pela dependência da existência de uma sequência referência. Programas como PHRED (EWING; GREEN, 1998), PHRAP (GREEN, 2009) e CONSED (GORDON; ABAJIAN; GREEN, 1998) foram desenvolvidos para otimizar as etapas de análise dos dados.

Uma outra alternativa é através do sequenciamento direto de determinados fragmentos do genoma que são amplificados por PCR. Com o desenvolvimento de novas tecnologias de genotipagem, surgiram plataformas baseadas em espectrometria de massa (HEATON

[8]disponível em http://www.hapmap.org

et al., 2005) e ensaios de microarranjos (do inglês, *microarrays*) (NEALE et al., 2008; KAMISKI et al., 2005; CHESSA et al., 2007), o que possibilitou a geração de dezenas de milhares de SNPs em um único ensaio.

Os microarranjos de DNA, ou chips de genotipagem de alta densidade, com dezenas de milhares de marcadores SNP proporcionaram avanços em estudos de identificação de genes bem como diminuíram os custos de genotipagem. Nessa técnica, um arranjo de moléculas de DNA é quimicamente ligada em uma superfície com compostos que possuem carga positiva Dentre as plataformas comerciais de genotipagem desse tipo estão: Illumina, Affymetrix, Agilent, AppliedBiosystems, entre outras. Cada uma possui diferentes *chips* com coberturas que abrangem número de marcadores distintos no mercado (CAETANO, 2009). Cada plataforma também apresenta diferentes métodos para medir a variação de SNPs. Por exemplo, os *chips* de densidade da Illumina são mais caros, porém oferecem melhor especificidade (BUSH; MOORE, 2012). A Figura 2.7 mostra o resultado dos sinais obtidos por um *chip* de alta densidade obtido pelo processo de microarranjos.

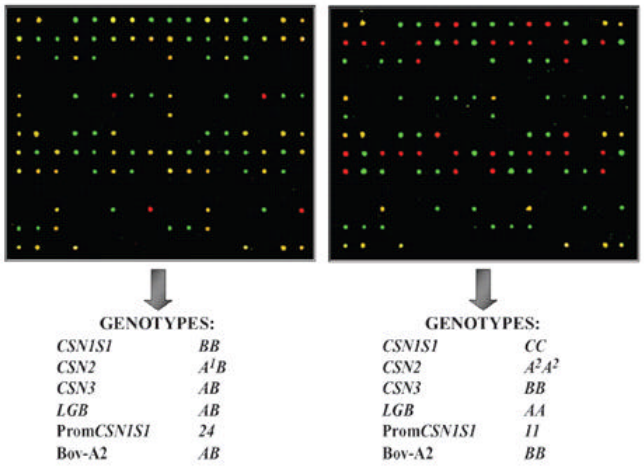


Figura 2.7: Exemplo dos sinais obtidos pelo processo de genotipagem de microarranjos. No exemplo, são demonstrados os sinais de dois indivíduos, onde são apresentadas variações alélicas nos seus genótipos. Figura extraída de (CHESSA et al., 2007).

O sequenciamento de nova geração, (do inglês, *Next-Generation Sequencing*-NGS) é considerado um sequenciamento de rendimento alto e automatizado que possibilitou uma nova abordagem em genotipagem. Inicialmente, é realizado o processo de particionamento do DNA em pequenos fragmentos e são feitas ligações de adaptadores nas extremidades dos fragmentos. Os adaptadores são utilizados como *primers* e são responsáveis pela distribuição espacial dos fragmentos de DNA em uma superfície. Os fragmentos são fixados

pela complementariedade das bases entre as bases do adaptador e da superfície. Os fragmentos de DNA são amplificados por PCR para geração de grupos de sequências idênticas. Finalmente, o sequenciamento é realizado através de diversos ciclos de incorporação, sendo que a sequência de cada *cluster* é obtida através dos *softwares* de cada plataforma em um processo conhecido como *base calling* (método responsável pela conversão dos sinais obtidos em cada ciclo em bases nitrogenadas) (PFEIFER, 2016; GOODWIN; MCPHERSON; MCCOMBIE, 2016).

Existem diversas plataformas de NGS, entre as mais conhecidas estão a Illumina e Ion Torrent. Apesar da similaridade entre elas, cada uma possui características únicas sobre a quantidade de material de DNA utilizada inicialmente no processo, diferenças na preparação da biblioteca e do método de amplificação por PCR. As plataformas de NGS permitiram o barateamento dos custos de sequenciamento e são mais comumentes utilizadas em laboratórios e centros de pesquisa, o que permite a realização de diversos projetos em larga escala (PFEIFER, 2016).

Os SNPs gerados como saída de *chips* de densidade podem ser codificados na forma de letras (*i.e*, AA, Aa, aa) ou números (*i.e*, 0, 1, 2), respectivamente. Esta codificação visa representar a variação dos alelos em uma determinada população, sendo que cada tecnologia utiliza uma abordagem de codificação singular para determinar as variantes do genótipo (NEALE et al., 2008).

2.5 Fenótipo

Fenótipo é um termo utilizado para representar todas as características observáveis de um indivíduo que resultam de uma interação do seu genótipo com o ambiente. Como características observáveis entende-se comportamentos, características morfológicas, fisiológicas e propriedades bioquímicas (por exemplo: cor dos olhos, tipo sanguíneo, doenças adquiridas) (CAETANO, 2009).

O fenótipo pode sofrer constantes modificações ao longo da vida de um indivíduo devido a mudanças no ambiente e modificações fisiológicas e morfológicas relacionadas a idade do mesmo. Dessa forma, ambientes diferentes também podem influenciar determinados traços de herança, com possível alteração das expressões de genótipos.

Os fenótipos são categorizados de duas formas distintas: discreto e contínuo. No

primeiro caso representam características qualitativas e no segundo características quantitativas. Como o fenótipo representa características que devem ser mensuradas, estas duas formas se adequam as possibilidades. Para exemplificar, considere uma doença complexa como diabetes do tipo II. O fenótipo de indivíduos que fazem parte do estudo pode ser categorizado por dois valores discretos e dicotômicos, como tendo a doença ou não, caracterizando um fenótipo discreto. Já para um fenótipo contínuo, pode-se considerar o valor do PTA como exemplo[9].

2.6 GWAS

Em GWAS, um conjunto de variáveis genéticas são observadas em um grupo de indivíduos para a identificação de quais variantes são responsáveis pela expressão do fenótipo que representa uma característica de interesse, como por exemplo uma doença, ou seja, determinar os mecanismos genéticos que dão origem à um fenótipo de interesse. Esse tipo de estudo se tornou viável com o surgimento e avanço das tecnologias de genotipagem em massa. Estudos de GWAS são divididos em duas principais categorias: estudos baseados em famílias e estudos baseados em populações. No primeiro tipo são coletados dados de indivíduos portadores do fenótipo de interesse e de seus pais. O objetivo é buscar os alelos que foram transmitidos de forma hereditária ou geracional de pai para filho com uma frequência maior do que à esperada ao acaso. O segundo tipo é focado em populações, que baseia-se na separação de indivíduos em dois grupos, os que apresentam e os que não apresentam o fenótipo avaliado. Esse tipo de abordagem também é conhecida como estudo de "caso-controle". A ideia é procurar por marcadores cujo a frequência dos alelos entre indivíduos caso e controle sejam diferentes. A identificação de marcadores e suas associações podem indicar de forma direta a causa ou uma ligação genética relacionada ao fenótipo. Desta forma, faz se necessário uma investigação detalhada em etapas ou estudos posteriores sobre regiões próximas aos marcadores identificados (BUSH; MOORE, 2012). Nas próximas seções, serão introduzidos os conceitos de GWAS e de componentes teóricos necessários ao entendimento do problema abordado neste trabalho.

[9] A capacidade prevista de transmissão, do inglês, *Predicted Transmission Ability*-PTA, é uma característica para avaliação de um touro à partir da produção de leite de sua prole.

2.6.1 Estudos de caso-controle

O objetivo dos estudos caso-controle é estudar o desenvolvimento de uma doença. Para isso, informações detalhadas da doença são desejadas nesse tipo de estudo para garantir que a doença em questão está de fato presente entre os indivíduos que estão sendo definidos no grupo de casos. Para garantir a robustez desse tipo de abordagem, fatores como uma boa correspondência entre o número de indivíduos no grupo de caso e controle devem ser levados em consideração, de forma que as diferenças genéticas entre os mesmos estejam relacionadas com a doença, não geradas a partir de uma amostragem incorreta. Os casos e controle devem pertencer a grupos étnicos similares, com a coleta das informações genéticas devendo seguir de preferência áreas geográficas correlacionadas, afim de se obter uma distribuição semelhante entre indivíduos de ambos os grupos (IOANNIDIS et al., 2001).

2.6.2 Etapas de um estudo de GWAS

Na literatura de GWAS, existem diversas metodologias e *workflows*[10] para descrever as etapas de um estudo de GWAS. Neste trabalho, como exemplo, foram consideradas as etapas propostas em (KINGSMORE et al., 2008):

1. a realização de um bom planejamento do estudo deve ser levada em consideração. A base de dados deve ser grande e a quantidade de indivíduos dos grupos caso e controle são fundamentalmente relevantes para a obtenção de resultados significativos.

2. a coleta do genótipo das amostras de indivíduos casos e controles (preferencialmente amostras maiores do que 1000, visto a importância da quantidade de dados no item anterior) deve ser feita a partir da genotipagem do DNA de cada indivíduo. Fatores em comum entre os indivíduos devem ser considerados tais como raça, etnia, sexo etc, para que a amostragem maximize os sinais de SNPs de interesse;

3. utilizar um número de aproximadamente 1 milhão de SNPs aleatórios ou 25000 não sinônimos;

[10]Fluxo de trabalho (do inglês, *workflow*) é um grafo direcionado acíclico ou cíclico em que cada tarefa é representada por um nodo e a dependência entre as tarefas é dada por um arco dirigido entre os nodos correspondentes. (BALA; CHANA, 2011)

4. controle de qualidade sobre os dados brutos gerados a partir dos *chips* de densidade. O objetivo desta etapa é a identificação, verificação e possíveis correções de erros gerados pelo processo de genotipagem;

5. derivação de blocos haplótipos [11];

6. realização de múltiplos testes estatísticos como por exemplo o χ^2 com o objetivo de verificar a associação entre os marcadores SNPs e o fenótipo;

7. SNPs com valores-p menores do que 10^{-7} devem ser considerados. Correções como a taxa de detecção de falsos SNPs ou correção de Bonferroni devem ser realizadas afim de assegurar a relevância de cada marcador em relação a doença;

8. refinamento do sinal de associação com a genotipagem adicional de SNP na região. O objetivo é identificar o desequilíbrio de ligação na região de interesse, bem como a associação entre os marcadores, derivação empírica de haplótipos e exame do efeito da estratificação, caso o mesmo esteja disponível;

9. verificação da confirmação dos sinais de associação entre marcadores e o fenótipo através da replicação dos resultados em amostras independentes de uma população (amostras maiores do que 1000 indivíduos), realizando a genotipagem de SNPs candidatos nomeados (menos do que 20 SNPs) e os seus respectivos testes χ^2 ou equivalentes;

10. finalmente, a última etapa do estudo deve ser realizada, que se trata da validação biológica da associação pela identificação dos alelos para o aumento de risco, exame da consequência funcional da variante encontrada e a determinação do mecanismo de aumento do risco.

2.6.3 Desequilíbrio de ligação

O desequilíbrio de ligação (do inglês, *linkage disequilibrium*-LD) é utilizado para a associação não-aleatória de SNPs, justamente para avaliar se um marcador SNP considerado pode não ser a causa da doença. Entretanto ele pode estar fortemente associado

[11] De acordo com (KINGSMORE et al., 2008), blocos haplótipos são uma combinação de alelos em *loci* ligados em uma única cromátide que podem ser transmitidos um número de vezes maior juntos do que aleatoriamente.

com a variante funcional, que acredita-se originar a doença. O LD ocorre quando dois ou mais alelos em *loci* distintos, em um mesmo cromossomo são mais frequentes em conjunto do que separado. Diz-se então que os *loci* estão em desequilíbrio. Caso contrário, diz-se que os alelos estão em equilíbrio de ligação (do inglês, *linkage equilibrium*-LE) (ARDLIE; KRUGLYAK; SEIELSTAD, 2002).

Quanto mais próximos estão os marcadores, maior a probabilidade de segregarem em conjunto durante o processo de recombinação gênica. A ligação pode ocorrer de duas formas: dentro de uma família e de uma população. No primeiro caso, a ligação ocorre quando dois alelos em *loci* distintos permanecem ligados em um cromossomo em vez de serem quebrados por eventos de recombinação que ocorrem durante o processo de meiose (identificados como uma linha vermelha na Figura 2.8). Em uma população, os segmentos do cromossomo que deram origem a geração inicial são sequencialmente reduzidos de tamanho através dos eventos de recombinação. Com o passar das gerações, um par de marcadores em um cromossomo movem do desequilíbrio de ligação para o equilíbrio de ligação, a medida que os eventos de recombinação ocorrem entre cada ponto viável dentro de um determinado cromossomo (BUSH; MOORE, 2012). A Figura 2.8 ilustra o desequilíbrio de ligação dentro de uma família e dentro de uma população.

2.6.4 Avaliação de desequilíbrio de ligação

Na literatura de GWAS, existem diversas métricas para realizar a avaliação de LD. As medidas mais utilizadas são D' e r^2.

Para o cálculo da medida D', considera-se dois *loci* na forma: Seja P_{AB} a frequência observada entre o par de *loci* e P_A e P_B a frequência esperada entre os alelos separados. O desequilíbrio é estimado pela Equação 2.1 (ARDLIE; KRUGLYAK; SEIELSTAD, 2002), que demostra o cálculo do escalar D.

$$D = P_{AB} - P_A \times P_B \tag{2.1}$$

Os valores relacionados com as probabilidades P_A e P_B podem ser reescalados de acordo com as Equações 2.3 (a) e 2.3 (b). Portanto, pode-se obter D'.

$$D' = \frac{|D|}{D_{max}}, \tag{2.2}$$

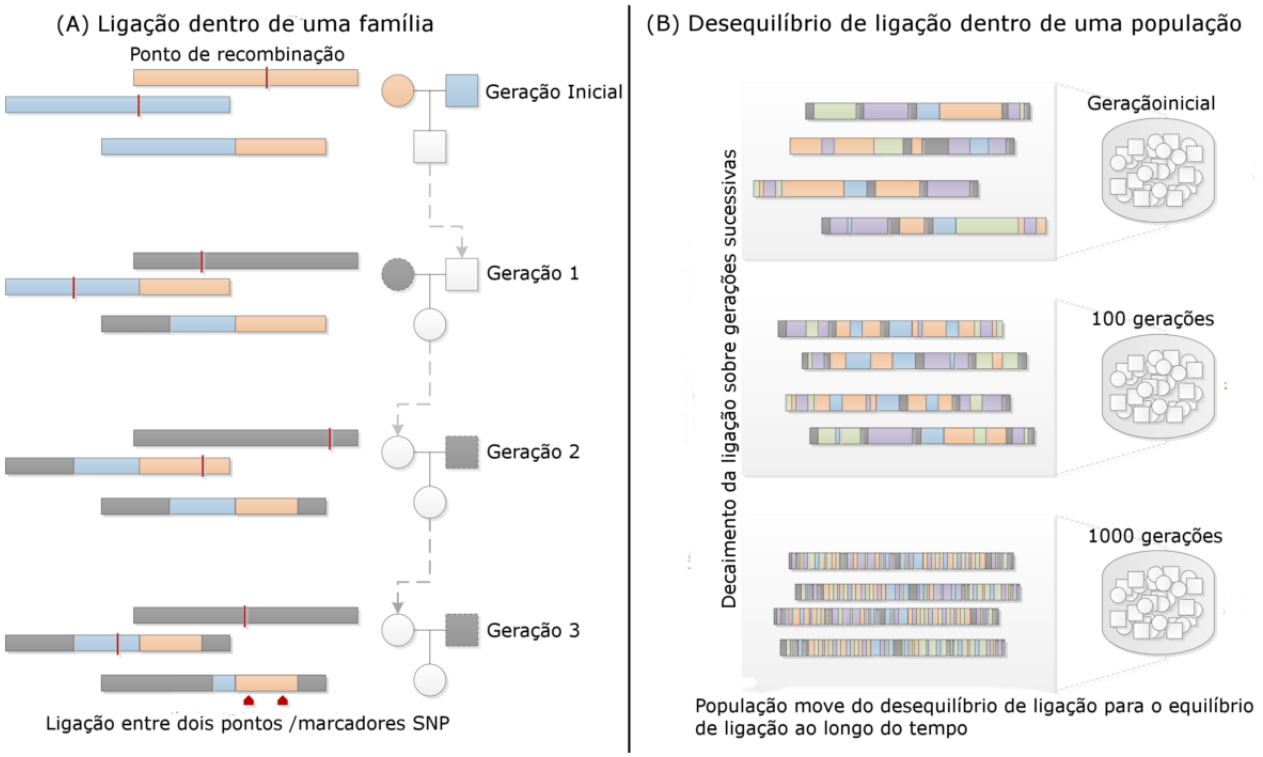


Figura 2.8: Exemplos de desequilíbrio de ligação. (A) ligação entre dois marcadores em uma família permanecendo ligados em um mesmo cromossomo, mesmo depois de eventos de recombinação, mostrados com uma linha vermelha. (B) SNPs em desequilíbrio de ligação que ao longo das gerações transformam-se em equilíbrio de ligação. Figura extraída de (OLIVEIRA, 2015) e adaptada de (BUSH; MOORE, 2012)

onde

$$D_{max} = \begin{cases} min(P_A P_b, P_a P_B) & \text{se } D > 0 (a) \\ min(P_A P_B, P_a P_b) & \text{se } D < 0 (b). \end{cases} \tag{2.3}$$

A segunda medida é denominada de r^2, é uma métrica na estatística χ^2 para o teste de associação entre uma tabela de contingência. Pode ser calculada pela Equação 2.4:

$$r^2 = \frac{\chi_1^2}{N}, \tag{2.4}$$

onde $N = 2n$ é o total de alelos B e b. O maior valor que pode se obter utilizando r^2 é determinado em função das frequências alélicas nos dois *lócus*. Quanto mais diferentes as frequências alélicas, menor o valor de r^2. Desta forma, se os variantes são raros, r^2 terá um valor baixo porque a maioria dos SNPs genotipados são comuns. Assim, um valor de r^2 baixo é adequado para a detecção de LD entre os *loci* (ARDLIE; KRUGLYAK;

SEIELSTAD, 2002).

2.6.5 Equilíbrio de Hardy-Weinberg

O equilíbrio de Hardy-Weinberg que diz que a partir de certas condições como tamanho da população, seleção, migração e mutação, as frequências alélicas permanecerão constantes ao longo das gerações (HOFFEE, 2000). É valida independentemente se um gene for raro ou frequente na população e, de acordo com a regra, sua frequência será mantida desde que essas condições sejam obedecidas.

Segundo o equilíbrio de Hardy-Weinberg, a reprodução não é a única causa da evolução. Há a necessidade de outros fatores como a pressão da seleção natural, migração, mutação para que as populações evoluam. Quando uma população está em equilíbrio de Hardy-Weinberg, as frequências genotípicas são determinadas pelas frequências alélicas. Para exemplificar os conceitos, considere um caso hipotético onde existe um único *lócus* com alelo dominante A e o alelo recessivo a. Com frequências alélicas p e q. Considere $freq(A) = p$ e $freq(a) = p$ onde $p + q = 1$. Deste modo, considerando que os alelos dos controles no conjunto de dados estão em equilíbrio de Hardy-Weinberg, temos a $freq(AA) = p^2$ para os homozigotos AA na população e a $freq(aa) = q^2$ para os homozigotos recessivos aa e $freq(Aa) = freq(aA) = 2pq$ para os heterozigotos (PIERCE, 2011). Em estudos de GWAS o genótipo dos indivíduos do grupo controle devem estar em equilíbrio de Hardy-Weinberg. Diversas fatores podem ocasionar desvios do equilíbrio na população como estratificação, seleção ou endogamia.

2.6.6 Herdabilidade

A ocorrência e desenvolvimento de uma doença em um indivíduo é determinada basicamente por dois aspectos: a carga genética e fatores ambientais. Neste último estão variáveis como estilo de vida que incluem dieta, tabagismo, exposição a agentes infecciosos, etc. Do ponto de vista da genética, a caracterização do fenótipo é dada pela Equação 2.5:

$$F = f(G + A), \tag{2.5}$$

onde F é o fenótipo, G a carga genética representada pelos genes e A é relativo ao fator ambiente. Isto indica que a determinação da expressão do fenótipo é dependente de ambos elementos (a expressão do gene e os fatores ambientais).

A herdabilidade pode ser estimada pela razão entre a variância do genótipo e a do fenótipo. O objetivo é medir a proporção da variação fenotípica que pode ser herdada em uma população. Em outras palavras, pode-se medir o quanto do fenótipo pode ser explicado somente pelo genótipo, sem levar em consideração os fatores ambientais. A herdabilidade pode ser calculada pela Equação 2.6:

$$H^2 = \frac{Var(G)}{Var(F)}, \tag{2.6}$$

onde $Var(G)$ e $Var(F)$ representam a variância do genótipo e fenótipo, respectivamente. Como exemplo: se uma doença ou fenótipo possui uma herdabilidade de 0.4, significa que 40% de toda a variação fenotípica da doença é dada somente pela carga genética (OLAZAR, 2013). Os outros 60% correspondem a fatores externos como variáveis ambientais por exemplo.

A herdabilidade pode interferir na capacidade de identificação correta dos marcadores. Dessa forma, quanto menor a herdabilidade, menor é a explicação obtida através do genótipo, sendo maior a influência dos fatores ambientais.

2.6.7 Epistasia

A epistasia é definida de duas formas, a primeira de forma qualitativa em (BATERSON, 1907) e a segunda de forma quantitativa em (CLARK; KEMPTHORNE, 1958). A primeira definição de epistasia é entendida como a inibição de um alelo em um *lócus* pela expressão de um outro alelo em um segundo *lócus* distinto. A segunda definição é estatística, e pode ser vista como qualquer desvio estatístico do efeito de uma combinação gênica aditiva de pelo menos dois *loci* em relação ao fenótipo. A segunda definição permite que a epistasia seja quantificada de formas diferentes à partir de seu significado biológico. Ainda, uma outra forma de entendimento da segunda abordagem, é a interpretação como desvio, uma relação não-linear entre os efeitos da combinação de cada *loci* envolvido na interação sobre o fenótipo (MOORE, 2014).

Epistasia é uma medida para determinar a força das interações poligênicas não-lineares.

Essas interações são não-aditivas entre marcadores, *loci*, mutações, etc. Dessa forma, epistasia determina como a combinação de genes pode interferir no fenótipo. A combinação dos genes pode inibir ou potencializar o efeito em outro gene para gerar uma nova característica. O gene que inibe é denominado de gene epistático, e o gene inibido é chamado de hipostático (AJF et al., 2000).

Suponha que existam dois genes A e B, com os fenótipos Bbaa = 8 e bbaa = 6. Para exemplificar, suponha que o alelo a seja substituído pelo alelo A produzindo as seguintes alterações no genótipo e fenótipo anterior: BbAa = 12 e bbAa = 4. Desta forma, no primeiro fenótipo pode-se observar que a substituição resultou em uma variação positiva igual a 4 . No segundo, a substituição resultou em uma supressão e a variação foi negativa igual a -2. De acordo com uma determinada combinação alélica os efeitos sobre o fenótipo podem ser relevantes.

Além do exemplo descrito anteriormente, outro exemplo clássico de epistasia é a cor da pelagem que resulta do cruzamento entre cães da raça labrador. A pelagem do cão labrador é determinada por epistasia recessiva, onde o alelo B é responsável por determinar a pigmentação da pelagem preta, bb determina o pigmento marrom, o alelo A determina a deposição de pigmentos, aa não determina a deposição, desta forma, dando origem a pelagem de coloração dourada. O esquema pode ser visto na Tabela 2.1.

	AA	Aa	aa
BB	preto	preto	dourada
Bb	preto	preto	dourada
bb	marrom	marrom	dourada

Tabela 2.1: Epistasia envolvendo a cor da pelagem em cães da raça labrador. Alelos no *lócus* A alteram o efeito sobre o fenótipo provocado pelos alelos no *lócus* B.

Os efeitos epistáticos podem ser definidos como principais e não principais, levando em consideração a influência de um ou mais SNPs sobre o fenótipo. O efeito principal é determinado pela força individual de um marcador que gera influência suficiente em alguma determinada característica do fenótipo. Logo, interações sem efeitos principais são aquelas em que os marcadores possuem influência semelhante sobre o fenótipo (WAN et al., 2009). Dessa forma, á aceito que doenças complexas são frequentemente poligênicas, ou seja, causadas pelo efeito combinado de múltiplos genes. Essas interações são epistáticas, com SNPs apresentando pouco ou nenhum efeito individual, mas podendo apresentar forte influência sobre uma doenças quando estão co-atuando.

2.6.8 Medidas de pré-processamento de dados de GWAS

Durante a etapa de pré-processamento dos dados de GWAS, algumas medidas devem ser levadas em consideração. As principais são apresentadas a seguir.

2.6.8.1 Menor frequência alélica

A menor frequência alélica ou frequência do menor alelo (do inglês, *minor allele frequency*-MAF) é uma métrica destinada a avaliar a variação de alelos em uma determinada população de estudo. Assim, alelos que apresentam pequenas variações são consequentemente menos informativos, representando pouca ou nenhuma relevância genética dentro da população de interesse. A MAF é calculada pela proporção do alelo A dado por $f(A)$ e pela proporção do alelo a, representado por $f(a)$. Assim, dentre os valores de $f(A)$ e $f(a)$, o menor é escolhido (LISTED, 2005), conforme apresentado na Equação 2.7.

$$MAF = min(f(A), f(a)) \tag{2.7}$$

onde $f(A)$ é definido pela seguinte razão:

$$f(A) = \frac{2 \times (\text{total de AA} \times \text{total de Aa})}{2 \times \text{total de indivíduos na população}}$$

e $f(a)$ é calculado por:

$$f(A) = \frac{2 \times (\text{total de aa} \times \text{total de Aa})}{2 \times \text{total de indivíduos na população}}$$

2.6.8.2 Call rate

A *call rate* (CR) é utilizada para a eliminação de SNPs que apresentam uma grande quantidade de valores perdidos (do inglês, *missing genotypes*). A medida de CR é calculada em proporção a relação ao número de observações válidas (do inglês, *nonmissing genotypes*). Frequentemente, trabalha-se com SNPs que apresentam CR maior do que 95%, ou seja, $0,95$. A Equação 2.8 apresenta o cálculo da CR (SILVA, 2013).

$$CR = 1 - \frac{mg}{ng} \tag{2.8}$$

onde mg representa o número de valores faltantes e ng o de não faltantes do SNP em

questão.

2.6.8.3 Teste para equilíbrio de Hardy-Weinberg

O teste é utilizado para verificar se as frequências genotípicas observadas estão de acordo com as esperadas conforme o equilíbrio de Hardy-Weinberg. Os problemas de seleção podem ocorrer em *lócus* que se desviam do equilíbrio caso as frequências não estejam conforme o esperado. Este desvio pode ser decorrente da influência de parâmetros como tamanho da população, mutação e migração (SILVA, 2013).

Para exemplificar, considere a Tabela 3.1. São realizados teste qui-quadrado para verificar se a frequência observada é próxima da esperada. Primeiramente, deve-se calcular o valor crítico da distribuição qui-quadrado, como pode ser visto na Equação 2.9.

$$\chi_c^2 = \sum_{i=1}^{3} \frac{(O-E)^2}{E} = \frac{(53-52,17)^2}{52,17} + \frac{(196-197,64)^2}{197,64} + \frac{(188-187,17)^2}{187,17} = 0,0304, \tag{2.9}$$

onde O e E representam as frequências observadas e esperadas respectivamente. Pelo teste qui-quadrado, consideramos que se $\chi_c^2 \geq \chi_\alpha^2$ rejeita-se a hipótese de nulidade H_0, caso contrário, H_0 é aceito de acordo com um limiar de α adotado previamente. Normalmente adota-se o valor de α de 5%.

Considerando ainda o exemplo adotado e extraído de (OLIVEIRA, 2015), o valor-p associado ao valor crítico é dado por: valor-p $= 1 - P(0,0304; df = 1) = 0,8615857$, onde o primeiro termo representa a probabilidade de se obter um valor menor ou igual a $0,0304$ a partir de uma distribuição qui-quadrado com um grau de liberdade.

2.7 Trabalhos correlatos

Para a detecção de epistasia, diversos métodos têm sido propostos na literatura. Em sua grande maioria, os métodos utilizam combinações de técnicas estatísticas e de aprendizagem de máquina. Nesta seção serão apresentados os principais métodos e suas diferentes categorias. Podem ser classificados como métodos exaustivos de busca por combinações de marcadores genéticos, filtros, combinatórios, heurísticas etc. A revisão aqui apresentada não tem como objetivo listar todos os modelos disponíveis para a detecção de epistasia,

mas introduzir o referencial teórico sobre os principais métodos desenvolvidos nos últimos anos.

2.7.1 Busca Exaustiva

As estratégias que utilizam busca exaustiva tem como objetivo testar todas as combinações de variantes genéticas. Como apresenta alto custo computacional, a maioria dos métodos que utilizam-se desta estratégia tem como foco a detecção de interações entre pares de marcadores, não sendo escalável para a detecção de interações de alta-ordem[12].

Modelos de regressão paramétricos apresentam um número fixo de parâmetros que são estimados à partir dos dados e fazem predições sobre a distribuição de probabilidade gerada por eles. Tais métodos funcionam relativamente bem quando as predições são suficientemente próximas da realidade, em contrapartida, deixam a desejar quando as predições tomadas são devidamente incorretas ou imprecisas.

Um dos métodos mais explorados é a regressão logística. Porém em problemas que apresentam alta dimensionalidade dos dados, a estimação de parâmetros se torna uma tarefa difícil e imprecisa, o que introduz erros devido ao tamanho pequeno das amostras comparadas ao tamanho real do conjunto de dados utilizado, gerando muitos falsos-positivos. A regressão logística foi utilizada em trabalhos como (CORDELL, 2009; STEEN, 2011). Outra desvantagem da regressão logística é avaliar todas as combinações possíveis entre SNPs de uma ordem definida, realizando uma busca exaustiva por essas combinações a serem testadas. Dessa forma, deve se calcular uma *odds ratio* para cada interação, o que pode elevar o custo computacional e inviabilizar a utilização do método em cenários com grandes quantidades de marcadores.

Técnicas de regressão com penalização como LASSO, BLASSO e SCAD ganharam notoriedade na detecção de interações entre SNPs. Entretanto, devido ao alto custo computacional de busca por combinações e tamanho dos conjuntos de dados, esses métodos são restritos a interações entre pares de SNPs.

[12]Interações entre grupos de mais de dois marcadores

2.7.2 Representação posição a posição e teste da razão de verossimilhanças

Desenvolvido para ser rápido, o BOOST(do inglês, *Boolean operation-based testing and screening*) realiza uma busca exaustiva de todos os potenciais pares de SNPs causais presentes em um conjunto de dados (WAN et al., 2010). Para cada par de SNPs, uma tabela de contingência é construída (sendo representada por uma matrix 3×3 que armazena a frequência de distribuição de todas as nove combinações possíveis entre os alelos dos dois marcadores). As tabelas são utilizadas no cálculo da razão de log-verossimilhança para avaliar os efeitos de cada interação. A tarefa do cálculo de todas as combinações possíveis pode gerar um alto custo computacional de processamento e armazenamento dos dados, principalmente em conjunto de dados razoavelmente grandes.

Com o objetivo de amenizar essa tarefa, os dados de GWAS são transformados em conjuntos de dados binários, onde cada linha representa um SNP e cada coluna um indivíduo. Nesta representação cada SNP é descrito por três linhas, onde cada uma descreve um alelo (*i.e*, 0, 1 ou 2) e duas colunas (maiores) representando indivíduos que pertencem aos grupos de caso e controle (Figura 2.9(A)).

Cada tabela contém um *bit* que representa um indivíduo e seu respectivo fenótipo: 1 se ele corresponde ao alelo codificado pela linha atual e 0 caso contrário (Figura 2.9(B)). Embora a matriz binária pareça maior que a representação usual de dados de GWAS, o espaço utilizado para o armazenamento de informação é menor e, como a representação dos dados é próxima da linguagem de máquina, operações *bit* à *bit* podem ser executadas de maneira rápida.

Depois desta primeira etapa de triagem, o número de marcadores é reduzido significativamente, então testes de significância como χ^2 são realizados. De acordo com (GOUDEY et al., 2013), isso pode resultar na geração de muitos falsos positivo, com pouco ou nenhum efeito de interação epistática, devido a estatística χ^2 favorecer a captura de efeitos marginais altos.

2.7.3 MDR

Um método referência em detecção de SNPs é conhecido como Redução de Dimensionalidade Multifator (do inglês, *Multifactor Dimensionality Reduction*-MDR), inicialmente

A

	Controle			Caso		
	I1	I2	I3	I4	I5	I6
SNP 1	0	**1**	**1**	2	2	0
SNP 2	0	1	2	2	0	2
SNP 3	1	1	0	1	**2**	**2**

B

		Controle			Caso		
		I1	I2	I3	I4	I5	I6
SNP 1	=0	1	0	0	0	0	1
	=1	0	**1**	**1**	0	0	0
	=2	0	0	0	1	1	0
SNP 2	=0	1	0	0	0	1	0
	=1	0	1	0	0	0	0
	=2	0	0	1	1	0	1
SNP 3	=0	0	0	1	0	0	0
	=1	1	1	0	1	0	0
	=2	0	0	0	0	**1**	**1**

Figura 2.9: (A) Representação usual de dados de GWAS, onde cada célula representa o alelo (0, 1 ou 2) de cada indivíduo. (B) Representação binária, onde cada linha de cada célula indica se o individuo possui o respectivo alelo, representado como 1 ou 0, caso contrário. Para um melhor entendimento, a marcação em cinza indica a mesma representação em cada uma das tabelas. Figura extraída de (NIEL et al., 2015).

proposto por (RITCHIE et al., 2001) para a busca de interações formadas por pares de SNPs. Em (MOORE et al., 2006),o MDR foi modificado para procurar por interações entre SNPs de ordens mais altas. É considerado um método não paramétrico que não faz nenhuma suposição a respeito da distribuição de probabilidade dos dados.

Basicamente, o MDR particiona o conjunto de dados para realizar um processo de validação cruzada (do inglês,*Cross-validation*-CV) de 10 partes. Depois de definida a ordem da interação a ser buscada, o algoritmo lista todas as combinações possíveis. Para cada combinação, uma tabela de contingência é construída a partir de cada combinação alélica da interação, sendo realizada a contagem de casos e controles de acordo com a seguinte razão c:

$$c = \frac{c_1}{c_2} \tag{2.10}$$

onde c_1 representa o número de casos que dividem a mesma combinação de genótipos e c_2 o número de controles que dividem a mesma combinação. Cada combinação é definida como sendo de alto ou baixo risco, de acordo com a Equação 2.10 (*i.e*, se o resultado ficar abaixo ou acima de 1.0, sendo esse um valor de limite). O modelo de classificação é

gerado acoplando as células que possuem alto-risco em um grupo e baixo risco em outro, reduzindo assim a dimensionalidade do problema.

Posteriormente, a partir da redução de dimensionalidade, o processo CV é executado. Um classificador estima o erro sobre cada iteração da validação cruzada para realizar a predição. Os modelos resultantes que apresentam menor erro de predição são armazenados. A Figura 2.10 mostra os passos executados pelo MDR.

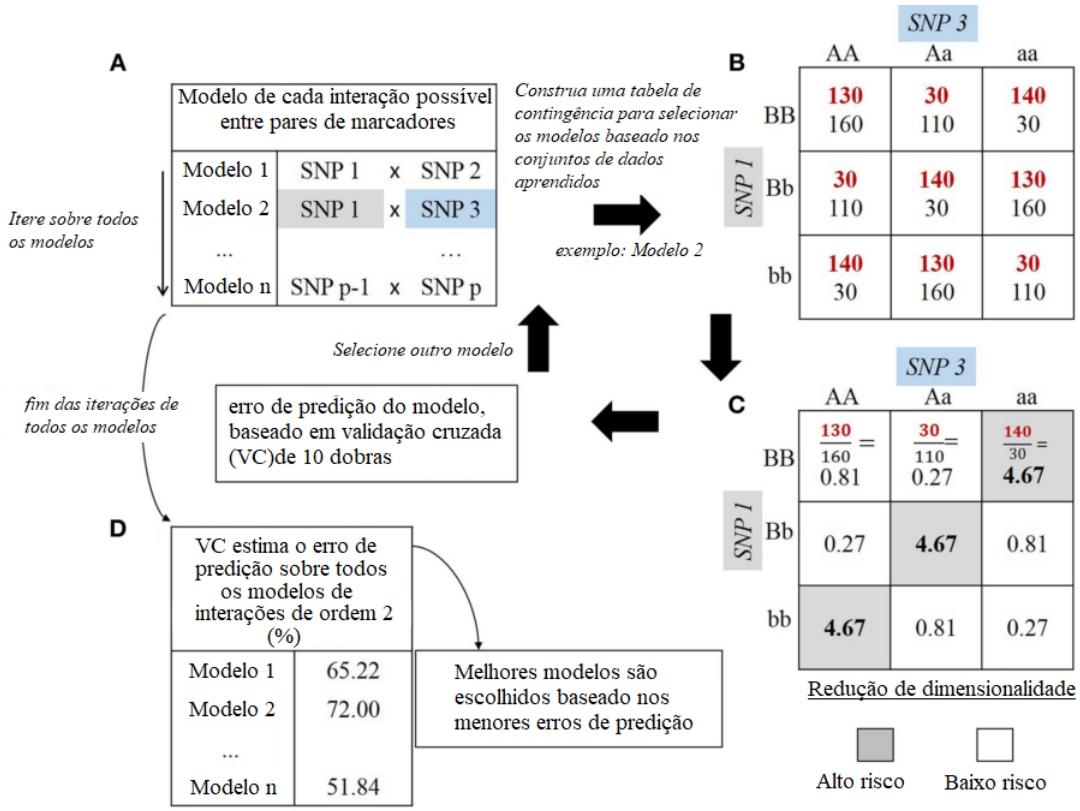


Figura 2.10: Etapas do modelo MDR: (A) Listagem de todas as combinações de segunda ordem possíveis de SNPs de um conjunto de dados. (B) Montagem das tabelas de contingência para cada combinação SNP-SNP dada pela contagem de casos (em vermelho) e controles (em preto). (C) Taxa de proporção de casos e controles é dada em cada célula. (D) Erro de predição do modelo (estimado por VC de 10 partes). Os genótipos com o menor erro de predição são armazenados e selecionados como melhores modelos.

O método vêm sendo desenvolvido desde a última década e pode ser facilmente combinado com outros modelos de classificação e filtros (MOORE, 2014). Um dos exemplos recentes foi proposto por (OLSON; MOORE, 2016). Um dos problemas do MDR se dá devido a utilização de busca exaustiva pra tentar encontrar as interações, tornando-o impraticável para busca de interações de alta-ordem e grandes conjuntos de dados.

2.7.4 Filtros

Os métodos apresentados anteriormente consideram todo o espaço de busca para tentar encontrar possíveis interações entre SNPs. Eles podem ser combinados com métodos de filtros para selecionar pequenos grupos de marcadores com a finalidade de reduzir o espaço de busca original. Um dos algoritmos mais conhecidos é o ReliefF (KONONENKO, 1994; ROBNIK-ŠIKONJA; KONONENKO, 2003). A ideia base por trás do ReliefF é aprender características informativas do conjunto de dados sem qualquer conhecimento *à priori*. Para isto, utiliza uma medida de cálculo de proximidade entre indivíduos para identificar similaridades genéticas. Posteriormente, é realizada uma avaliação da qualidade das variantes genéticas de acordo com o quanto elas podem ser úteis para distinguir indivíduos que estão próximos um dos outros. Existem diversas variações do algoritmo original. Dentre as mais utilizadas estão: RReliefF, desenvolvida para estudos de traços quantitativos, como epistasia de eQTL; *Tuned* ReliefF (TuRF), proposto para eliminar SNPs com nenhuma ou baixa importância em possíveis interações epistáticas (SNPs que não servem para discriminar indivíduos com diferentes genótipos); ReliefF espacialmente uniforme (do inglês, *Spatially Uniform ReliefF*- SURF), utiliza todos os vizinhos em uma determinada distância no lugar de usar um número fixo para comparar sua similaridade; SURF(∗), é uma variação do SURF que utiliza indivíduos vizinhos mais distantes em sua medida de avaliação. O Algoritmo 1, extraído de (MOORE; WHITE, 2007) apresenta como a importância de cada SNP pode ser calculada utilizando o filtro ReliefF.

Algoritmo 1: ReliefF

Entrada: Conjunto de dados de GWAS

n : número de indivíduos mais próximos para considerar (de acordo com os dados do genótipo)

t : número de iterações

Saída: Importância de todos os SNPs do conjunto de dados de GWAS

```
1  início
2      Selecione um conjunto de dados de GWAS
3      para cada i até t faça
4          aleatoriamente selecione um indivíduo I
5          encontre n vizinhos mais próximos S
6          encontre n vizinhos mais próximos O
7          para cada SNP ∈ conjunto de dados de GWAS faça
8              diminua a importância do SNP cada vez que o SNP no genótipo difere
                entre I e os vizinhos S
9              aumente a importância do SNP cada vez que o SNP no genótipo difere
                entre I e os vizinhos O
10         fim
11     fim
12 fim
```

Esses métodos têm a tendência de capturar efeitos marginais, sendo que dificilmente conseguem capturar interações de efeito combinatório (onde nenhum dos marcadores envolvidos na interação apresenta um valor significativo para o fenótipo sozinho).

2.8 Métodos de busca não-exaustiva

Essa categoria de métodos combina técnicas de aprendizagem de máquina com otimização combinatória e frequentemente envolvem o uso de meta-heurísticas, especialmente para detecção de interações de alta ordem. Técnicas de otimização combinatória consideram o espaço de soluções (combinações de potenciais interações entre SNPs) determinado para identificar as combinações mais relevantes.

A maioria dos métodos que fazem parte dessa categoria realizam testes para associação de variantes que permitem possíveis interações, em oposição à testar as interações em si.

Um modelo de associação que permite interações comumente considera SNPs com efeitos marginais, o que pode levar à descoberta de múltiplos marcadores com efeito independente em relação ao fenótipo, no lugar de encontrar interações epistáticas (NIEL et al., 2015). A seguir os principais métodos serão apresentados.

2.8.1 Redes bayesianas

As redes bayesianas são compostas de duas partes principais: uma probabilística e um grafo acíclico dirigido (do inglês, *directed acyclic graph*-DAG). As variáveis são representadas por nós e as dependências entre elas por arestas. A parte probabilística do modelo está associada à distribuição de probabilidade de cada nó do DAG. A rede é baseada na propriedade de Markov, onde cada variável é independente das caraterísticas do seus nodos não descendentes, dado por seus nós pais no DAG.

No contexto de dados genômicos, as variáveis são SNPs e valores de fenótipo. A ideia é oferecer uma forma intuitiva de capturar as relações existentes entre os marcadores e o fenótipo. Devido à quantidade de combinações possíveis, técnicas específicas devem ser utilizadas para reduzir o espaço de busca para tornar o modelo computacionalmente viável para avaliar grandes conjuntos de dados.

O principal modelo que utiliza redes bayesianas é conhecido como mapeamento de associação de epistasia bayesiana (do inglês, *Baysian epistasis association mapping*-BEAM) (ZHANG; LIU, 2007). O BEAM utiliza um algoritmo de cadeia de Markov Monte Carlo (MCMC) para testar cada marcador condicional com o estado natural dos outros marcadores. Para cada marcador, o algoritmo gera uma saída que corresponde com a probabilidade posterior de associação com a doença. Os marcadores são distribuídos em três grupos diferentes: marcadores que não contribuem para o fenótipo; marcadores que contribuem para o fenótipo de forma independente (efeitos marginais ou modelo aditivo) e marcadores que podem influenciar o fenótipo de acordo com suas interações (modelo epistático). Depois da fase de separação de grupos, o algoritmo utiliza testes estatísticos para filtrar os SNPs nos grupos selecionados e identificar as interações.

Outro modelo que utiliza redes bayesianas é o EpiBN, proposto em (HAN et al., 2012), que utiliza um processo iterativo chamado *Branch and Bound*-BB no lugar do MCMC. Neste modelo, a cada iteração o algoritmo adiciona, deleta ou inverte arestas. Na sequência, uma função calcula a melhor estrutura da rede dada pela iteração anterior.

Outro método alternativo é o *Markov blanket-based*. O princípio desse método é encontrar um conjunto mínimo de variáveis que guardam o estado da doença de outras variáveis, resultando em uma fração da rede bayesiana que tem o nodo do fenótipo noas bordas do grafo. Esse conjunto é denominado de *blanket*.

Estratégias baseadas em *Markov blanket* dependem fortemente da hipótese de fidelidade, definida com relação a amostra como segue: 'Toda independencia condicional na rede bayesiana também existe na distribuição de probabilidade das variáveis'. Na prática, essa é uma hipótese que raramente ocorre em estudos de GWAS.

2.8.2 Otimização por colônia de formigas

A visão geral de algoritmos de otimização por colônia de formiga (do inglês, *ant colony optimization*-ACO) é baseada em observações do mundo real. A ideia é que formigas procuram por caminhos aleatórios até encontrar comida. Ao retornar para a colônia, elas deixam um rastro de feromônio. A quantidade de feromônio serve para comunicar as outras formigas o caminho descoberto até a comida.

Baseado nesta premissa, o algoritmo de ACO foi proposto por (DORIGO; GAMBARDELLA, 1997). O método mais utilizado para detecção de epistasia que utiliza ACO é conhecido como AntEpiSeeker (WANG et al., 2010). O método realiza uma busca por múltiplos grupos de marcadores associados à doença de forma paralela. A cada iteração, o algoritmo produz formigas artificiais que cooperam entre si para atualizar o quanto os SNPs estão relacionados com o fenótipo. As formigas representam conjuntos de SNPs com potenciais efeitos de epistasia e a concentração do feromônio avalia a significância da interação epistática de cada grupo.

A comunicação entre as formigas é dada por uma função de distribuição de probabilidade (do inglês, *probability distribution function*-PDF) compartilhada pela colônia. A PDF descreve a probabilidade de seleção de um SNPs específico em uma determinada iteração. A cada iteração, múltiplos SNPs são selecionados de acordo com a PDF e testes χ^2 são utilizados para medir a associação entre cada formiga e o fenótipo. Os resultados do valor-p, obtido de cada associação, atualiza a PDF para a execução da próxima iteração do modelo (WANG et al., 2010). A Figura 2.11 indica as etapas do modelo.

A limitação do modelo deve-se a necessidade de determinação dos diversos parâmetros do modelo, que devem ser definidos *à priori*, e afetam sensivelmente a análise de GWAS

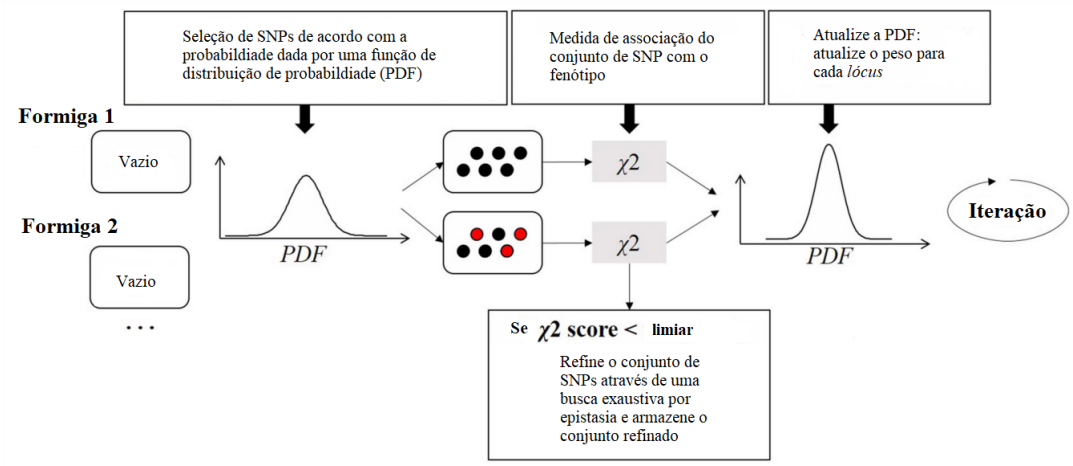


Figura 2.11: Modelo do AntEpiSeeker: grupos de SNPs são alocados em cada formiga. A PDF oferece a probabilidade de cada SNP ser alocado em uma determinada iteração. Uma vez que uma formiga define um conjunto de SNPs, são realizados testes de associação χ^2 entre esse conjunto e o fenótipo. Para cada formiga, a PDF é atualizada de acordo com o valor-p de cada associação χ^2 para a próxima iteração do modelo. Figura extraída de (NIEL et al., 2015).

e os resultados gerados.

2.9 Métodos baseados em programação genética

A programação genética (KOZA, 1992) (do inglês, *genetic programming*-GP) em estudos de GWAS têm sido utilizada com duas finalidades principais: seleção de atributos e construção de modelos discriminativos. Para o primeiro caso, estudos têm demonstrado que a GP consegue selecionar SNPs de interesse dentro de um conjunto de com centenas até poucos milhares de marcadores, com o uso de conhecimento especialista[13] aumentando significativamente a performance do algoritmo de GP. O segundo grupo têm a finalidade de descriminar e predizer amostras que são mais susceptíveis à uma doença ou fenótipo (SZE-TO et al., 2013).

Três dos principais modelos que pertecem a segunda categoria são o GPAS (do inglês, *Genetic Programming for Association Studies*) (NUNKESSER et al., 2007), o GPDTI (do inglês, *Genetic Programming Decision Tree Induction*) (ESTRADA-GIL et al., 2007) e o GP-Pi (do inglês, *Genetic Programming with penalization and initialization*) (SZE-TO et al., 2013). Os três modelos adotam soluções baseadas em árvores de decisão, que têm como

[13]neste contexto, utiliza-se o termo "conhecimento especialista" ou "inicialização " como métodos que selecionam grupos de marcadores previamente e os inserem na população inicial do algoritmo de programação genética.

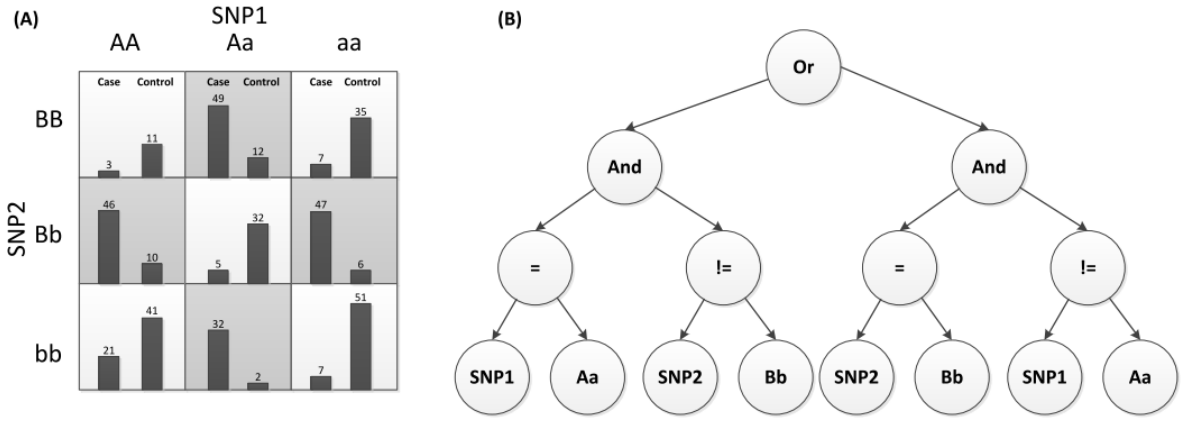


Figura 2.12: Representação de um indivíduo do modelo GP-Pi: Cada célula nas barras à esquerda (A) representam o número de casos e controles para cada combinação alélica da tabela de contingência. As células mais escuras representam as combinações com alto risco e as mais claras com baixo risco de apresentaram uma o fenótipo de interesse. Em (B) é dada a representação lógica preditiva das células que apresentam um alto risco. Dessa forma, o algoritmo evolui a cada iteração da população buscando soluções que podem explicar e modelar melhor as interações entre SNPs. Figura extraída de (SZE-TO et al., 2013).

objetivo oferecer uma interpretação mais legível e compreensiva dos resultados obtidos, além de permitir que as estruturas sejam avaliadas de forma recursiva para produzir resultados de expressões multivariadas . A principal diferença entre eles é que o GPAS e GPDTI não utilizam métodos de conhecimento especialista ou inicialização, sendo que o GPDTI suporta apenas pequenos grupos de marcadores. O GP-Pi por sua vez, apresenta um método de conhecimento especialista baseado no algoritmo ReliefF e um termo de penalização em sua função objetivo, que tem como objetivo manter a população da GP heterogênea. A Figura 2.12 exemplifica uma indivíduo da população do GP-Pi e ilustra como é representada uma interação SNP-SNP.

Os algoritmos de GP apresentam uma dificuldade em trabalhar com grandes quantidades de marcadores e volume de dados, possuindo limitações de escalonamento, dificultando análises em dados reais.

2.10 Filtros

Os métodos apresentados anteriormente consideram todo o espaço de busca para tentar encontrar possíveis interações entre SNPs. Eles podem ser combinados com métodos de filtros para selecionar pequenos grupos de marcadores com a finalidade de reduzir o espaço

de busca original. Um dos algoritmos mais conhecidos é o ReliefF (KONONENKO, 1994; ROBNIK-ŠIKONJA; KONONENKO, 2003). A ideia base por trás do ReliefF é aprender características informativas do conjunto de dados sem qualquer conhecimento *à priori*. Para isto, utiliza uma medida de cálculo de proximidade entre indivíduos para identificar similaridades genéticas. Posteriormente, é realizada uma avaliação da qualidade das variantes genéticas de acordo com o quanto elas podem ser úteis para distinguir indivíduos que estão próximos um dos outros. Existem diversas variações do algoritmo original. Dentre as mais utilizadas estão: RReliefF, desenvolvida para estudos de traços quantitativos, como epistasia de eQTL; *Tuned* ReliefF (TuRF), proposto para eliminar SNPs com nenhuma ou baixa importância em possíveis interações epistáticas (SNPs que não servem para discriminar indivíduos com diferentes genótipos); ReliefF espacialmente uniforme (do inglês, *Spatially Uniform ReliefF*- SURF), utiliza todos os vizinhos em uma determinada distância no lugar de usar um número fixo para comparar sua similaridade; SURF(∗), é uma variação do SURF que utiliza indivíduos vizinhos mais distantes em sua medida de avaliação. O Algoritmo 1, extraído de (MOORE; WHITE, 2007) apresenta como a importância de cada SNP pode ser calculada utilizando o filtro ReliefF.

Algoritmo 2: ReliefF

Entrada: Conjunto de dados de GWAS

n : número de indivíduos mais próximos para considerar (de acordo com os dados do genótipo)

t : número de iterações

Saída: Importância de todos os SNPs do conjunto de dados de GWAS

```
1  início
2      Selecione um conjunto de dados de GWAS
3      para cada i até t faça
4          aleatoriamente selecione um indivíduo I
5          encontre n vizinhos mais próximos S
6          encontre n vizinhos mais próximos O
7          para cada SNP ∈ conjunto de dados de GWAS faça
8              diminua a importância do SNP cada vez que o SNP no genótipo difere
                entre I e os vizinhos S
9              aumente a importância do SNP cada vez que o SNP no genótipo difere
                entre I e os vizinhos O
10         fim
11     fim
12 fim
```

Esses métodos têm a tendência de capturar efeitos marginais, sendo que dificilmente conseguem capturar interações de efeito combinatório (onde nenhum dos marcadores envolvidos na interação apresenta um valor significativo para o fenótipo sozinho).

3 Métodos de inteligência computacional

O campo de estudo da área de inteligência computacional é extenso e engloba uma grande diversidade de técnicas com componentes teóricos relacionados às disciplinas da área de matemática, ciência da computação e estatística. Este capítulo têm como finalidade apresentar os conceitos e métodos de inteligência computacional utilizadas no desenvolvimento do modelo proposto neste trabalho.

3.1 Árvores de decisão

Árvores de decisão são modelos estatísticos de inferência indutiva que podem ser utilizadas em tarefas de classificação e regressão. É um algoritmo muito difundido e largamente utilizado em diversas áreas (ROKACH; MAIMON, 2008). No contexto de aprendizagem supervisionada, o método utiliza dados de treinamento para construção do modelo e tem como objetivo obter uma hipótese de predição e/ou explicação dividindo o conjunto de dados por meio de regras baseadas nos atributos, utilizando estratégias de dividir para conquistar. As soluções baseadas nas regras geradas são agrupadas em forma de árvore gerando uma solução para o problema inicial (FACELI et al., 2011).

Uma árvore de decisão classifica as instâncias de acordo com as regras definidas nos nós da árvore. Cada nó especifica um teste de algum atributo da instância. Cada galho corresponde a um dos possíveis valores do atributo. A cada iteração do método, um novo nó é gerado (MITCHELL, 1997). Uma instância é classificada a partir do nó raiz, testando o atributo de acordo coma regra especificada pelo nó, em um processo hierárquico até que se encontre um nó folha ou de finalização que determina o rótulo da classe. A Figura 3.1 demonstra um exemplo de uma árvore de decisão hipotética em GWAS que classifica se uma instância possui ou não o fenótipo de interesse.

Para a construção das árvores de decisão, os nós inicial ou raiz representam o atributo mais importante (*i.e*, discriminativo em relação a alguma medida adotada) da instância para previsão do rótulo da classe. Os demais atributos devem ser inseridos sucessivamente,

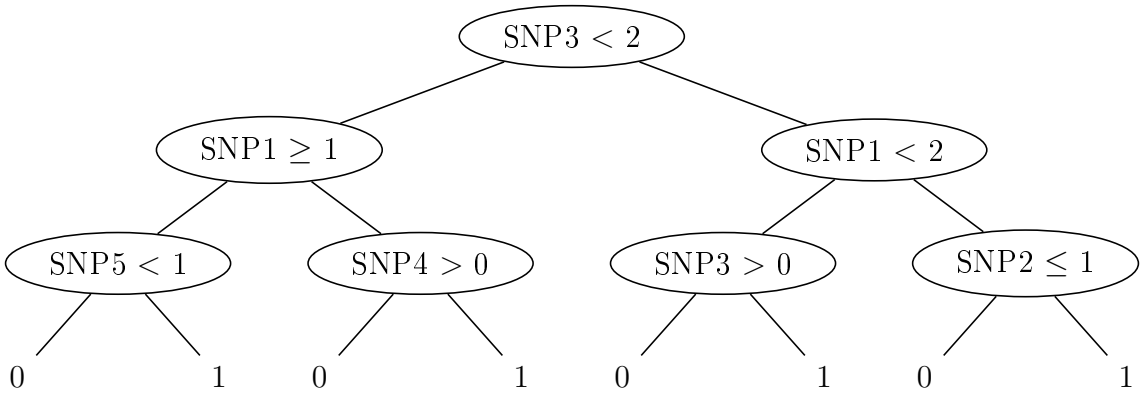


Figura 3.1: Exemplo de uma árvore de decisão. A árvore classifica se uma instância apresenta ou não um fenótipo. A instância é classificada ordenando-a através da raiz da árvore até o nó folha (neste caso, sim ou não). Por exemplo: Uma instância seria classificada como positiva se o SNP3 < 2, SNP1 ≥ 1 e SNP4 > 0.

em ordem decrescente de importância. O processo termina quando algum critério de parada é atendido.

Em (FOULKES, 2009), é apresentada uma descrição em alto nível do processo construtivo de uma árvore. Seja $D = \{(\boldsymbol{x}_i, y_i)\}$ um conjunto de dados com n amostras e m características, onde $\boldsymbol{x}_i$ representa uma instância, $x_{i,j}$ uma característica de uma instância e y_i a variável explicativa ou classe. Seja S todas as instâncias do conjunto de treinamento.

Suponha que todas as variáveis explicativas sejam binárias (0 ou 1), e que $x_{i,j}$, onde $1 \leq j \leq m$ seja a variável preditora mais importante em relação a y_i. O primeiro passo é dividir os indivíduos da amostra de treinamento S em dois grupos denominados S_1 e S_2 baseada em uma medida de informação como, por exemplo, entropia, que separa os dois grupos dado pela variável preditora. Os grupos podem ser definidos como $S_1 = \{j : x_{i,j=0}\}$ e $S_2 = \{j : x_{i,j=1}\}$ para os indivíduos dados por $i = 1, 2, ..., n$. O passo seguinte é identificar o próximo atributo com o maior potencial de predição no conjunto S_1 e o mesmo para S_2. Essa variável é dada por $x_{i,j_{a1}}$, $x_{i,j_{a2}}$, $x_{i,j_{an}}$ onde $a_1, ..., a_n$ representam os atributos potenciais para as próximas divisões. O processo é executado sucessivamente até que o critério de parada estabelecido seja alcançado.

O problema de encontrar a árvore com o menor número de nós é um problema NP-completo. Dessa forma, são utilizadas meta-heurísticas para o tornar o tempo de busca linear em função do número de atributos. As medidas mais utilizadas são: entropia, o índice gini e o erro de classificação (ROKACH; MAIMON, 2008). As medidas são dadas

pelas Equações 3.1, 3.2 e 3.3 respectivamente. O símbolo $p(i|t)$ indica a fração de registros pertencentes à classe i de um determinado nó t.

$$Entropia(t) = \left[-\sum_{i=1}^{c} p(i|t) log_2 p(i|t) \right] \tag{3.1}$$

$$Gini(t) = \left[1 - \sum_{i=1}^{c} [p(i|t)]^2 \right] \tag{3.2}$$

$$Erro(t) = [1 - max_i [p(i|t)]] \tag{3.3}$$

Em (FACELI et al., 2011) são abordadas as principais vantagens das árvores de decisão:

- Flexibilidade: O método é não paramétrico, ou seja, o modelo de árvore não possui a premissa quanto à distribuição dos dados;
- Robustez: Pequena sensibilidade a distribuições com "caudas longas" e *outliers*. As árvores são invariantes as transformações estritamente monótonas nas variáveis da base de dados inicial;
- Seleção de atributos: A escolha das variáveis é realizada durante o processo de construção da árvore. Dessa forma, a árvore gerada é robusta na presença de variáveis redundantes e irrelevante na base de dados inicial;
- Interpretabilidade: A estrutura da árvore pode ser traduzida naturalmente na forma de regras lógicas de conjunções e disjunções. A relação entre as variáveis explicativas e a variável explicada pode ser identificada mesmo em casos onde ela é complexa;
- Eficiência: O tempo computacional gasto na construção das árvores é proporcional aos algoritmos lineares em relação ao tamanho da bases de dados de treinamento.

Em (FACELI et al., 2011) também são discutidas as principais desvantagens da utilização dos modelos de árvores de decisão, listadas a seguir:

- Replicação: Uma subárvore pode ser replicada n vezes em níveis distintos em uma árvore de decisão, tornando-a mais complexa do que o necessário e reduz a interpretabilidade do modelo;

- Ausência de valores: A árvore toma uma decisão baseada nos valores de uma determinada variável escolhida para realizar a dicotomia. Valores faltantes podem causar problemas na tomada de decisão sobre qual ramo seguir;

- Atributos contínuos: Neste caso, há a necessidade da ordenação para cada nó de decisão devido a existência de variáveis explicativas contínuas. Estima-se que essa tarefa é responsável por até 70% do tempo necessário para ajustar uma árvore de decisão em grandes conjuntos de dados;

- Instabilidade: Variações pequenas no conjunto de treinamento podem gerar grandes variações na geração da árvore final. Tal fato ocorre se para a escolha da variável usada na quebra há similaridade entre duas ou mais variáveis.

3.2 Comitê de classificadores

A combinação de vários classificadores base (classificadores individuais) para melhorar a predição é denominada de comitê de classificadores (do inglês, *ensembles*). O método têm sido utilizado em problemas onde um único classificador ou especialista não generaliza o problema de forma eficiente. Com a aplicação de comitê de classificadores espera-se um refinamento na capacidade preditiva do modelo de forma que bons resultados possam ser encontrados em aplicações das mais diversas áreas e cenários.

De acordo com (TAN; STEINBACH; KUMAR, 2005), um método de comitê de classificadores deve utilizar o conjunto de treinamento dos dados de forma que seja possível construir um conjunto de classificadores base e viabilizando a classificação através dos votos de cada classificador. Dessa forma, cada voto representa a previsão feita por cada um dos classificadores.

Em (POLIKAR, 2006) são discutidos as principais vantagens para a utilização de métodos de comitê:

- Razões estatísticas: Experimentalmente têm sido demonstrado que comitê de classificadores apresentam melhor desempenho do que modelos selecionados utilizando validação cruzada, *bootstrap* e *leave-one-out*. Comitês de classificadores reduzem a variância do modelo (erro particular de um modelo treinado), sendo que, quanto

maior o número de classificadores compondo o comitê, maior a redução de sua variância;

- Grandes volumes de dados: Em inúmeros cenários, a quantidade de dados a serem analisados pode ser muito expressiva para que um único classificador realize a tarefa, visto que treinar um classificador com uma grande quantidade de dados não é prático nem trivial. Utilizando comitês de classificadores específicos pode-se quebrar o conjunto de treinamento em pequenos subconjuntos onde cada classificador é responsável por treinar somente uma parcela dos mesmos. O resultado de todos os classificadores podem ser combinados para tornar a tarefa de classificação mais eficiente;

- Conjunto de dados muito pequenos: comitês podem ser utilizados em contraposição do modelo discutido anteriormente. Conjuntos de dados muito pequenos apresentam baixa representatividade do espaço amostral. Técnicas de reamostragem e imputação podem ser utilizadas para gerar subconjuntos de treinamento aleatórios onde cada um pode ser treinado por cada classificador criando um comitê;

- Divisão e conquista: Independente da quantidade de dados disponível, alguns problemas são muito difíceis de serem resolvidos por um único classificador. A fronteira de decisão que separa as classes pode ser muito complexa ou pode estar fora do espaço de funções que podem ser mapeadas pelo classificador. A Figura 3.2 e Figura 3.3 ilustram um cenário onde um único classificador não consegue isoladamente só aprender a regra de classificação. A ideia é dividir o espaço dos dados para que os classificadores presentes no comitê consigam realizar a tarefa de classificação. O espaço de dados é particionado em subgrupos menores que devem ser mais fáceis de serem aprendidos pelos classificadores. Finalmente, a fronteira de decisão pode ser aproximada através da combinação dos resultados de diferentes classificadores.

Para exemplificar, considere um comitê com 25 classificadores binários. Cada classificador apresenta uma taxa de erro de $\epsilon = 0.35$. A classificação final dada pelo classificador de grupo é igual a maioria dos votos sobre as previsões realizadas pelos classificadores base. Dessa forma, a taxa de erro permanece em 0.35. Caso os classificadores básicos forem independentes, ou seja, o erro dos mesmos não estiver correlacionado, o grupo irá

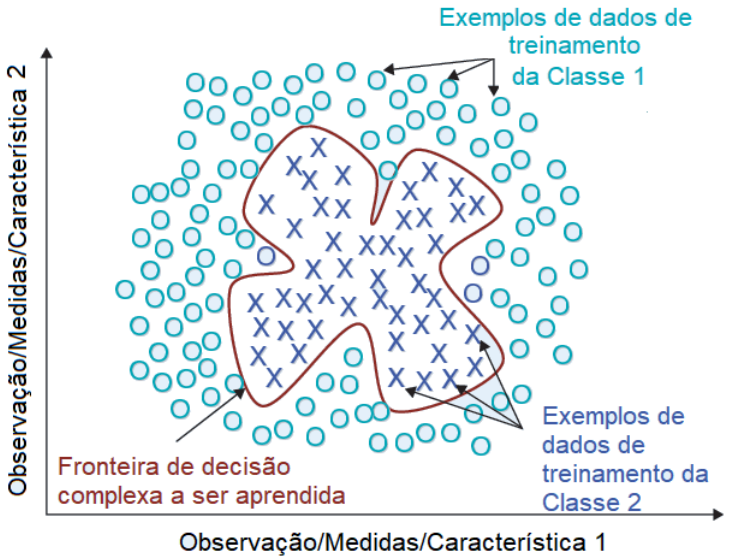


Figura 3.2: Exemplo de uma fronteira de decisão complexa que não pode ser generalizada por um classificador linear. Figura adaptada de (POLIKAR, 2006).

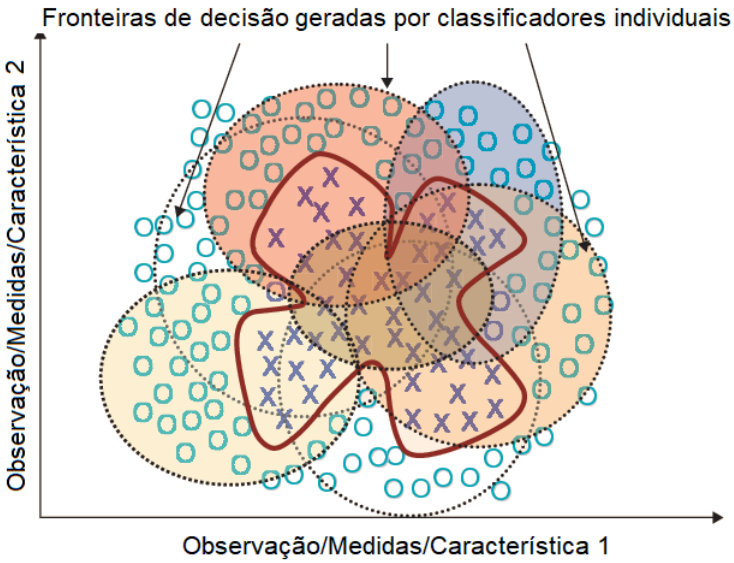


Figura 3.3: Demonstração de como o espaço de características pode ser dividido. Neste contexto, cada classificador fica responsável por classificar uma parcela dos dados. Figura adaptada (POLIKAR, 2006).

realizar uma previsão errônea se pelo menos a metade mais um dos classificadores realizarem uma previsão incorreta. A taxa de erro do comitê de classificadores é dada pelo somatório da probabilidade de 13 (metade mais um) classificadores errarem, a classificação mais a probabilidade de 14 classificadores e sucessivamente até a probabilidade de todos os 25 errarem a classificação. A Expressão 3.4 mostra o somatório total do comitê de classificadores que nesse caso se torna 0.06. O número de termos da Expressão 3.4 é

$13(25 - 13 + 1)$, ou seja, mais da metade dos classificadores base.

$$e_{grupo} = \sum_{13}^{25} \binom{25}{i} \epsilon^i (1-\epsilon)^{25-i} = 0.06. \tag{3.4}$$

Assim, para que um comitê de classificadores obtenha resultados mais promissores do que um único classificador duas condições são necessárias: os classificadores base devem ser independentes entre si e devem ser mais eficiente do que um classificador aleatório (TAN; STEINBACH; KUMAR, 2005).

3.2.1 Tipos de comitês de classificadores

Nesta seção, são introduzidos os principais tipos de comitê de classificadores disponíveis na literatura de inteligência computacional. Este é um assunto extenso, e tem-se aqui como objetivo somente dar uma introdução aos modelos utilizados na construção do modelo proposto neste trabalho. Sendo assim, para um entendimento mais aprofundado sobre o tema como métricas de ponderação de votação, combinação de classificadores, dentre outros assuntos relacionados, o leitor deve consultar trabalhos específicos na literatura.

3.2.1.1 Bagging

O método de *bagging* (abreviação de amostragem aleatória com reposição, do inglês, *bootstrap aggregating*) é um dos tipos mais antigos de comitê de classificadores, sendo também o mais intuitivo, tendo sido desenvolvido por Breiman em (BREIMAN, 1996). O método gera conjuntos de dados utilizando amostragem *bootstrap* do conjunto de dados original. Cada conjunto de dados é utilizado para treinar classificadores, geralmente do mesmo modelo, que apresentam hipóteses de indução distintas. A partir daí, os classificadores individuais são combinados por meio de votação majoritária ou ponderada de votos de suas decisões. Portanto, para cada amostra em particular, a classe atribuída à ela, é dada pela decisão da maioria dos classificadores que compõem o comitê.

Formalmente, as amostras sucessivas de *bootstrap* geram conjuntos na forma (x^b, y^b) e uma previsão $\hat{f}(x^b)$ é derivada de cada uma das amostras. Em problemas de regressão, a previsão final ou rótulo atribuído a amostra é dada por $\frac{1}{B}\sum_{b=1}^{B} \hat{f}(x^b)$, ou seja, através

da média das B previsões dos classificadores independentes ou pela maioria dos votos, no caso de problemas de classificação (GOLDSTEIN et al., 2010).

De acordo com (BREIMAN, 1996), os classificadores no método de *bagging* são treinados de maneira independente. Para a sua construção, é necessário a montagem de k subconjuntos de treinamento de forma idêntica, e a partir daí, replicar os dados de treinamento por re-amostragem com reposição.

Um dos usos mais efetivos do método é quando o conjunto de dados disponível é de tamanho limitado. Para garantir amostras de treinamento de dimensões mais adequadas, cada subconjunto gerado utiliza porcentagens relativamente altas do conjunto de treinamento no dimensionamento dos subconjuntos base (entre 75% e 100%). Essa medida, faz com que os subconjuntos de treinamento apresentem um maior nível de sobreposição. Assim, grande parte das amostras faz parte da maioria dos subconjuntos, assegurando a diversidade dos dados porém em bases com dimensões mais representativas (POLIKAR, 2006).

3.2.1.2 Boosting

O método de *Boosting* foi originalmente proposto em (SCHAPIRE, 1990). Neste trabalho Schapire provou que um classificador fraco (do inglês, *weak learner*), ou seja, um modelo que gera classificadores que não fazem mais do que adivinhações aleatórias, pode tornar-se um classificador forte (do inglês, *strong learner*) capaz de realizar predições que podem classificar corretamente senão todas, mais do que uma boa fração das amostras disponíveis.

No método de *boosting*, os classificadores são construídos sequencialmente, a ideia é gerar novas hipóteses subsequentes nas instâncias que estão sendo classificadas erroneamente. Dessa forma iterativa, cada classificador têm como objetivo reduzir os erros gerados pelos classificadores anteriores, aprendendo a partir do seu predecessor e atualizando os erros residuais. Dessa forma, o próximo classificador da sequência irá aprender sobre uma versão atualizada dos resíduos. (SCHAPIRE, 1999).

Como mencionado anteriormente, os classificadores base no método de *boosting* podem ser classificadores fracos que apresentam viés alto. Cada um contribui com ajustes essenciais para o processo de predição, permitindo que o método produza um classificador forte pela efetiva combinação dos classificadores fracos. O classificador forte, ao final do pro-

cesso, reduz o viés e a variância. O método de *boosting* utilizando classificadores do tipo árvores de decisão, frequentemente geram árvores de profundidade pequena e portanto, altamente interpretáveis. O método de *boosting* consiste de três etapas:

- Considere um modelo F_0, definido para predizer uma variável y. Esse modelo será associado com o resíduo $(y - F_0)$;
- Um novo modelo h_1 é ajustado aos resíduos pela etapa anterior;
- Agora, F_0 e h_1 são combinados para dar origem a F_1, ou seja, a versão atualizada de F_0. Dessa forma, o erro quadrático médio de F_1 será menor que o de F_0:

$$F_1(x) = F_0(x) + h_1(x) \tag{3.5}$$

Para melhorar a performance de F_1, pode-se modelar um novo modelo F_2 à partir dos resíduos de F_1, na forma:

$$F_2(x) = F_1(x) + h_2(x) \tag{3.6}$$

Assim, para o iterações, até que os resíduos tenham sido minimizados, têm-se:

$$F_o(x) = F_{o-1}(x) + h_o(x) \tag{3.7}$$

Para ilustrar a ideia do exemplo do método de *boosting* apresentado anteriormente, a Figura 4.1 demonstra a visualização da criação de modelos e ajuste de resíduos em um problema de regressão.

3.2.1.3 Adaboost

O *Adaboost* foi proposto em (FREUND; SCHAPIRE, 1997) sendo visto como uma versão mais geral do método de *boosting* original. Suas versões mais utilizadas são o Adaboost.M1 e Adaboost.R. Ambas são capazes de lidar com problemas de classificação multi-classe e regressão, respectivamente.

A ideia do *Adaboost* é gerar um conjunto de hipóteses e combiná-las através da ponderação da maioria de votos da classe predita por cada hipótese individual, onde cada

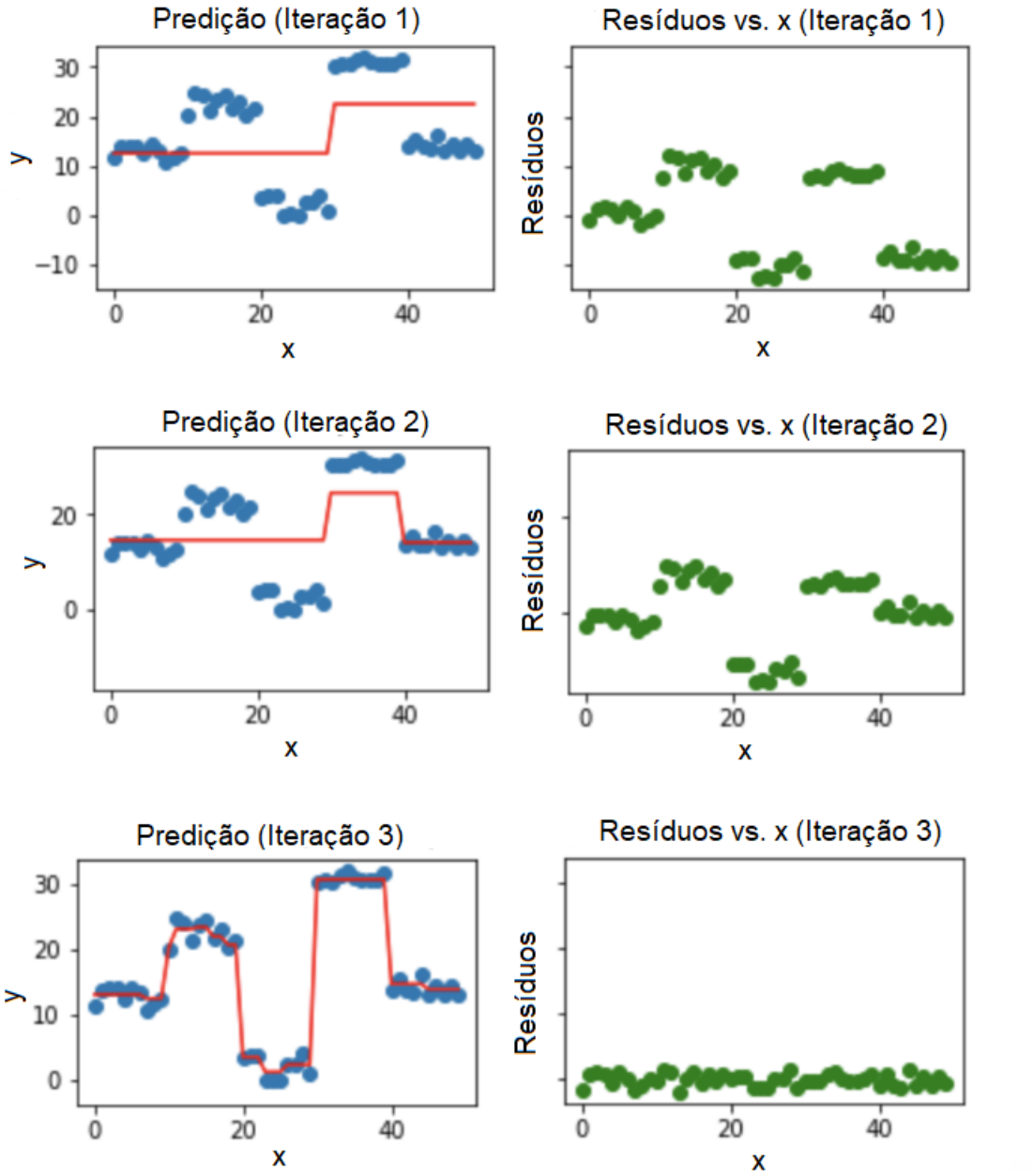


Figura 3.4: Visualização das iterações de algoritmo de *boosting*. Os pontos azuis são as amostras $\boldsymbol{x}$ plotadas com as saídas y. As linhas vermelhas representam valores preditos por um algoritmo construtivo como por exemplo, árvores de decisão. Os pontos verdes mostram os resíduos de acordo com as amostras $\boldsymbol{x}$ na i-ésima iteração. Cada iteração representa uma ordem sequencial do ajuste do modelo de *boosting*

.

hipótese é gerada por um classificador fraco, utilizando amostras determinadas por uma atualização iterativa da distribuição dos dados de treinamento. Essa distribuição atualizada garante que amostras não classificadas pelos classificadores predecessores tenham mais chances de serem incluídas no conjunto de treinamento utilizado pelo próximo classificador. Portanto, classificadores sequenciados têm mais chances de classificar corretamente

amostras que se apresentam mais difíceis de serem classificadas corretamente(FREUND; SCHAPIRE, 1999).

3.2.1.4 Mistura de especialistas

Mistura de especialistas é um tipo de comitê em que um conjunto de classificadores $C_1, ..., C_t$ o constituem. Segue-se, então para um segundo nível de classificadores C_{t+1}, utilizado para atribuir ponderações para combinações consecutivas de especialistas. Uma mistura de especialista pode ser vista com um algoritmo de seleção de classificadores, onde cada classificador individual é especialista em alguma fração do do espaço de características, e uma regra de combinação é utilizada para selecionar o classificador mais apropriado, ou a ponderação mais adequada dos classificadores de acordo com sua especialidade.

3.3 Extreme Gradient Boosting

Os classificadores base possuem a tendência de ter um poder preditivo limitado, mas quando são selecionados cuidadosamente utilizando algoritmos de *boosting*, eles formam um modelo geralmente mais acurado. Mais formalmente, algoritmos de *boosting* são algoritmos de aprendizagem que ajustam hipóteses de predição (SCHAPIRE; FREUND, 2012).

O *XGBoost* (do inglês, *Extreme gradient boosting*) é uma implementação particular de um *gradient boosting* (FRIEDMAN, 2000) de árvores de decisão ou regressão desenvolvido visando um aumento no desempenho computacional (e de qualidade de predição). O *gradient boosting* é uma técnica de aprendizagem de máquina que utiliza o *ensemble* do tipo *boosting* para construir classificadores. O método implementa um algoritmo de otimização e diversas funções de perda podem ser utilizadas. O *XGBoost* é pelo menos dez vezes mais rápido do que outras soluções existentes na literatura quando executados em uma única máquina e pode manipular conjunto de dados com bilhões de amostras em sistemas de memória distribuída (CHEN; GUESTRIN, 2016). A escalabilidade do *XGBoost* deve-se a um processo construtivo eficiente e a otimização de algoritmos que fazem parte do método. De acordo com (CHEN; GUESTRIN, 2016), dentre as características do *XGBoost*, destacam-se:

- capacidade de lidar com dados esparsos, ou seja, quando os dados possuem uma

grande quantidade de elementos que valem zero ou valores não presentes;

- um procedimento de seleção de atributos que são utilizados nas divisões das árvores que possibilita ao método lidar com instâncias com pesos diferentes;
- alto nível de paralelização, o que permite um processo rápido de aprendizagem e construção do modelo.

Para descrever o método do *XGBoost*, faz-se necessário introduzir os conceitos sobre *gradient boosting*. O primeiro deles é chamado objetivo de aprendizado regularizado. A regularização de modelos de árvores aditivas pode ser realizada de diversas forma, não se restringe somente a *boosting* de árvores, podendo ser utilizado qualquer algoritmo de *boosting* (FRIEDMAN, 2002).

Para um conjunto de dados com n amostras e m características: $D = \{(\boldsymbol{x}_i, y_i)\}$ ($|D| = n, \boldsymbol{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R}$), um modelo de comitê de árvores utiliza K funções aditivas para predizer a saída, dada pela Equação 3.8:

$$\hat{y}_i = \phi(\boldsymbol{x}_i) = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i),\ f_k \in F, \tag{3.8}$$

onde $F = \{f(\boldsymbol{x}) = w_{q(\boldsymbol{x})}\}(q : \mathbb{R}^m \mapsto T, w \in \mathbb{R}^T)$ é o espaço de árvores de decisão ou regressão e q representa a estrutura de cada árvore que mapeia uma amostra para um índice correspondente a uma folha da árvore. A variável T é o número de folhas em uma árvore. Cada f_k corresponde à uma estrutura de árvore independente q e o peso de uma folha w_i, onde w_i representa o valor da $i-$ésima folha.

Para uma determinada amostra, as regras de decisão dadas por q são utilizadas para classifica-lá dentro das folhas e são responsáveis pelo valor final de sua predição através da soma dos valores das folhas correspondentes em cada árvore (dada por w) (CHEN; GUESTRIN, 2016). Para aprender o conjunto de funções utilizadas pelo modelo, deve-se minimizar a seguinte função objetivo regularizada:

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{3.9}$$

onde l representa uma função de perda convexa, diferenciável que mede a diferença entre o valor de predição do conjunto de treinamento $\hat{y}_i$ e o rótulo da classe ou valor esperado y_i. O termo dado por Ω estima a complexidade do modelo para evitar *overfitting*.

O modelo inclui funções como parâmetros, então, métodos tradicionais de otimização para encontrar soluções no espaço euclidiano não podem ser utilizados diretamente. Por esse motivo, treina-se o modelo aditivamente na forma: a cada iteração t, o algoritmo primeiramente procura no espaço de funções F para encontrar uma nova função f_t que otimiza a função objetivo, então, ela é adicionada ao comitê de árvores. Formalmente, seja $\hat{y}_i^{(t)}$ a predição da $i-$ésima instância na $t-$ésima iteração. Dessa forma, encontra-se f_t que otimiza a seguinte função objetivo (CHEN; HE, 2015):

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t)} + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i) \tag{3.10}$$

Dessa forma, otimizar esse objetivo significa adicionar a função mais adequada para melhorar o modelo a cada iteração. Esse objetivo, na forma mostrada, ainda permanece difícil de ser otimizado. Assim, aproxima-se o mesmo por uma expansão em série de Taylor de segunda ordem, na forma:

$$L^{(t)} = \sum_{i=1}^{n} \left[l(y_i, \hat{y_i}^{(t)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \sum_{i=1}^{t} \Omega(f_i) \tag{3.11}$$

onde $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ e $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ representam o gradiente de primeira e segunda ordem. g_i se assemelha a medida de impureza utilizada para avaliar árvores de decisão, a diferença é que pode ser derivado para uma vasta quantidade de funções objetivo. Removendo os termos constantes, obtêm-se a seguinte função objetivo aproximada na iteração t:

$$\hat{L^{(t)}} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \sum_{i=1}^{t} \Omega(f_i) \tag{3.12}$$

O diferencial do *XGBoost* em relação a um algoritmo geral de *gradient boosting* que iterativamente adiciona funções para otimizar a equação acima é a utilização do termo de regularização que auxilia na prevenção do modelo de apresentar sobreajuste [1]. O termo de regularização, segundo (CHEN; HE, 2015), utiliza-se de $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$.

Definindo $I_j = \{i | q(x_i) = j\}$ como sendo o conjunto de instâncias da folha j. A Equação 3.12 pode então ser reescrita pela expansão do termo de regularização Ω, na

[1] sobreajuste, do inglês *overfitting* é utilizado para descrever quando um modelo se ajusta muito bem ao conjunto de dados utilizado, porém não obtém bons resultados para predizer novas observações

forma:

$$L^{\hat{(t)}} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 = \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2 \right] + \gamma T \tag{3.13}$$

Para uma estrutura fixada $q(x)$, pode-se computar o peso ótimo dado por w_j^* da j-ésima folha :

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{3.14}$$

O valor da função objetivo ótima correspondente é calculado por:

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{3.15}$$

Para exemplificar, a Tabela 3.1 mostra um conjunto de dados com cinco instâncias, a Figura 3.3 representa um comitê de classificadores composto por duas árvores (representação de árvores de classificação) a Figura 3.5 ilustra como a Equação 3.15 pode ser calculada.

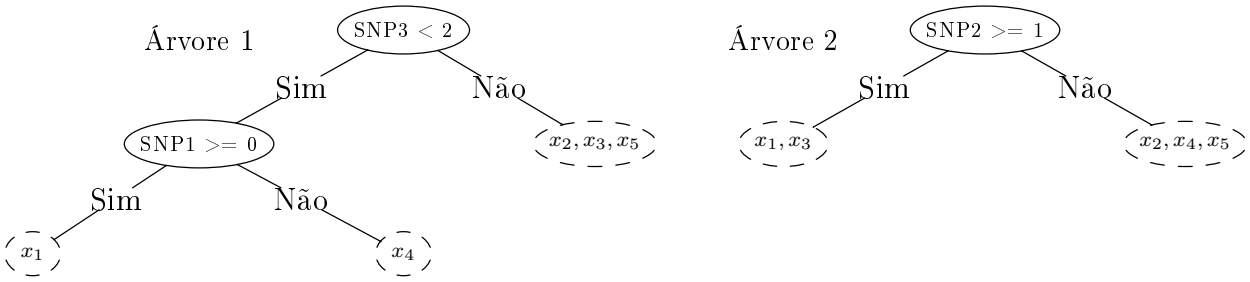
Um algoritmo de *greedy* é utilizado devido a impossibilidade de enumerar todas as estruturas de árvores determinadas por q. O algoritmo procura iterativamente dentre as características m_i em cada folha, os possíveis candidatos a divisão ou quebra da árvore. Encontrando o melhor candidato, o modelo executa a quebra na árvore na posição indicada até que a sua profundidade máxima seja alcançada. Sejam I_L e I_R nós do conjunto de instâncias da esquerda e direita respectivamente após uma quebra. Assumindo $I = I_L U I_R$, a função de perda reduzida depois da quebra é dada pela Equação 3.12, que na prática é utilizada para avaliar possíveis candidatos a quebra (CHEN; GUESTRIN, 2016).

$$L_{\text{quebra}} = \frac{1}{2}[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}] \tag{3.16}$$

Assim, podemos observar que a primeira classificação foi feita pela árvore 1 e a segunda pela árvore 2, respectivamente

Instância	Gradiente	Classe
x_1	g_1, h_1	1
x_2	g_2, h_2	0
x_3	g_3, h_3	1
x_4	g_4, h_4	0
x_5	g_5, h_5	0

Tabela 3.1: Representação de um conjunto de dados hipotético composto por cinco instâncias de duas classes diferentes. A instância x_1 e x_3 possuem o rótulo da classe 1 e as demais instâncias o rótulo da classe 0.



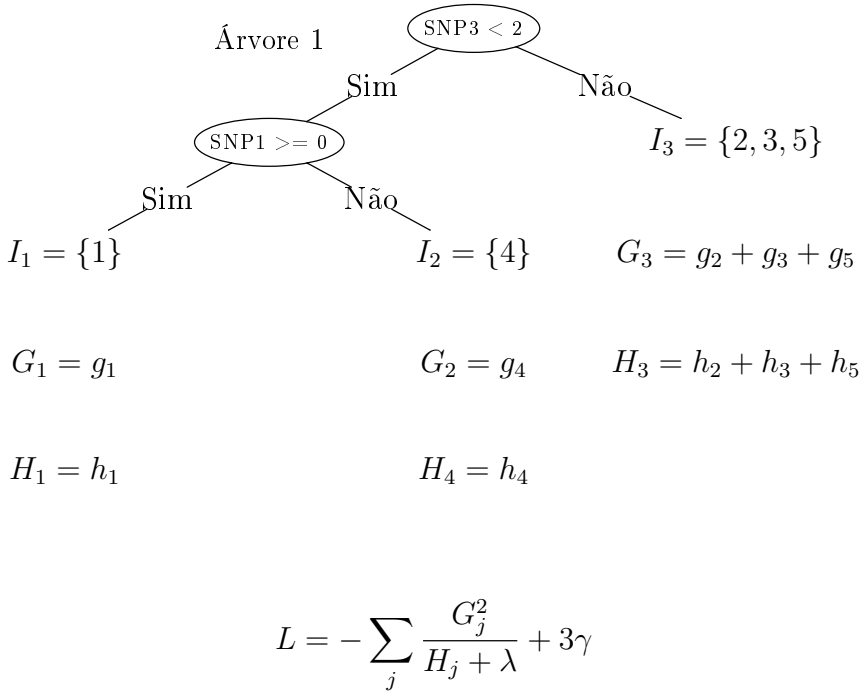
$$f(x_1) = 2 + 0.9 = 2.9$$
$$f(x_3) = -1 - 0.9 = -1.9$$

Figura 3.5: Exemplo de predição utilizando comitê classificadores de regressão. Para a árvore 1, considere que os valores atribuídos para as amostras durante o processo de treinamento: $x_1 = +2$, $x_4 = +0.1$ e $x_2, x_3, x_4 = -1$ e para árvore 2: $x_1, x_3 = +0.9$ e $x_2, x_4, x_5 = -0.9$. Assim, para x_1 ser classificado como caso, a árvore 1 indicou que ele deve apresentar 'SNP3 < 2' e 'SNP1 ≥ 0' e pela árvore 2 'SNP2 ≥ 1', a árvore 1 produziu um valor de saída para classificação de $+2$ e árvore 2 produziu um valor de 0.9. Assim o valor combinado das árvores resulta em $f(x_1) = 2 + 0.9 = 2.9$. O mesmo raciocínio vale para $f(x_3)$.

3.3.1 Características do XGBoost

Como mencionado anteriormente, o *XGBoost* é uma implementação particular do algoritmo de *gradient boosting*. De acordo com (CHEN; GUESTRIN, 2016), o seu desenvolvimento levou em consideração algumas características únicas para esse tipo de algoritmo, tais como:

- Regularização: o algoritmo possui a opção de penalização de modelos complexos. A utilização da regularização ajuda na prevenção de *overfitting*;



$$L = -\sum_j \frac{G_j^2}{H_j + \lambda} + 3\gamma$$

Figura 3.6: Exemplo do cálculo do *score* do *XGBoost*. Para uma árvore, é utilizado o gradiente de primeira e segunda ordem g_i e h_i nos nós folhas, soma-se os valores e utiliza-se do resultado para calcular a eficiência da árvore. Esta medida é similar a impureza de uma árvore de decisão, porém leva em consideração a complexidade do modelo.

- Manipulação de dados esparsos
- Esquema de quantil ponderado com dispersão: O *XGBoost* possui a implementação de um algoritmo eficiente para ponderar nós, de forma que dados ponderados sejam mais facilmente manipulados.
- Estrutura de blocos para aprendizagem paralela: Para um processo mais rápido dos dados, o *XGBoost* pode utilizar múltiplos núcleos da CPU. Isso é possível devido à uma estrutura de blocos no seu projeto de sistema. Os dados são ordenados e armazenados em unidades de memória denominadas de blocos. Isso permite que o *layout* de dados possa ser reutilizado por iterações subsequentes, ao invés de computá-los novamente. A Figura 3.7 mostra a estrutura em blocos do algoritmo. Essa estrutura também é útil para o processamento dos passos necessários para encontrar os pontos de quebra e sub-amostragem de colunas dos conjuntos de dados.
- Consciência de *cache*: O acesso de memória não-contínuo é necessário para calcular os gradientes através do índice da linha da matriz de um conjunto de dados. O *XGBoost* têm sido desenvolvido visando o uso ótimo do *hardware*. Essa tarefa é

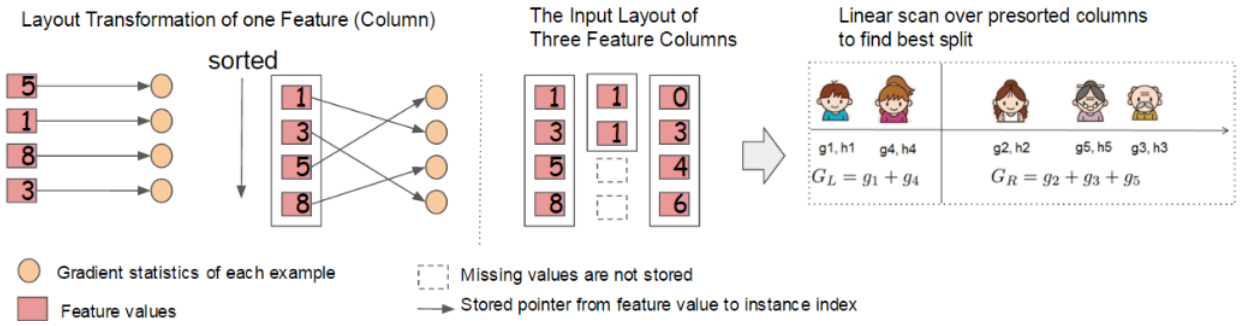


Figura 3.7: Representação da estrutura de blocos para aprendizagem paralela. Cada coluna em um bloco é ordenada pela valor da característica correspondente. Uma varredura linear em uma coluna em um bloco é suficiente para enumerar todos os pontos de quebra. Figura extraída de (CHEN; GUESTRIN, 2016).

realizada através da alocação de *buffers* internos em cada *thread*, onde o gradiente estatístico pode ser armazenado.

- Computação fora do núcleo: Essa característica otimiza o espaço em disco disponível e maximiza o seu uso ao lidar com grandes conjuntos de dados que não são alocados diretamente na memória.

O *XGBoost* é um modelo de *boosting* construído com dois objetivos claramente definidos: obter uma alta eficiência computacional por meio de implementações que utilizam todo o potencial do *hardware* bem como modelar um comitê de classificadores que apresente elevado nível de predição em bases de dados consideradas complexas devido a questões de dimensionalidades tanto quanto da padronagem de dados.

3.4 Floresta randômica

Uma floresta randômica ou aleatória(do inglês, *random forest*-RF) é uma combinação de preditores de árvores em que cada árvore depende dos valores de um subconjunto de treinamento amostrado de forma independente e com a mesma distribuição em todas as árvores na floresta. Esse subconjunto é gerado à partir do conjunto de treinamento original pelo método de *bagging*. Dessa forma, os subconjuntos utilizados por cada árvore na floresta são formados por amostragem *bootstrap* e cada nó de uma árvore classificadora é selecionado através de um conjunto de variáveis explicativas que foram amostradas aleatoriamente do conjunto de treinamento original que compõem todos os atributos das

amostras. O método foi desenvolvido em (BREIMAN, 2001). Em RF não existe a realização de poda nas árvores, pois geralmente a sua profundidade é pequena e pré-definida. A ideia é que através de um conjunto de árvores classificadoras ou regressoras que apresentam um pequeno viés e alta variância, ao se combinarem, gerem um classificador mais robusto que minimize a variância (TAN; STEINBACH; KUMAR, 2005).

Utilizando o método de *bagging*, sabe-se que quando o tamanho do conjunto de dados inicial é suficientemente grande, uma amostra *bootstrap* pode aparecer múltiplas vezes no mesmo subconjunto de treinamento, em contrapartida, outras amostras podem não aparecer. Assim, a probabilidade de uma instância não ser selecionada é de $1 - \frac{1}{n}$. Reescrevendo a probabilidade para n repetições, tem-se $\left(1 - \frac{1}{n}\right)^n$. Já para uma amostra ser selecionada dentre as n repetições a probabilidade é de $1 - \left(1 - \frac{1}{n}\right)^n$. Assim, utilizando o limite fundamental dado por $\left[\lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^n = e \approx 2.718\right]$, tem-se $\left[\lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^n = 1 - \frac{1}{e} \approx 0,63\right]$. Logo, quando o tamanho do conjunto de treinamento inicial dado por n é suficientemente grande, uma amostra *bootstrap* terá aproximadamente 63% do conjunto original, portanto, uma amostra do conjunto de teste terá 37% do conjunto original. Esse conjunto de teste é chamado de amostra *out-of-bag* (OOB). Uma das vantagens do método de *bagging* é apresentar um meio computacionalmente eficiente para estimar o erro de predição em um conjunto de teste independente (BREIMAN, 1996).

Ainda em (BREIMAN, 1996), foi demonstrado que a amostra OOB pode ser utilizada para estimar a medida de erro denominada taxa de erro OOB. A Figura 3.8 demonstra um exemplo de uma amostra de teste que compõe as amostras OOB sendo aplicada em cada uma das árvores em uma floresta randômica hipotética, na sequência, o processo de classificação da RF por maioria de votos.

Formalmente, a RF é um classificador que contém um conjunto de árvores de classificação ou regressão dada por $h(x, \Omega_k), k = 1, ...$ onde Ω_k são subconjuntos formados por *bagging*, distribuídos de forma idêntica e onde cada árvore contribui com apenas um único voto. Ao final do processo, a classe mais votada, será atribuída como rótulo para a amostra de entrada dada por x (BREIMAN, 2001).

A correlação de um conjunto de classificadores é dada pelo desempenho médio dos mesmos, de forma, que esse desempenho é medido de maneira probabilística em termos da margem do classificador, dada pela Equação 3.17 (TAN; STEINBACH; KUMAR,

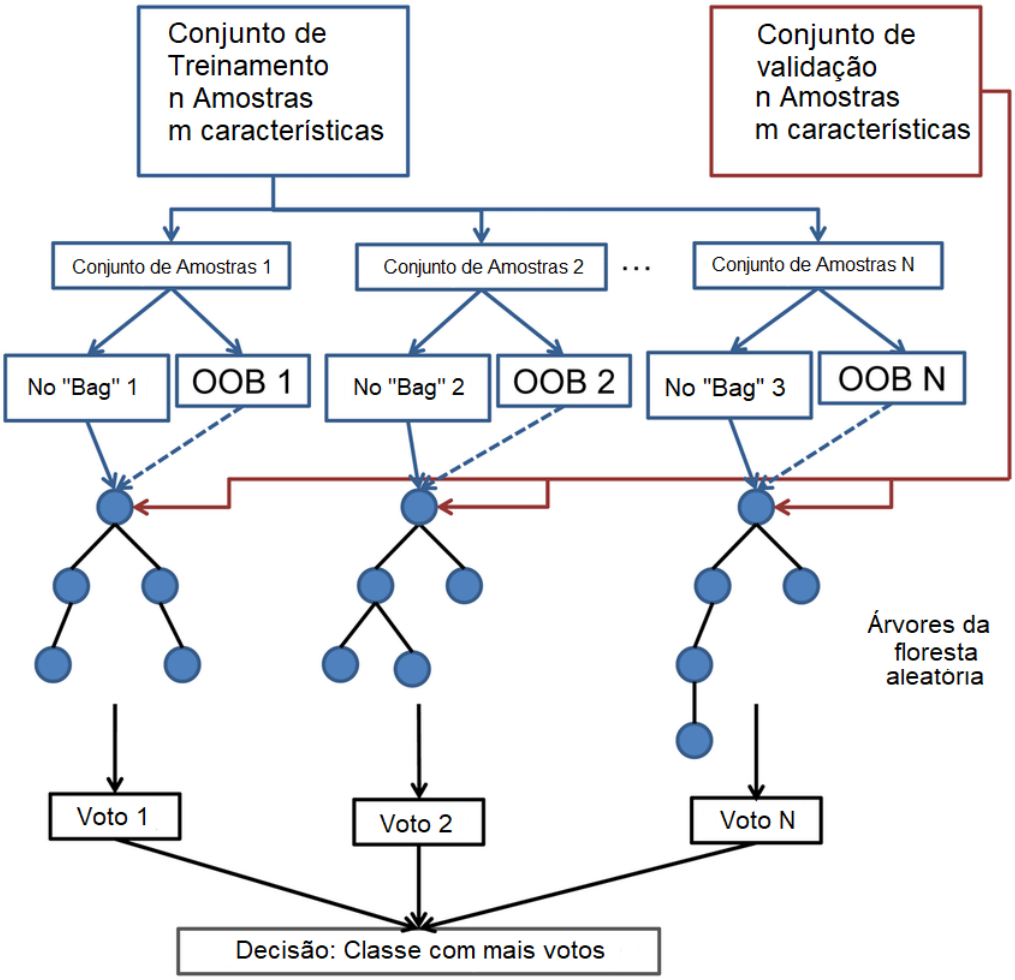


Figura 3.8: Representação da predição de uma amostra dada por x realizada por cada árvore presente em uma RF hipotética. Figura extraída de (LIAROKAPIS et al., 2013).

2005):

$$mr(X,Y) = P_{\Omega}(h(X,\Omega) = Y) - max_{j \neg Y}(P_{\Omega}(h(X,\Omega) = j) \tag{3.17}$$

onde $h(x,\Omega)$ é classe prevista para a amostra x de acordo com um classificador gerado a partir de um vetor aleatório Ω. Portanto, quanto maior for a margem, mais provável será a previsão correta para uma determinada amostra x.

A inicialização de um algoritmo de RF se dá a partir da seleção de uma amostra aleatória no conjunto de dados. O conjunto aleatório de amostras é selecionado realizando uma busca até encontrar um atributo da amostra ideal para ser tomado como nó de quebra da árvore. Este processo é repetido para cada nó até que uma árvore seja completamente gerada. As amostras que não fazem parte da amostragem de *bootstrap* são aplicadas em cada árvore para deverivar a taxa de erro OOB e medidas de importância de variável

(VI). O procedimento é repetido até que um número de árvores, definidos *à priori* sejam geradas, compondo uma floresta. Depois de treinadas, uma observação nova é atribuída à classe que recebe a maioria dos votos (BREIMAN, 2001).

O erro de generalização de uma RF é fortemente relacionado com a correlação entre das árvores e do seu poder de predição individual. Para melhorar a sua precisão, uma aleatoriedade é imposta na RF para a minimização da correlação. De acordo com (BREIMAN, 2001), as principais características de uma RF são:

- precisão comparável ao método *Adaboost*;
- robusto para *outliers* e ruídos;
- mais rápido que os métodos de *bagging* original e *boosting*;
- como os classificadores são independentes em sua geração, permite paralelização;
- oferece estimativas de erro, correlação e importância de variável.

3.4.1 Importância de variável

Em floresta randômica, existem duas medidas mais utilizadas para o cálculo de importância de variável (do inglês, *Variable Importance*-VI), são elas: importância de permutação (pVI) e importância de gini (gVI). A pVI representa o aumento no erro de classificação para a amostra OOB dada pela i-ésima amostra, depois do atributo j ser permutado na árvore k (LIAW; WIENER, 2002a). Formalmente, como analisado em (GOLDSTEIN; POLLEY; BRIGGS, 2011), para se definir pVI, é necessário apresentar as seguintes medidas:

- s_{ijk} representa o número de árvores k que utilizam o atributo j em algum nó e que erram na obervação da amostra i;
- r_{ijk} representa o número de árvores k que não utilizam o atributo j em algum nó e que erram na obervação da amostra i;
- ps_{ijk} representa o número de árvores k que utilizam o atributo j em algum nó e que erram na obervação da amostra i quando j é permutado;

- pr_{ijk} representa o número de árvores k que não utilizam o atributo j em algum nó e que erram na obervação da amostra i quando j é permutado;

Dessa forma, pVI_{ijk} é definido pela Equação 3.18

$$pVI_{ijk} = (ps_{ijk} + pr_{ijk}) - (s_{ijk} + r_{ijk}) = ps_{ijk} - s_{ijk}, \text{ sendo que } pr_{ijk} = r_{ijk} \tag{3.18}$$

Então, a Equação 3.19 indica como calcular pVI_{ij},pVI_{jk} e pVI_j

$$\begin{aligned} pVI_{ij} &= \frac{1}{ntree}\sum_{k=i}^{ntree}(ps_{ijk} - s_{ijk}) \\ pVI_{jk} &= \frac{1}{np}\sum_{i=i}^{np}(ps_{ijk} - s_{ijk}) \\ pVI_j &= \frac{1}{np \times ntree}\sum_{i=i}^{np}\sum_{k=i}^{ntree}(ps_{ijk} - s_{ijk}) \end{aligned} \tag{3.19}$$

Sendo np e $ntree$ parâmetros que representam o número de amostras OOB e o número de árvores na floresta, respectivamente. pVI pode ser visto como a qualidade preditiva do atributo j, porque é calculada para as amostras OOB. Dessa forma, se uma variável não apresenta importância para o processo de predição ($E(pVI) = 0$), a permutação não aumentaria nem diminuiria o erro de classificação.

A medida de importância gVI utiliza-se do índice gini (GI), apresentado anteriormente, para o crescimento de árvores na RF em tarefas exclusivas de classificação binária. A Equação 3.20 representa a medida gVI:

$$GI = 2p(1-p) \tag{3.20}$$

onde o parâmetro p representa a proporção da segunda classe. O atributo de quebra que minimiza GI é a mais adequada. sendo n o índice para um nó em uma árvore. Assim, a Equação 3.21 define as medidas $gVIU_{kmn}$, ou seja, a importância do atributo j no nó n na árvore k, gVI_{jk}, que representa a importância do atributo j na árvore k e gVI_j, a

importância do atributo j é dado por:

$$
\begin{aligned}
gVI_{jkn} &= (GI_{\text{pai}} - GI_{\text{filho da esquerda}} + GI_{\text{filho da direita}})np_{kn} \\
gVI_{jk} &= \frac{1}{np} \sum_{n_j \in \text{ tree k}}^{N} (gVI_{jkn}) \\
gVI_{j} &= \frac{1}{np \times ntree} \sum_{k=i}^{\text{ntree}} gVI_{jk}
\end{aligned}
\tag{3.21}
$$

Portanto, quanto maior o valor de gVI, melhor o desempenho da variável j em dividir os dados. Essa medida tem um enfoque distinto do pVI em relação a medição da qualidade preditiva no conjunto de testes das amostras OOB. Assim, a medida gVI pode ser vista como um teste χ^2, hierarquicamente aos eventos ocorridos no crescimento das árvores (GOLDSTEIN; POLLEY; BRIGGS, 2011).

Ainda em (GOLDSTEIN; POLLEY; BRIGGS, 2011), indica-se que a medida pVI é a mais comumente utilizada. Entretanto, quando o erro na amostra no conjunto de teste OOB é próxima de 50%, (ou seja, casos em que a predição do classificador não se difere de uma escolha ou predição aleatória) o índice gVI pode ser preferível.

3.5 Programação genética

Esta seção têm como objetivo descrever os conceitos sobre programação genética essenciais para o entendimento do modelo desenvolvido como parte deste trabalho. Modelos de programação genética são frequentemente atualizados e propostos na literatura, portanto, aborda-se aqui, somente conceitos fundamentais referentes à esse campo de estudo.

A programação genética é uma técnica da computação evolucionista (do inglês, *Evolutionary Computation*-CE) proposta por Koza (KOZA, 1992) que visa resolver problemas de otimização de forma automática sem a necessidade do usuário conhecer ou especificar a sua estrutura de forma aprofundada. De maneira formal, é uma meta-heurística que se difere dos demais modelos evolucionistas pela maior flexibilidade e complexidade na codificação dos indivíduos que compõem a população, o que traz a necessidade de um novo esquema sistemático na estrutura do método bem como na construção e uso dos chamados operadores genéticos para resolução computacional dos problemas de interesse.

Em um modelo de GP, inicialmente uma população com um determinado número de indivíduos é criada para dar inicio a um processo de evolução geracional. Os indivíduos

são avaliados e em cada geração, novos indivíduos são inseridos na população por meio dos operadores genéticos de recombinação e mutação, gerando uma nova população até que um individuo que atenda os requisitos para ser uma solução adequada seja encontrado ou até que algum critério de terminação seja atingido (POLI; LANGDON; MCPHEE, 2008). O fluxograma da Figura 3.9 de um algoritmo de programação genética indica os seguintes passos a serem executados (KOZA, 1994):

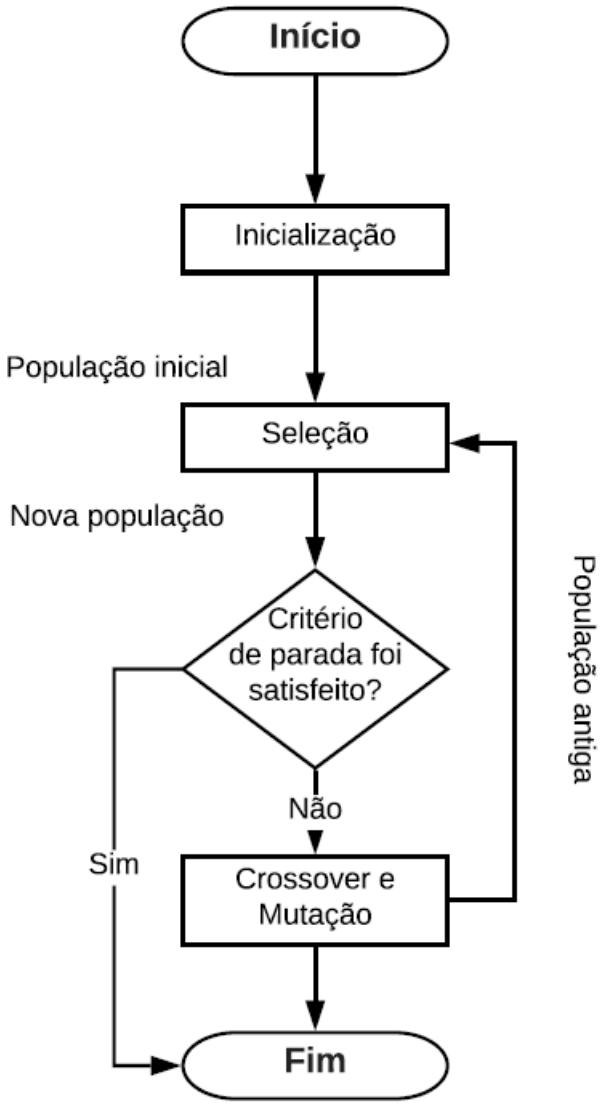


Figura 3.9: Fluxograma de um algoritmo de programação genética.

- Criação da população inicial de indivíduos: frequentemente, a população inicial de um sistema de programação genética é criada a partir de indivíduos completamente aleatórios ou baseando-se em algum padrão relacionado a estrutura de codificação utilizada como, por exemplo, árvores de decisão. O conjunto de possíveis soluções

dos indivíduos deve ser definida *apriori*, levando em consideração os elementos dos conjuntos de funções e terminais distintos para gerar a população. Define-se a profundidade máxima da árvore na codificação. Essa variável é utilizada para gerar populações com padrões distintos (seção 3.6.2;)

que definem a saída esperada da solução;

- Avaliação dos novos indivíduos: a cada geração, todos indivíduos da nova população devem ser avaliados para a determinação do sua respectiva aptidão;

- Critério de parada: Caso o critério de parada seja satisfeito, o algoritmo deve interromper a sua execução. Normalmente utiliza-se dois tipos de critérios de parada: (a) quando uma solução aceitável é encontrada. (b) quando o número de gerações limite é atingido. Neste último caso, o melhor individuo encontrado ao longo das gerações é o escolhido como a solução do problema;

- Seleção dos indivíduos mais promissores: os indivíduos que obtém os melhores resultados de avaliação tem maior probabilidade de gerarem proles mais eficientes para a resolução do problema;

- Aplicar os operadores genéticos: operados genéticos como reprodução e mutação são aplicados aos indivíduos selecionados para gerarem diversidade na população;

- Os novos indivíduos gerados são adicionados na população através dos métodos de reprodução.

3.5.1 Representação

Os indivíduos em um sistema de GP podem ser representados por diversas formas. Em programação genética gramatical, são frequentemente representados por árvores sintáticas [2]. Essas árvores são compostas por variáveis e constantes nos nós folhas que são conhecidos como nós "terminais". As operações aritméticas como (+,-,*) constituem, por exemplo, os nós internos. Elas podem receber parâmetros e são conhecidos como nós "funções". O conjunto formado por todos os terminais é denominado de conjunto terminal e o conjunto

[2]Uma árvore sintática é uma estrutura de dados em forma árvore, que representa a estrutura sintática de uma cadeia seguindo uma gramática formal. Os nós internos são chamados de não-terminais da gramática e os nós folha são rotuladas pelos símbolos terminais da gramática (LOUDEN, 1997).

formado por todas as funções de conjunto função. A combinação de ambos os conjuntos é denominada de conjunto primitivo de um sistema de programação genética (POLI; LANGDON; MCPHEE, 2008).

Entretanto, existem também variações de PG que utilizam estruturas lineares (OLTEAN, 2005; FERREIRA, 2001), grafos e até estruturas mais complexas formadas por múltiplos componentes. A relação é então dada por um conjunto de galhos ou subárvores agrupados sobre um nó especial (nó raiz).

Na literatura de PG é comum que as expressões sejam representadas em notação de prefixo, frequentemente adotada por programas em linguagem Lisp. Por exemplo, a expressão max(x+x,x+3*y) se torna (max (+ x x) (+ x (* 3 y))). A ideia dessa representação é a visualização dos relacionamentos das subárvores de cada expressão. A Figura 3.10 representa o seguinte programa (+ 2(* 10 4)) que tem como resultado de sua execução o valor 42.

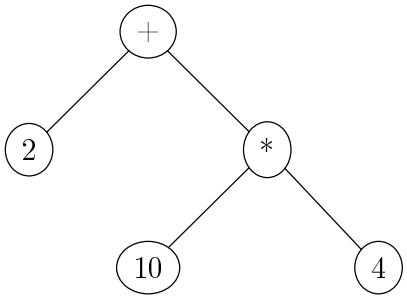


Figura 3.10: Representação de um individuo dado pelo seguinte programa: (+ 2(* 10 4))

3.5.2 População inicial

Similarmente à outros algoritmos da CE, os indivíduos iniciais de um sistema de PG são frequentemente gerados de forma aleatória. Esses indivíduos formam a população inicial do sistema. Os dois métodos convencionais de gerar os indivíduos da população inicial são conhecidos como *full* e *grow*, e um terceiro método formado pela combinação de ambos é denominado de *Ramped half-and-half* (POLI; LANGDON; MCPHEE, 2008). Em ambos os casos, os indivíduos são gerados de forma a não exceder a especificação adotada sobre a profundidade máxima dá árvore [3].

[3] O nível de um nó é dado pela distância deste nó até o nó raiz. A profundidade da árvore é dada pela sua altura, ou seja, a distância entre o nó terminal mais distante até o nó raiz.

O método *full* gera árvores completas em que todas as folhas têm a mesma profundidade. Os nós função são tomados de forma aleatória do conjunto função até que um nível a menos que a profundidade máxima da árvore seja atingida. A partir daí, somente os terminais são selecionados. A Figura 3.11 exemplifica os passos do processo de inicialização *full*.

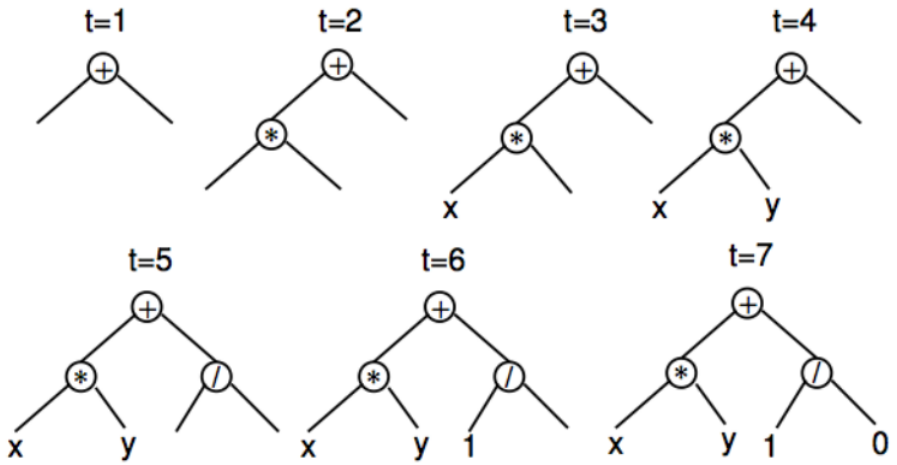


Figura 3.11: Criação de um indivíduo representado por uma árvore que possui profundidade máxima de tamanho 2 utilizando o método de inicialização *full*. A variável t representa cada passo da geração da árvore. Figura extraída de (POLI; LANGDON; MCPHEE, 2008)

Apesar o método *full* gerar árvores em que todos as folhas têm a mesma profundidade, não significa necessariamente que todas as árvores inicias possuem o mesmo número de nós ou a mesma forma. Essa fato somente acontece quando todas as funções no conjunto primitivo possuem aridade iguais.

O método *grow* permite a criação de árvores mais variadas com formas e tamanhos diferentes. Os nós são selecionados à partir de todo o conjunto primitivo até que a profundidade máxima seja alcançada. Uma vez que tal fato acontece, somente os terminais podem ser selecionados. A Figura 3.12 demonstra o processo da criação de uma árvore pelo método *grow*.

Uma combinação desses dois métodos denominada de *ramped half-and-half* foi proposta em (KOZA, 1992). Neste método metade da população inicial é construída usando o método *grow* e a outra metade o *full*. Nesse método diversos limites de profundidade podem ser escolhidos para assegurar que as árvores possuam tamanhos e formas variadas.

Existem também casos em que a população inicial não precisa ser completamente

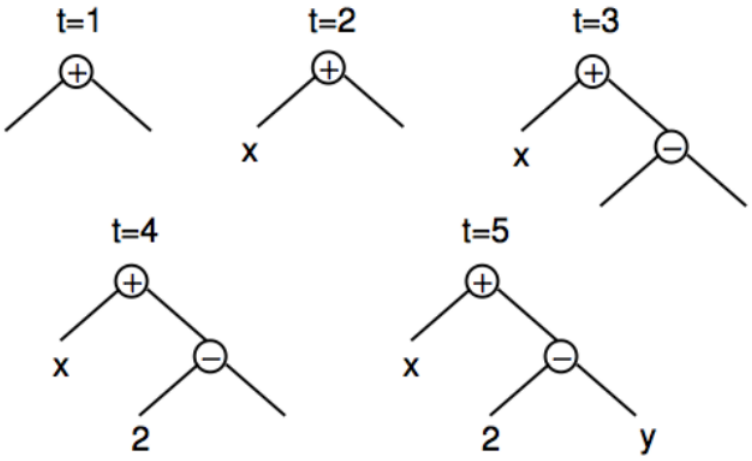


Figura 3.12: Criação de um indivíduo representado por uma árvore que possui profundidade máxima de tamanho 2 utilizando o método de inicialização *grow*. O terminal selecionado no passo $t = 2$ causa o fechamento da subárvore esquerda, já que a partir desse ponto, nenhum outro nó pode ser selecionado. Isso faz com que a profundidade máxima não possa ser alcançada nesse ponto. Figura extraída de (POLI; LANGDON; MCPHEE, 2008)

aleatória. Há situações em que o espaço de busca pode conter elementos que devem ter mais pesos que outros e com a necessidade de haver uma proporção ou a garantia da existência de sua ocorrência na população inicial para direcionar o processo de evolução (POLI; LANGDON; MCPHEE, 2008)s.

3.5.3 Seleção

Na programação genética, assim como na maioria dos algoritmos de CE, os indivíduos são selecionados baseados em alguma métrica probabilística definida em sua aptidão para compor uma nova geração. Dessa forma, indivíduos mais aptos têm chances maiores de gerarem proles do que indivíduos menos aptos. O método de seleção mais comum utilizado em PG é o de seleção por torneio. Nesse método, um número pré-definido de indivíduos são selecionados da população. Há uma comparação entre o valor de avaliação de cada um e o melhor é selecionado para a geração de proles. Dois torneios são necessários para a realização da recombinação onde dois pais são selecionados (KOZA, 1994).

Além do método de seleção por torneio, outro método comum é o de seleção por roleta. Este método de seleção é bastante utilizado em algoritmos genéticos e adota um mecanismo que simula o funcionamento de uma "roleta". Cada "fatia" da roleta é ocupada de maneira proporcional ao valor de cada aptidão. Um número aleatório é gerado e o

indivíduo que ocupar a fatia em que o número foi gerado é selecionado (KOZA, 1992).

Além dos dois métodos citados, outros métodos de seleção utilizados em CE podem ser adaptados para a GP (MüHLENBEIN; SCHLIERKAMP-VOOSEN, 1993).

3.5.4 Avaliação

Através de uma analogia com o processo de evolução descrito por Darwin, os seres vivos mais aptos são selecionados naturalmente com base em sua adaptabilidade ao meio ambiente em que estão inseridos. O mesma ideia se aplica a GP, ou seja, à cada indivíduo é atribuído um valor de aptidão ou *fitness*. Os melhores indivíduos receberão melhores valores de aptidão e terão maiores chances de serem selecionados para gerarem proles.

A forma de avaliação dos indivíduos irá depender fundamentalmente do domínio da aplicação. Frequentemente, é dado um conjunto de casos de treinamento ou *fitness cases*, com os valores de entrada e saída a serem aprendidos, ou seja, no padrão de aprendizagem supervisionada. Dessa forma, para cada indivíduo é fornecido os valores de entrada, a sua resposta é então comparada com o valor de saída, e quanto mais próximo do valor de saída estiver a resposta do indivíduo, melhor será o seu valor de aptidão (POLI; LANGDON; MCPHEE, 2008).

Em sistemas de PG, os indivíduos passam por um processo construtivo para que possam ser avaliados, o que ocorre em todo processo evolutivo. Uma árvore é interpretada pela execução de seus nós em uma ordem que garante que os argumentos de cada nó função sejam respeitados. O processo é realizado normalmente executando a árvore recursivamente, começando pelo nó raiz até os seus nós folhas.

3.5.5 Operadores genéticos

Depois dos indivíduos serem selecionados, os operadores genéticos devem ser aplicados. A aplicação dos operadores genéticos têm como finalidade a construção de indivíduos mais aptos para composição da próxima geração da população. Esse processo é realizado através da combinação ou variações sobre os indivíduos, permitindo que o espaço de busca seja melhor explorado. De acordo com (KOZA, 1992), os três principais operadores genéticos são:reprodução, cruzamento ou recombinação, que atua sobre dois indivíduos, e a mutação, que atuam em um único indivíduo. Os três operadores são descritos a seguir:

3.5.5.1 Reprodução

Neste tipo de operador genético, um ou mais indivíduos são selecionados e copiados sem sofrer nenhuma alteração em sua estrutura para a próxima geração.

3.5.5.2 Cruzamento

No operador de cruzamento (do inglês, *crossover*), dois indivíduos são selecionados para gerarem duas proles. Nesse operador, um ponto é aleatoriamente escolhido na árvore de cada um dos indivíduos selecionados como pais e as subárvores abaixo destes pontos são trocadas. Pode-se observar na Figura 3.13 um exemplo da aplicação do cruzamento. Para tal, foram escolhidos dois indivíduos: (3 + (x * x)) e (2 - (1 - x)). Foram escolhidos aleatoriamente um nó em cada árvore (serrilhado). As proles geradas são (3 + (1 - x)) e (2 - (x * x)).

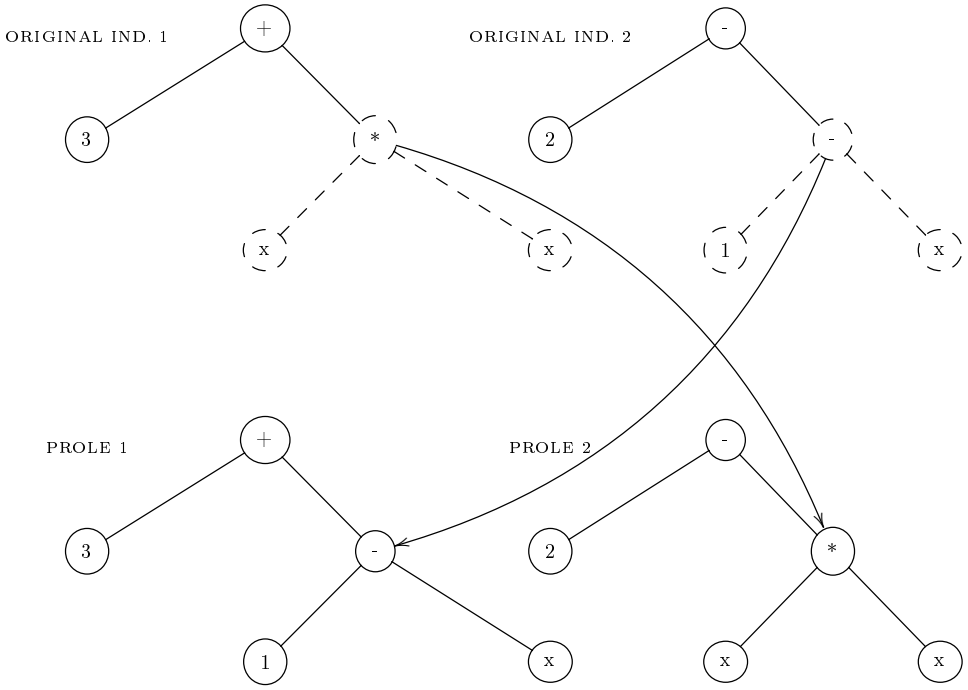


Figura 3.13: Exemplo do cruzamento entre dois indivíduos.

3.5.5.3 Mutação

O operador de mutação á aplicado somente em um indivíduo por vez. O novo descendente é criado a partir de modificações na estrutura do indivíduo de origem. Normalmente um nó na árvore é selecionado de forma arbitrária. A partir desse nó uma nova subárvore

aleatória é criada e alocada nesta posição. A antiga subárvore é descartada e um novo indivíduo é gerado.

3.5.6 Critérios de terminação

Os critérios de terminação aplicados em GP podem incluir o número máximo de gerações atingida ou quando um ou mais indivíduos alcançam uma faixa aceitável próxima da aptidão esperada para a solução ótima, ou seja, com baixo nível de erro. O processo pode terminar também quando a execução do sistema alcança um tempo limite ou pela estagnação do processo, ou seja, quando o algoritmo não consegue mais encontrar soluções mais promissoras que as atuais no decorrer das gerações (POLI; LANGDON; MCPHEE, 2008).

3.6 Métodos para avaliação de classificadores

3.6.1 Validação cruzada

A validação cruzada (do inglês, *Cross-Validation*-CV) é uma medida utilizada para avaliação e comparação de modelos de classificadores através da estimação da acurácia. A ideia é dividir o conjunto de dados com n observações em k partes ou dobras iguais (do inglês, *k-folds*) (nos casos onde n pode ser particionado igualmente em k partes ou $k-1$ partes iguais onde a k-ésima parte é constituída pelo número de observações iguais ao resto da divisão entre n e o total das partes k). Dessa forma, através do método de CV, cada observação é utilizada um mesmo número de vezes para o treinamento e somente uma vez para o teste.

Para exemplificar, a Figura 3.14 apresenta o processo de validação cruzada. No exemplo, o conjunto de dados é particionado em 5 partes, ou seja $k = 5$. Depois de particionado, o conjunto de dados utiliza 4 partes como treinamento e uma parte é separada para cada aplicação. A parte separada é utilizada como conjunto de teste. O processo é repetido k vezes até que todos as aplicações sejam executadas e todas as k partes tenham sido utilizadas como conjunto de teste. Dessa forma cada parte é utilizada somente uma vez para teste e 4 vezes para compor o conjunto de treinamento. No final do processo, o erro de predição é estimado a partir da combinação do erro de todas as aplicações. Através

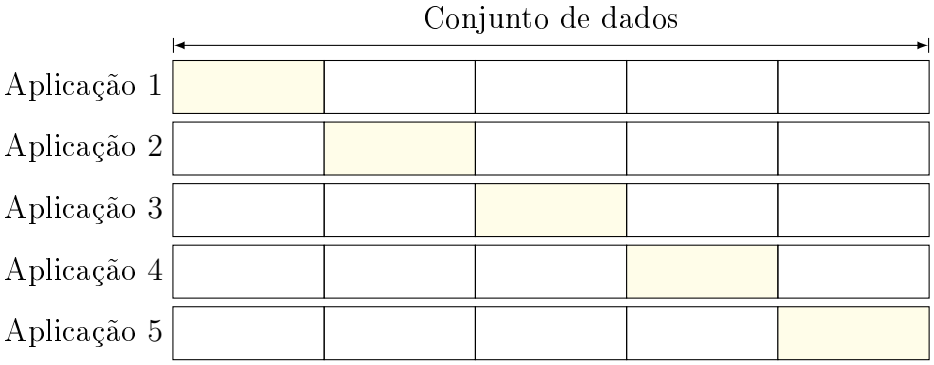


Figura 3.14: Exemplo de validação cruzada de 5 partes. Para cada aplicação, uma parte é selecionada pra ser o conjunto de teste e as demais são combinadas para formar o conjunto de treinamento. Pode-se observar que os conjuntos de treinamento são representados pelas células em brancos e os conjuntos de teste em amarelo. A predição final do modelo é dada pela combinação das 5 predições, uma de cada aplicação. Dessa forma, o erro de predição pode ser determinado

desse processo, o viés de cada k parte é reduzido ou eliminado em comparação a utilização de se dividir o conjunto em apenas dois subconjuntos (como por exemplo a divisão entre dois blocos: treinamento e teste).

O método de validação cruzada foi uma medida utilizada para os modelos de classificação construídos no presente trabalho. Dessa forma, através da utilização dessa medida, é possível comparar as predições geradas por cada modelo em cada partição para a medida de avaliação de classificação adotada no trabalho.

3.6.2 Área abaixo da curva ROC

Para discutir essa medida de avaliação, primeiramente sabe-se que no problema abordado têm-se duas saídas para classificar os indivíduos. A saída $y = 1$ indica pertencimento ao grupo de casos e a saída $y = 0$ ao grupo de controle (ou em alguns casos, pode-se considerar duas classes, uma positiva e uma negativa). Assim, para predizer a saída, uma regra de predição pode ser definida como: $\hat{y} = 1$ ou $\hat{y} = 0$, e a saída de um classificador gera uma valor entre 0 e 1. Essa medida é baseada em observações do conjunto de teste, dessa forma o objetivo é maximizar o poder de generalização do classificador avaliado.

Utilizando-se de um erro de predição comum, frequentemente entende-se que $\hat{y}_i = 1$[4] se a saída do classificador for próxima de 1 e $\hat{y}_i = 0$ se a saída for próxima de 0. A curva ROC (do inglês, *Receiving Operating Characteristics*) auxilia na identificação do ponto

[4] $\hat{y}_i$ representa a predição atribuída pelo classificador depois de selecionado um valor entre 0 e 1 para indivíduo do conjunto de teste.

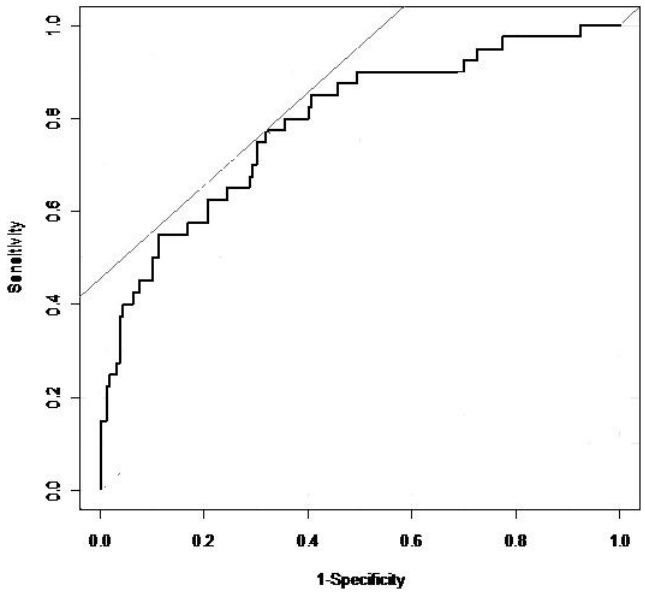


Figura 3.15: Exemplo hipotético do plot de uma curva ROC. O ponto de corte é escolhido pelo que mais se aproxima do canto superior esquerdo do gráfico.

de corte em que se pode atribuir um individuo ao grupo caso ou controle. Dessa forma a curva ROC plota $P(\hat{y}_i = 1|y = 1)$ sobre $1 - P(\hat{y}_i = 0|y = 0)$, ou seja, a sensibilidade sobre 1−especificidade para todos os possíveis pontos de corte entre 0 e 1 (FACELI et al., 2011).

O ponto de corte é definido baseado na combinação ótima da sensibilidade e especificidade. A ideia é que classificar o indivíduo como falso positivo ou falso negativo pode prejudicar o processo de predição do classificador. O gráfico de uma curva ROC é bidimensional e os eixos X e Y representam respectivamente a taxa de falsos positivos e a taxa de verdadeiros positivos. Dessa forma, o desempenho de um classificador sobre um conjunto de dados pode ser plotado em um gráfico como um ponto no espaço bidimensional (FACELI et al., 2011). Uma medida associada a curva ROC é denominada área abaixo da curva ROC (do inglês, *Area under ROC curve*-AUC). Assim, o classificador que possui a maior AUC é considero como sendo o que obteve o melhor desempenho. Os valores de AUC variam entre 0 e 1 e quanto mais próximo de 1, melhor é o classificador. A Figura 3.15 apresenta o gráfico de uma curva ROC.

Depois de definido o ponto de corte, o poder discriminativo do modelo é avaliado. Para isso 5 métricas são utilizadas: Acurácia, sensibilidade, especificidade, taxa de falso positivo e taxa de verdadeiro positivo. Essas medidas podem ser desdobradas com base em uma matriz de confusão, como pode ser observado na Tabela 3.2 :

		valor observado	
valor predito		y=1	y=0
	y=1	VP	FP
	y=0	FN	VN

Tabela 3.2: Exemplo de uma matriz de confusão. Figura adaptada de (FACELI et al., 2011).

A AUC é uma alternativa robusta para se avaliar uma classificação binária em comparação com a utilização da acurácia do modelo. Uma outra vantagem da AUC é a sua independência da distribuição de saída. Dessa forma, mesmo as observações positivas forem duplicadas duas vezes em relação as linhas no conjunto de dados, a AUC do modelo permanecerá sendo a mesma. Além disso, ela é uma medida da capacidade discriminativa de um teste, ou seja, representa a capacidade de um teste classificar corretamente indivíduos que possuem e que não possuem um fenótipo. Em outras palavras, se sortearmos ao acaso uma amostra com a $y = 1$ e outra com $y = 0$ da população do conjunto de dados, a AUC nos oferece a probabilidade de classificarmos corretamente este par de amostras.

3.7 Considerações do capítulo

Este capítulo teve como objetivo introduzir o referencial teórico a respeito da metodologia desenvolvida nesta tese, sendo apresentado somente o considerado necessário para o entendimento e acompanhamento do processo construtivo do modelo que será desenvolvido. Cada técnica apresentada foi utilizada como parte de uma das etapas da metodologia proposta. Em relação aos primeiros temas apresentados como árvores de decisão, comitê de classificadores e suas variantes, é importante ressaltar que formam a base construtiva para a primeira etapa e uma das mais importantes da metodologia, que será responsável pela escalabilidade do modelo, ou seja, pela possibilidade de se trabalhar com grandes conjuntos de dados e de selecionar subgrupos através de critérios específicos para etapas posteriores. O algoritmo de floresta randômica e o modelo evolutivo de programação genética serão utilizadas na metodologia responsável pela geração de modelos descritivos que mapeiam as relações entre marcadores. No próximo capítulo são apresentadas as etapas do modelo. Dessa forma, os temas abordados neste capítulo são fundamentais e no auxilio e compreensão dos desenvolvimentos de uma ferramenta robusta para detecção de epistasia.

4 Metodologia proposta

4.1 Introdução

Esse capítulo apresenta o modelo proposto para o problema abordado nessa tese. A solução proposta trata-se do desenvolvimento e implementação de um modelo computacional aplicado a investigação de SNPs envolvidos em interações epistáticas em conjuntos de dados massivos e de baixa herdabilidade, como foi discutido nos capítulos anteriores. O modelo é composto de três etapas fundamentais e distintas, com finalidades específicas para os seguintes problemas: (1) seleção de potenciais subgrupos de marcadores; (2) ranqueamento de marcadores; (3) geração das soluções que apresentem padrão de epistasia e interpretabilidade. Diferente da maioria dos modelos computacionais existentes na literatura, apresentados no Capítulo 2, o objetivo é que o resultado final gerado pelo modelo represente uma solução compreensível para pesquisadores com formação multidisciplinar e dos campos de pesquisa que permeiam o tema abordado no trabalho.

O embasamento teórico de aprendizagem de máquina e computação evolucionista que compõem parte da estrutura do modelo, foi apresentado e está disponível no Capítulo 3. Devido a grande existência de material disponível para consulta destinado a esse conteúdo teórico e extenso, e como não é finalidade do trabalho discutir detalhadamente cada algoritmo abordado no presente texto, parte-se do pressuposto que o leitor possua compreensão do que foi abordado para um melhor entendimento do material apresentado a seguir.

4.1.1 Seleção de subconjuntos

Um dos maiores desafios do problema estudado nesta tese é encontrar um grupo reduzido de marcadores de interesse dentre dezenas ou centenas de milhares que compõem um conjunto de dados. O estudo em questão traz uma dificuldade intrínseca que o coloca dentro de uma classe de problemas de mineração de dados conhecida como encontrar "agulhas no palheiro"(do inglês, *The Needles-in-Haystack Problem*) (MORELAND; TRUEMPER, 2009). Neste contexto, o problema é definido de forma que cada indivíduo seja representado por um vetor de característica, onde os atributos contém informações sobre

cada marcador, indicando se o mesmo é homozigoto dominante, recessivo ou heterozigoto. Em problemas do tipo caso-controle, existem dois grupos de indivíduos, sendo que um dos grupos possui um fenótipo de interesse, ou seja, uma conjectura de marcadores que explica parcialmente ou totalmente esse fenótipo. O objetivo é identificar esse grupo de marcadores explicativos para predizer quais indivíduos são susceptíveis ao fenótipo estudado.

Formalmente têm-se um conjunto de vetores (associados aos marcadores) conhecidos como "palheiro" e um vetor "agulha" adicional. Alguns desses vetores no "palheiro" são similares ao vetor "agulha" de acordo com uma relação desconhecida envolvendo um subconjunto de seus atributos. Um "oráculo" disponível aceita qualquer vetor do conjunto de "palheiros" e diz se ele é ou não uma agulha. O objetivo é identificar todos os vetores "agulha" escondidos no "palheiro" enquanto o número de chamadas ao "oráculo" é minimizada (MORELAND; TRUEMPER, 2009).

Encontrar os atributos que fazem parte dessa relação torna o problema mais complexo do que uma seleção de atributos. Neste caso, uma variável isolada pode não ser suficiente para explicar a expressão do fenótipo, mas um grupo das mesmas pode ser mais informativo.

Através dessa perspectiva, a ideia inicial destinou-se a determinar uma forma de tornar o problema computacionalmente viável para encontrar tais relações dentro de um conjunto com uma grande quantidade de atributos. Para tal, o classificador *XGBoost* apresenta propriedades interessantes no manuseio e trato de grandes volumes de dados. Como foi discutido no Capítulo 3, a implementação do *XGBoost* é direcionada para trabalhar com conjuntos grandes de vetores e criar modelos robustos de forma rápida.

Dessa forma, considere um conjunto de dados composto de N marcadores. Esse conjunto é dividido em subgrupos de tamanho $r < N$ sem repetição de SNPs. O processo é realizado até que todos os SNPs do conjunto de dados inicial estejam alocados nos subgrupos disponíveis. A Figura 4.1 demonstra um exemplo hipotético da criação de subgrupos de tamanho três.

Depois de serem formados n grupos de tamanho r, são criados pares desses grupos sem repetições que são posteriormente combinados. Esse processo pode ser visto como a criação de um *power set* (conjunto de todos os subconjuntos de um conjunto) em que

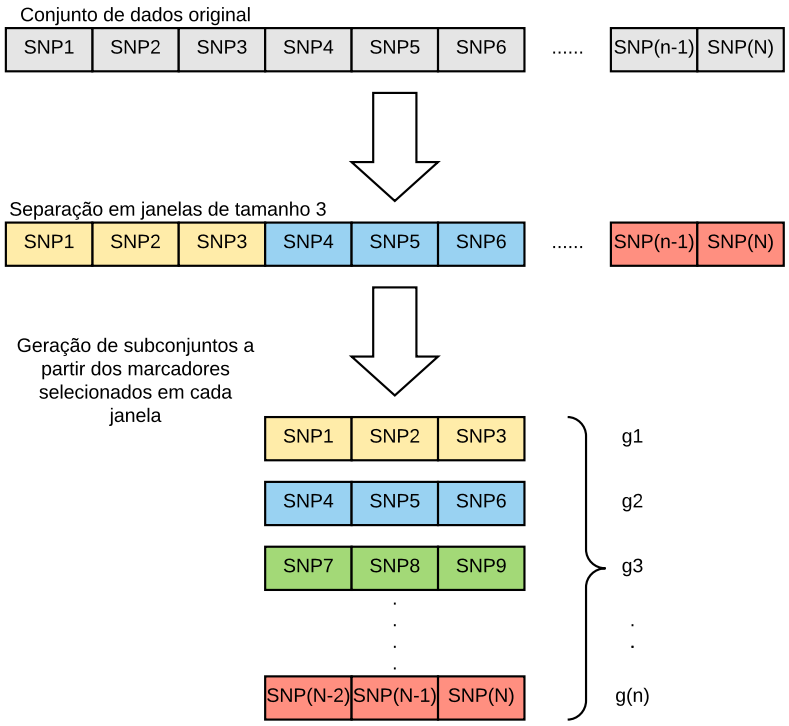


Figura 4.1: Criação dos subgrupos g_i, onde $i = 1, .., n$, onde $n = N/r$. Os subgrupos são formados a partir do conjunto de dados original. O tamanho da janela é definido e os subgrupos do tamanho da janela são criados até que todos os SNPs estejam alocados nos respectivos subgrupos.

cada combinação possui apenas dois subconjuntos no exemplo[1]. Por exemplo, seja A um conjunto com 3 elementos ($A = \{1, 2, 3\}$). Um *power set* do conjunto A é dado por: $P(A) = \{\{\}, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$. Pode-se observar que $P(A)$ possui um total de 8 elementos. No cenário hipotético discutido aqui, utiliza-se somente os subgrupos de tamanho dois. Dessa forma o mesmo conjunto A seria composto por: $P(A) = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$, gerando um total de apenas 3 elementos. A Figura 4.2 mostra a criação do conjunto $P(A)$ a partir dos subgrupos de A.

Ao final do processo de geração dos conjuntos de todas os possíveis pares de elementos de tamanho 2, são criadas as combinações, denominadas aqui de c_i, onde $i = 1, ..., k$ e k é o total de combinações. Cada c_i contém o dobro da cardinalidade de marcadores dos subgrupos, dada por $c_i = g_j + g_{j+1}$ onde j representa o índice de cada subgrupo g. Assim, o número total de combinações de k sugbrupos tomados dois a dois é $c_{k,2} = \left(\frac{k!}{2!(k-2)!}\right)$.

[1] no exemplo são discutidos buscas por pares de marcadores, para interações de ordens mais altas, as combinações devem ser feitas com combinações de mais grupos

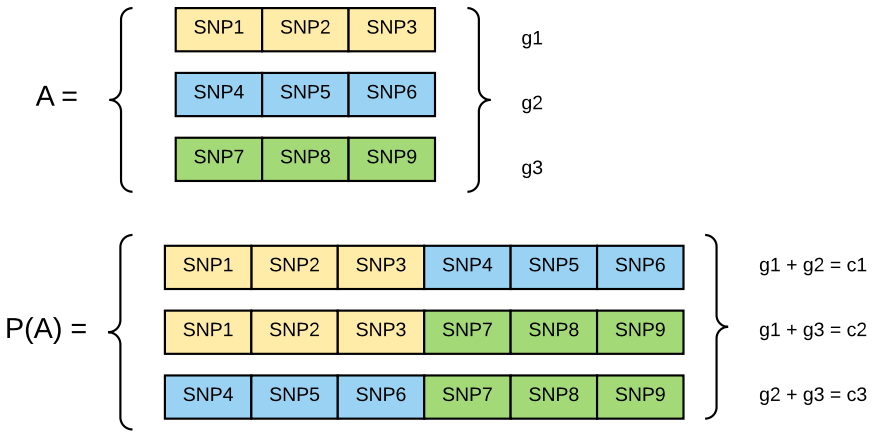


Figura 4.2: Exemplo da formação de $P(A)$. O conjunto de todos os possíveis de pares de subgrupos de A. Geração das combinações dos elementos do conjunto A são denominadas de c_i, onde neste exemplo $i = 1, ..., 3$.

Em seguida, é utilizada validação cruzada de k partes. No modelo proposto foi utilizado $k = 5$ para avaliar o conjunto de dados e diminuir o custo computacional nessa etapa do processo. Para cada combinação c_i o algoritmo é executado 5 vezes, na i-ésima execução, onde $i = 1, 2, 3, 4, 5$, a i-ésima parte será utilizada como subconjunto de teste e as outras 4 são combinadas e utilizadas como conjunto de treinamento para cada execução.

A ideia do processo de validação cruzada é eliminar o viés existente quando se particiona o conjunto de dados em apenas duas partes(conjunto de treino e conjunto de teste). Essa divisão simples pode subestimar ou superestimar o classificador ou modelo avaliado. Para cada processo da execução de validação cruzada em c_i, um modelo é criado pelo algoritmo *XGBoost*. A predição final é dada pela média da predição das cinco aplicações geradas pela validação cruzada.

O *XGBoost* utiliza classificação binária sobre a validação cruzada de cada subconjunto c_i. Para isso utiliza-se uma função de perda baseada em regressão logística, definida pela Equação 3.10, apresentada no capítulo 3, que é aqui replicada:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t)} + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i) \tag{4.1}$$

onde l representa uma função de perda convexa e diferenciável responsável pela medição da diferença entre o valor de predição do conjunto de treinamento $\hat{y}_i$ e o rótulo da classe

y_i. Para o modelo proposto, utiliza-se a função l na forma:

$$l = y_i(\hat{y_i}^{(t)} + f_t(x_i)) - log(1 + exp(\hat{y_i}^{(t)} + f_t(x_i))) \tag{4.2}$$

Dessa foma, a medida utilizada em cada avaliação do classificador *XGBoost* é a área sobre a curva ROC (AUC), discutida no capítulo anterior. Para concluir, pode-se definir o custo computacional da etapa. O número de avaliações é dada por $O\left(\frac{k!}{r!(k-r)!}\right)$, onde r é a ordem da interação de SNPs e k é o número de subgrupos. Dessa forma, é necessária a criação de um classificador XGBoost para cada combinação que pode conter os SNPs de interesse. Em contrapartida, para o método MDR que foi discutido no Capítulo 2, é necessário a criação de n^k classificadores. O número total de combinações demonstra que houve uma redução do espaço de busca original, onde os marcadores de interesse estão presentes nesse espaço reduzido representado pela combinação de subconjuntos.

4.2 Ranqueamento

Uma tarefa importante em diversos problemas é a avaliação do potencial de predição de uma variável sobre um conjunto de variáveis preditoras. Em muitas situações, o objetivo não é somente tornar o processo de predição mais acurado através dessa avaliação mas também identificar quais variáveis preditoras são mais importantes para o processo de predição. Nesse contexto em que o conjunto de dados original foi reduzido para uma combinação de subgrupos selecionados através da etapa anterior, o objetivo agora é determinar os marcadores mais importantes para o processo de predição presentes nesse conjunto restrito de marcadores.

Para a etapa de ranqueamento, os marcadores dos subgrupos selecionados na etapa anterior são ordenados utilizando uma medida de importância de variável do algoritmo da RF. Frequentemente para essa tarefa, deve se otimizar os parâmetros da RF, como número de árvores na florestas e número de variáveis para a escolha dos nós em cada árvore. Assim, a etapa de otimização de parâmetros acarreta em um grande consumo de tempo e complexidade computacional. Entretanto, para vários tamanhos de conjuntos de marcadores, essa tarefa não significa que o ranqueamento correto para os marcadores SNPs possa sempre ser obtido. Dessa maneira, definiu-se um procedimento baseado na fixação dos valores para estes parâmetros testando empiricamente nos experimentos se os valores

utilizados são satisfatórios e suficientes para o correto ranqueamento dos marcadores. Então, para o número total de árvores na floresta foi utilizado o número de indivíduos presentes no subconjunto de GWAS selecionado na etapa anterior e para o número de variáveis utilizados para quebra das árvores foi determinado o número de marcadores presentes no subgrupo.

A métrica de importância de variável definida como padrão de utilização do modelo foi a importância de permutação (pVI), que representa o aumento no erro de classificação para a j-ésima amostra OOB , depois da variável j ser permutada na árvore k. O ranqueamento é armazenado pelo modelo sendo utilizado como processo de inicialização da população inicial do algoritmo de programação genética. O processo é discutido na sessão 4.3.4. Para a construção da floresta randômica e cálculo das medidas de ranqueamento foi utilizado o pacote do R denominado de *randomForest* em conjunto com a função *importance* (LIAW; WIENER, 2002a).

4.3 Geração das soluções e interpretabilidade - Programação genética

Nesta seção é discutida a última etapa do modelo. Essa etapa constitui-se de um modelo de programação genética que utiliza conhecimento especialista ou inicialização da população inicial baseada no processo da etapa anterior. Ao final do processo são geradas soluções que podem ser facilmente interpretadas. São discutidas nas seções à seguir as estruturas dos indivíduos, função de avaliação, operadores genéticos e os parâmetros do modelo.

4.3.1 Estrutura dos indivíduos da PG

A estrutura dos indivíduos foi baseada em (NUNKESSER et al., 2007), onde são utilizadas expressões lógicas multi-valoradas na forma normal disjuntiva (do inglês, *disjunctive normal form*-DNF). Uma expressão lógica em DNF é uma disjunção de um ou mais monômios, onde um monômio é constituído de um único literal ou um conjunto dos mesmos. A Figura 4.3 apresenta uma árvore genérica com expressões lógicas em DNF representando um indivíduo da GP. A gramática utilizada é simples e o conjunto de funções é dado pelas expressões E e OU, sendo o conjunto de terminais composto pelos marcadores SNPs com

seus respectivos alelos, por exemplo $SNP1 = 0$.

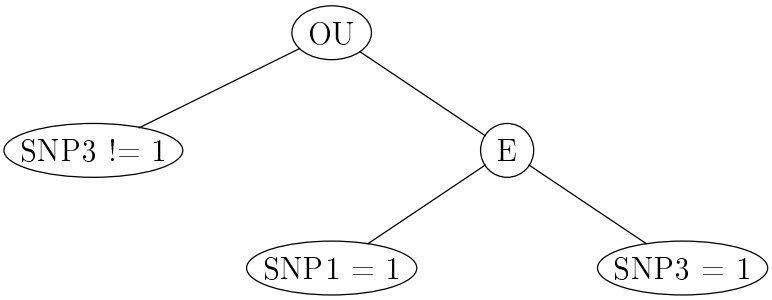


Figura 4.3: Exemplo de um indivíduo que representa o seguinte programa (SNP3 != 1 Ou (SNP1 = 1 E SNP3 = 1)).

O resultado da análise de cada indivíduo pode ser representada pelas seguintes expressões lógicas definidas pela Equação 4.3 e Equação 4.4.

$$R_1 = ((SNP_0 = 0))\ \text{E}\ ((SNP_2 = 1)\ \text{OU}\ (SNP_2 \neq 1)) \tag{4.3}$$

$$R_2 = ((SNP_3 \neq 1))\ \text{OU}\ ((SNP_1 = 1)\ \text{E}\ (SNP_2 = 0)) \tag{4.4}$$

Dessa forma, um indivíduo será interpretado como pertencente ao grupo "caso" se R_1 ou R_2 forem verdadeiras se por exemplo, todos os SNPs em que pelo menos um dos dois monômios demonstram o genótipo indicado pelo literal correspondente. Caso contrário, o indivíduo será dado como pertencente ao grupo "controle".

4.3.2 Avaliação e seleção dos indivíduos

Para determinar os indivíduos que irão compor a nova geração da população, os indivíduos da geração atual precisam ser avaliados. Para esta finalidade, a função de avaliação têm dois objetivos: definir as expressões DNF que melhor expliquem os indivíduos do grupo de caso com enfoque em maximizar a classificação correta dos mesmos. O segundo objetivo é minimizar o tamanho da expressão com a finalidade de determinar a regra mais compacta que melhor explica o fenótipo dos indivíduos do grupo caso.

Para avaliar cada indivíduo de uma geração da PG, algumas funções de avaliação foram definidas. A primeira é denominada de Precisão (do inglês, *Precision*). Essa métrica calcula quais amostras são efetivamente corretas dentre as amostras que foram classificadas

com sendo corretas. A expressão é dada por:

$$\text{Precisão} = \frac{VP}{VP + VN} \tag{4.5}$$

Outra medida avaliada foi a de Revocação (do inglês, *recall*) e F1 *score* respectivamente. A revocação avalia a frequência que as amostras de uma classe são realmente classificadas como pertencentes a ela. A F1 *score* combina precisão e revocação para trazer um único índice capaz de indicar a qualidade geral do modelo.

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{4.6}$$

$$\text{F1} = 2 * \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} \times \text{Revocação}} \tag{4.7}$$

Na composição da função objetivo, além das métricas definidas, foi adicionado o termo $\frac{N_i}{\alpha}$, definido a seguir. As métricas foram avaliadas em relação as regras geradas. Foram utilizados os conjunto de dados dos exemplos anteriores. Os resultados podem ser encontrados no Apêndice A. Apesar das métricas utilizadas serem bem estabelecidas na literatura, a função objetivo do modelo que gerou as regras com maior nível de interpretabilidade é dada pela seguinte Equação 4.8:

$$f_i = \frac{T}{(VP + VN)} + \frac{N_i}{\alpha}, \tag{4.8}$$

onde i é o índice do indivíduo, T representa o número total de indivíduos caso-controle (total de indivíduos na população), VP é número de verdadeiros positivos, VN a quantidade de verdadeiros negativos classificados corretamente. N_i representa o número de nós de um indivíduo da PG e α é uma constante de parcimônia (introduzido em (BLEULER et al., 2001)).

A função de objetivo apresentada pela Equação 4.8 foi baseada em (NUNKESSER et al., 2007) e (SZE-TO et al., 2013), que são dadas respectivamente por:

$$f_i = E_i + \frac{N_i}{\alpha}, \tag{4.9}$$

$$f_i = E_i + \frac{N_i}{\alpha} + P_i, \tag{4.10}$$

onde E_i representa o erro de classificação e P_i é um termo que penaliza indivíduos semelhantes na população. No modelo proposto não foi considerado um termo de penalização e o erro de classificação de ambas as classes, visto que tem-se o interesse somente na identificação da regra DNF que melhor explica o fenótipo dos indivíduos do grupo caso. Essa modificação na função de avaliação ajuda a reduzir o número de monômios falsos positivos na explicação da regra, em contra partida, o número de indivíduos do grupo de interesse deve ser de pelo menos metade da quantidade da base de dados para que o modelo consiga predizer a regra com maior precisão.

Depois de avaliados, os indivíduos são selecionados via método de torneio para se reproduzirem. Definiu-se 7 indivíduos da população são selecionados aleatoriamente pelo método de torneio (KOZA, 1994), sendo realizada uma comparação entre o valor da aptidão de cada um e o melhor é selecionado para a geração de proles, conforme pode ser visto na próxima seção. Dois torneios são necessários para a realização da recombinação para a seleção dos dois pais necessários. Caso contrário, o indivíduo não sofre o processo de cruzamento e é diretamente copiado para a próxima geração.

4.3.3 Operadores de manipulação de indivíduos

Para a criação de uma nova geração, os operadores genéticos de cruzamento e mutação, apresentados no capítulo 3 são utilizados no modelo. A Figura 4.4 mostra novamente os dois indivíduos exemplificados anteriormente, supondo que ambos são selecionados para realizarem o cruzamento.

Um nó é selecionado aleatoriamente (indicado em serrilhado na Figura 4.4) em cada um dos dois indivíduos. Com esse ponto de referência, duas proles são geradas a partir da combinação de ambas as árvores. A Prole 1 é uma combinação do primeiro indivíduo até o ponto de corte com o segundo indivíduo depois do seu ponto de corte, respectivamente. A Prole 2 é gerada a partir da junção de forma inversa ao da Prole 1.

Para cada indivíduo, uma probabilidade é definida para a aplicação do operador de mutação. De acordo com (POLI; LANGDON; MCPHEE, 2008; NUNKESSER et al., 2007), os possíveis resultados do processo de mutação são:

- Alteração de um terminal: um terminal é simplesmente trocado por outro;
- Exclusão de um terminal: um terminal é excluído, dando origem a uma expressão

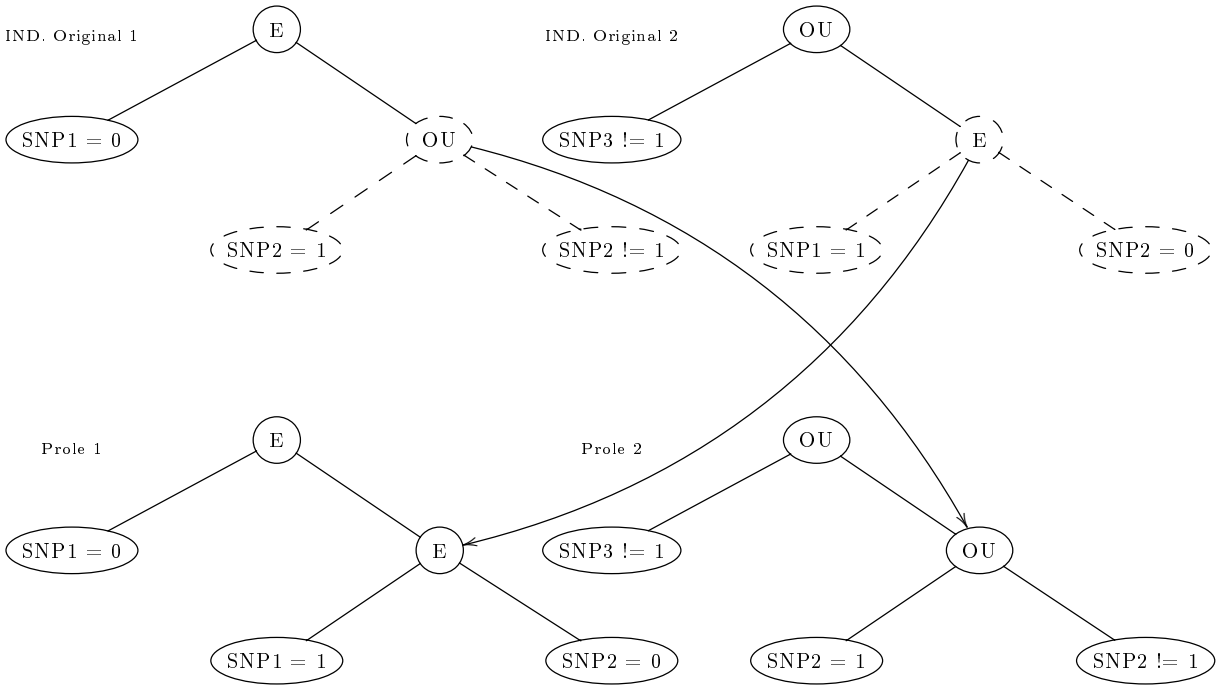


Figura 4.4: Exemplo de recombinação utilizada pelo modelo.

diferente da anterior;

- Alteração de um nó função por terminal: este processo realiza a poda indireta da árvore, provocando a terminação dos galhos da árvore a partir do ponto alterado;
- Exclusão de um nó função: dá origem a uma expressão mais compacta, eliminando o segmento da expressão afetado pela remoção do nó função.

4.3.4 Geração da população inicial

O mecanismo de geração da população inicial foi baseado no ranqueamento discutido previamente. O processo gerou uma lista de marcadores de acordo com a medida pVI. Através do processo de ranqueamento cada nó terminal de cada indivíduo da população é submetido a um processo de torneio com tamanho definido de 7, onde um marcador é selecionado dentre os 7 obtidos pela seleção de torneio aplicada nos marcadores que compõem a população. É realizada uma comparação entre o valor atribuído a cada marcador pela floresta randômica, sendo o que apresentar o maior valor é selecionado para gerar o nó terminal. Cada indivíduo só pode possuir uma cópia de um marcador, dessa forma, se um terminal apresenta um determinado SNP, ele não pode mais aparecer

na árvore de solução e outro torneio é realizado até que um SNP não inserido previamente seja encontrado. A Figura 4.5 montra o fluxograma do processo. Note que a execução da floresta randômica é um passo prévio discutido anteriormente e a realização do torneio é feita baseada no valor atribuído a cada marcador de acordo com a medida pVI. Para cada nó selecionado, há um probabilidade de 50% de apresentar o sinal de igual ou diferente e um terço de chance de conter cada um dos alelos (0, 1 ou 2). Dessa forma, nó terminais resultantes do processo poderiam se apresentar como: '$SNP1 = 2$ ou $SNP1! = 0$'.

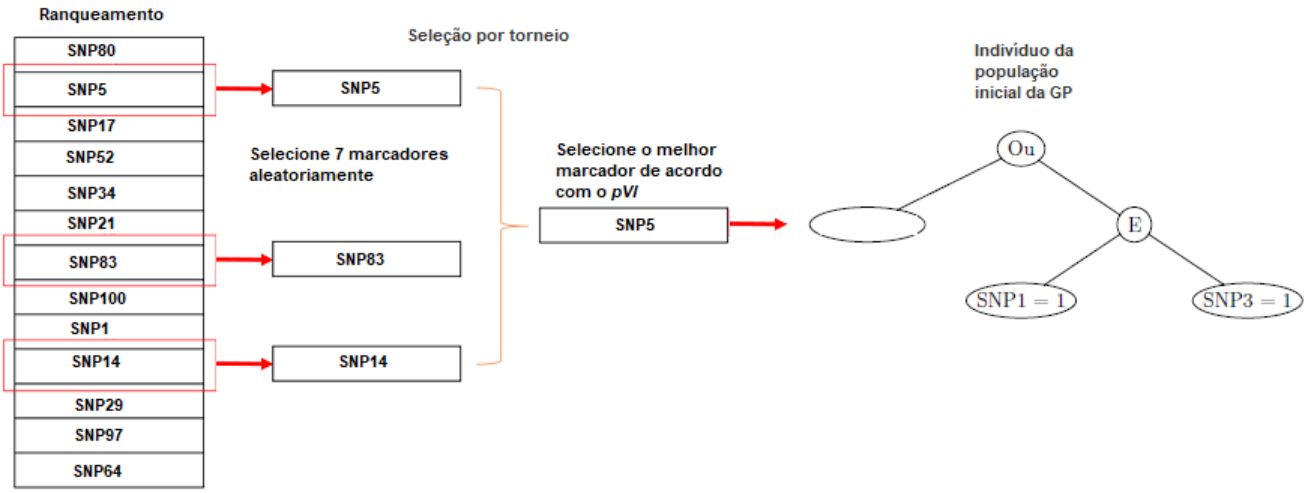


Figura 4.5: Fluxograma do modelo de geração da população inicial proposto baseado no ranqueamento da floresta randômica.

4.3.5 Critérios de terminação

Foram definidos três critério de terminação do algoritmo. O primeiro é se uma solução satisfatória dada pela aptidão de um indivíduo for encontrada. O segundo é se não houver melhora das soluções ao longo de 10 gerações. O terceiro é definido pelo número máximo de gerações de que o algoritmo irá executar. Os critérios de terminação foram baseados em (SZE-TO et al., 2013).

4.4 Configuração de parâmetros

A Tabela 4.1 sumariza a configuração dos parâmetros utilizados pelos algoritmos que compõem o modelo proposto neste trabalho. Os parâmetros tamanho da população, probabilidade de cruzamento, probabilidade de mutação e número de gerações foram

baseadas nos experimentos realizados em (SZE-TO et al., 2013) e o conjunto de funções e terminais adotados de (NUNKESSER et al., 2007). A função de avaliação utilizada foi definida na sessão 4.4.2. As árvores do algoritmo utilizam o método *grow*, descrito no capítulo anterior e a profundidade máxima de cada uma é 6.

Parâmetro	Valor
Tamanho da população	4096
Gerações	50
Cruzamento	Cruzamento de um ponto
Frequência de cruzamento	0.9
Mutação	inserção, deleção (50% de chance para cada)
Frequência de mutação	0.05
Função de aptidão	$f_i = \frac{T}{(VP+VN)} + \frac{N_i}{\alpha}$ com $\alpha = 10$
Método de seleção	Torneio
Conjunto de funções	E, OU
Conjunto de terminais	marcadores e alelos
Profundidade máxima das soluções	6

Tabela 4.1: Configuração dos parâmetros do algoritmo de GP utilizados nos experimentos.

4.5 Algoritmo e implementação

O algoritmo de GP foi desenvolvido em JAVA e para a sua implementação foi utilizado um pacote de computação evolucionista chamado ECJ (LUKE et al., 2007). A etapa de seleção de subgrupos e de inicialização da população inicial foram desenvolvidas em *R* (Team. R.C.R,, 2008) e tiveram como base o algoritmo de *XGBoost* e de floresta randômica do pacote randomForest (LIAW; WIENER, 2002b). Para a etapa de seleção de subgrupos foi desenvolvido um *script* para paralelizar o processo de classificação do *XGBoost*. As permutações são igualmente divididas e enviadas para cada *thread* disponível pelos processadores. Para avaliar as árvores pela função de aptidão, foi criado uma tabela em MySQL para cada base de dados. Foram desenvolvidos *scripts* em *Shell script* para automatizar o processo de criação das tabelas a partir dos arquivos de dados gerados pelos simuladores. Finalmente, o modelo proposto foi nomeado de xGPi para a comparação com os outros métodos nos experimentos, sendo xGPi a abreviação de *XGBoost Genetic Programming with initialization*. O Algoritmo 2 descreve o processo desenvolvido em todas as etapas do modelo proposto.

Algoritmo 3: Modelo Proposto

Entrada: Conjunto de dados de GWAS
t : tamanho do subgrupo (quantidade de SNPs em cada subgrupo)
c : número de combinações
cv : quantidade de partes da validação cruzada do *XGBoost*
it : número de iterações do *XGBoost*
Saída: Regra de associação gerada pela GP

início
  Selecione um conjunto de dados de GWAS
  divida o conjunto de dados de GWAS em subgrupos de t tamanhos iguais
  calcule as c combinações possíveis de todos os subgrupos
  para *cada* i *até* z **faça**
    execute o algoritmo xgboost(cv,it) para a combinação c_i
    armazene o resultado da AUC de cada z_i
  fim
  selecione as c combinações de acordo com a área sobre a curva ROC
  elimine as redundâncias de SNPs nas c combinações
  execute o algoritmo de floresta randômica e calcule a importância de cada variável SNP nas c combinações utilizando métrica pVI
  crie uma lista ordenada dos SNPs de acordo com valor de pVI
  crie a população inicial da GP utilizando a lista
  para cada nó de cada árvore da GP, execute o método de torneio e selecione um marcador da lista sem repetição
  para *cada população* j *até* n **faça**
    avalie os indivíduos da população j
    selecione os melhores indivíduos da população j
    aplique os operadores genéticos insira os novos indivíduos na população $j = j + 1$
  fim
fim

5 Experimentos preliminares

5.1 Introdução

Esse capítulo apresenta experimentos preliminares para a justificativa da construção das etapas 1 e 2 do modelo proposto no presente trabalho. Portanto, são discutidos nesse capítulo suas respectivas importâncias para a proposta de solução do problema, bem como os principais obstáculos encontrados durante o desenvolvimento. Para a discussão de cada componente e para detalhar o seu processo, o capítulo apresenta exemplos ilustrativos usando bancos de dados simulados como os que são discutidos no capítulo de Resultados.

5.2 Experimentos para a seleção de subconjuntos

A ideia inicial foi utilizar o algoritmo *XGBoost* para identificar os atributos mais relevantes para o fenótipo. Um dos grandes benefícios de se utilizar algoritmos de gradiente *boosting* é que depois que as árvores são construídas, é relativamente simples a recuperação dos valores de importância para cada variável, bastando retornar aos pontos de quebra da árvore. Geralmente, esse valor de importância indica o quanto a variável foi útil para que a árvore gerasse a quebra naquele nó, assim, quanto mais alto o valor, mais importante é a variável para o modelo. A importância é calculada para uma única árvore de decisão pela quantidade de pontos de quebra que cada atributo melhorou a medida de desempenho, ponderada pelo número de observações obtidas pelo nó responsável. A medida de performance pode ser o índice Gini ou uma função de erro específica. Ressalta-se que a medida de importância gerada pelo *XGBoost* é a média de todas as árvores de decisão presentes no modelo (HASTIE; TIBSHIRANI; FRIEDMAN, 2016).

Para ilustrar a ideia, será gerado um conjunto de dados sintético com dois mil indivíduos, com dados balanceados entre o grupo de casos e controle. A herdabilidade do conjunto de dados é de 0.4 e a MAF de 0.2 e mil marcadores estão disponíveis. Uma interação epistática foi simulada entre os marcadores $SNP999$ e $SNP1000$. Para a geração dos dados foi utilizado o simulador GAMETES (URBANOWICZ et al., 2012). Esse cenário foi utilizado porque representa um dos cenários mais simples abordados nesta

tese. Assim, pode-se utilizá-lo de forma ilustrativa para o entendimento do processo de composição dessa etapa do modelo.

O algoritmo *XGBoost* foi executado sobre esse conjunto de dados com os seguintes parâmetros: função objetivo binária para classificação, validação cruzada de dez partes e profundidade máxima das árvores com nível seis, esses parâmetros foram definidos através da utilização de uma *grid-search*. Dessa forma, através do algoritmo, um classificador "fraco" foi inicialmente criado e foram definidos mil passos para refinamento do modelo. Assim, a cada passo do algoritmo um outro classificador "fraco" é adicionado para aumentar o desempenho e construir um classificador mais "forte". A medida em que o algoritmo é iterativamente executado, ele minimiza a função de perda baseada em regressão logística. Esse perda é dada pelo erro residual, ou seja, a diferença entre o valor do rótulo real do vetor e o seu valor predito. A cada etapa, o algoritmo utiliza-se da predição anterior para atualizar a minimização dos resíduos (CHEN; GUESTRIN, 2016). A medida de importância dos atributos foi calculada para este conjunto de dados e pode ser observada no gráfico da Figura 5.1.

Pode-se notar que nenhum dos dois marcadores de interesse estão presentes do grupo das variáveis que foram melhores avaliadas pela medida de importância do modelo. O marcador $SNP999$ ficou na posição 30 e o marcador $SNP1000$ na posição 484. Isso indica que a relação de epistasia não foi bem capturada pelo modelo. Entretanto, uma avaliação das variáveis melhor selecionadas pode ajudar a entender o comportamento da ferramenta. Observando o valor-p em relação ao fenótipo, que pode-se definir como a menor escolha que teríamos feito para o nível de significância, de forma que rejeitaríamos H_0. Pode-se notar que dentre as variáveis melhores ranqueadas, algumas possuem um valor-p muito baixo, como observado no gráfico de Manhattan presente na Figura 5.2. Isso significa que esses marcadores podem ter um efeito marginal grande sobre o problema, o que indica que efeitos isolados ou aditivos podem ser capturados por essa medida de avaliação. Na Figura 4.2, os marcadores melhores ranqueados pelo modelo estão mostrados na cor verde, onde pode-se observar que o marcador melhor qualificado, o $SNP837$ é o que possui o menor valor-p dentro do conjunto. Entretanto, para o objetivo do estudo, que é o de investigar interações epistáticas, essa medida mostrou-se não ser a mais adequada.

Uma outra alternativa é utilizar as caraterísticas do *XGBoost* para classificação. Em (SOHN; OLSON; MOORE, 2017) o algoritmo foi comparado com regressão logística e

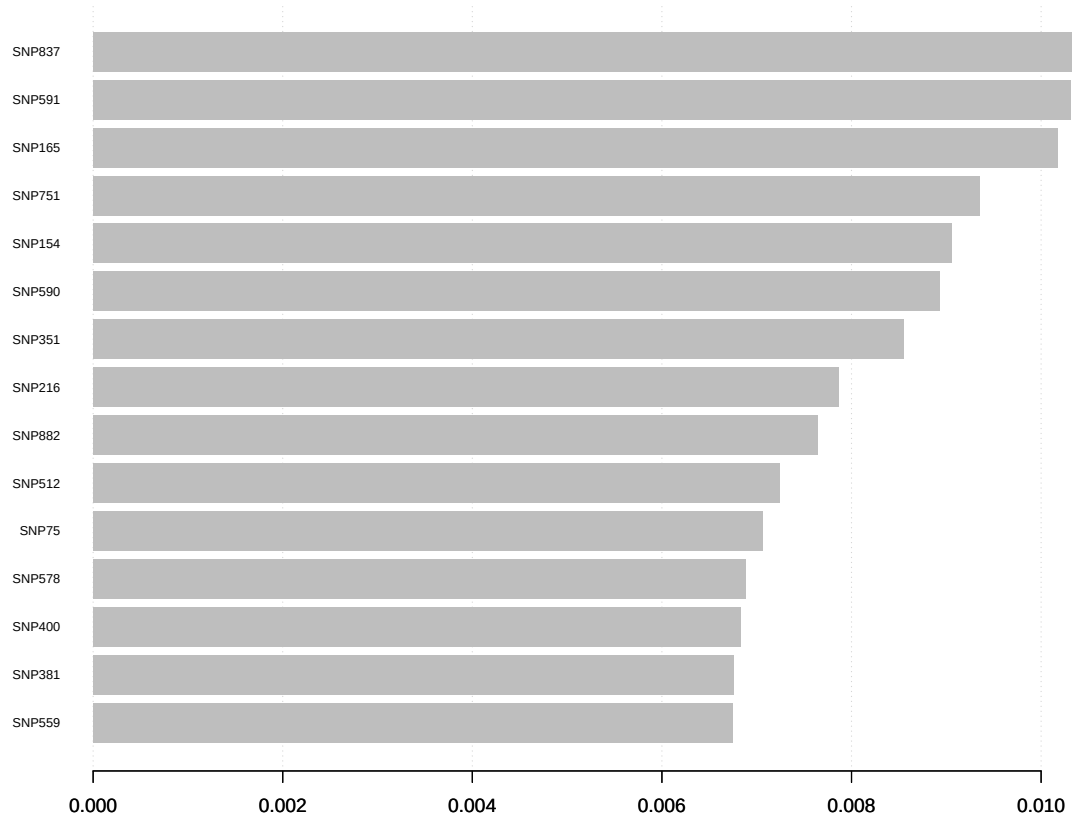


Figura 5.1: Gráfico com a medição de valores de importância das variáveis de um conjunto de dados sintético dado pelo *XGBoost* com mil marcadores e herdabilidade de 0.4. Os marcadores de interesse que simulam a interação epistática são o $SNP999$ e $SNP1000$. Como pode-se observar, ambos não estão presentes no grupo com os 15 melhores valores de importância.

com versões do MDR para predição. A Figura 4.3 demonstra a comparação dos resultados obtidos pelo trabalho. Para a comparação dos algoritmos, foi adotado o mesmo simulador do modelo utilizado como exemplo nesta seção, o GAMETES. Cada gráfico mostra a distribuição da médias das acurácias sobre uma validação cruzada de 10 partes para cada experimento, utilizando intervalo de confiança de 95%. De acordo com (SOHN; OLSON; MOORE, 2017), uma acurácia de 50% é equivalente á uma predição aleatória. O experimento foi realizado com diferentes números de marcadores, com quantidades variando entre 10 e 5000, onde cada painel representa níveis de herdabilidade diferentes (ruídos nos dados), com valores entre $0.05 - 0.4$.

O leitor atento pode notar que o *XGBoost* teve resultados relevantes somente em alguns cenários, principalmente quando o conjunto de marcadores é pequeno. Quando

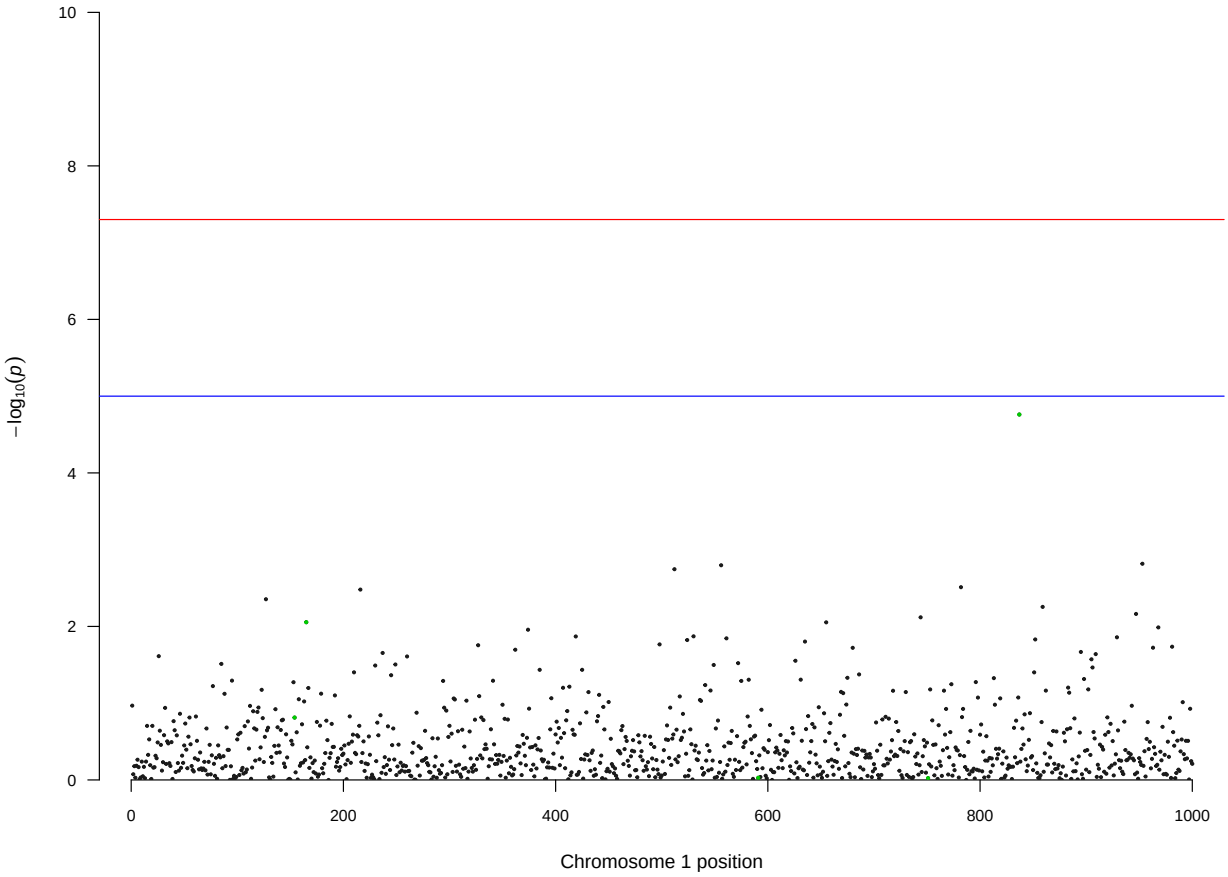


Figura 5.2: Gráfico de Manhattan com os valores-p de cada marcador no conjunto de dados. Os marcadores em verde são os melhores ranqueados pelo *XGboost*. Pode-se observar que variável escolhida como a mais importante pelo modelo, o $SNP837$ foi a que obteve o menor valor-p em relação ao fenótipo.

têm-se apenas 10 marcadores, o algoritmo teve quase 70% de acurácia em quase todos os cenários de herdabilidade, exceto com herdabilidade 0.05, onde o modelo apresentou cerca de 60% de acurácia, quando o número de marcadores é igual a 100. Pode-se notar que ainda assim, o algoritmo ficou pouco acima dos 50% de acurácia. Nos demais cenários, o modelo não obteve bons resultados em comparação com os outros algoritmos apresentados. Ainda no trabalho (SOHN; OLSON; MOORE, 2017), os autores mencionam que o *XGBoost* pode algumas vezes encontrar um bom modelo se os dados forem fortemente filtrados previamente, como por exemplo, casos em que se tem 10 marcadores, mas que seu desempenho decai drasticamente na medida em que ruídos são adicionados aos conjuntos de dados. Entretanto, os autores não mencionam os parâmetros utilizados no experimento, o que dificulta a replicação dos mesmos. Têm-se somente as informações

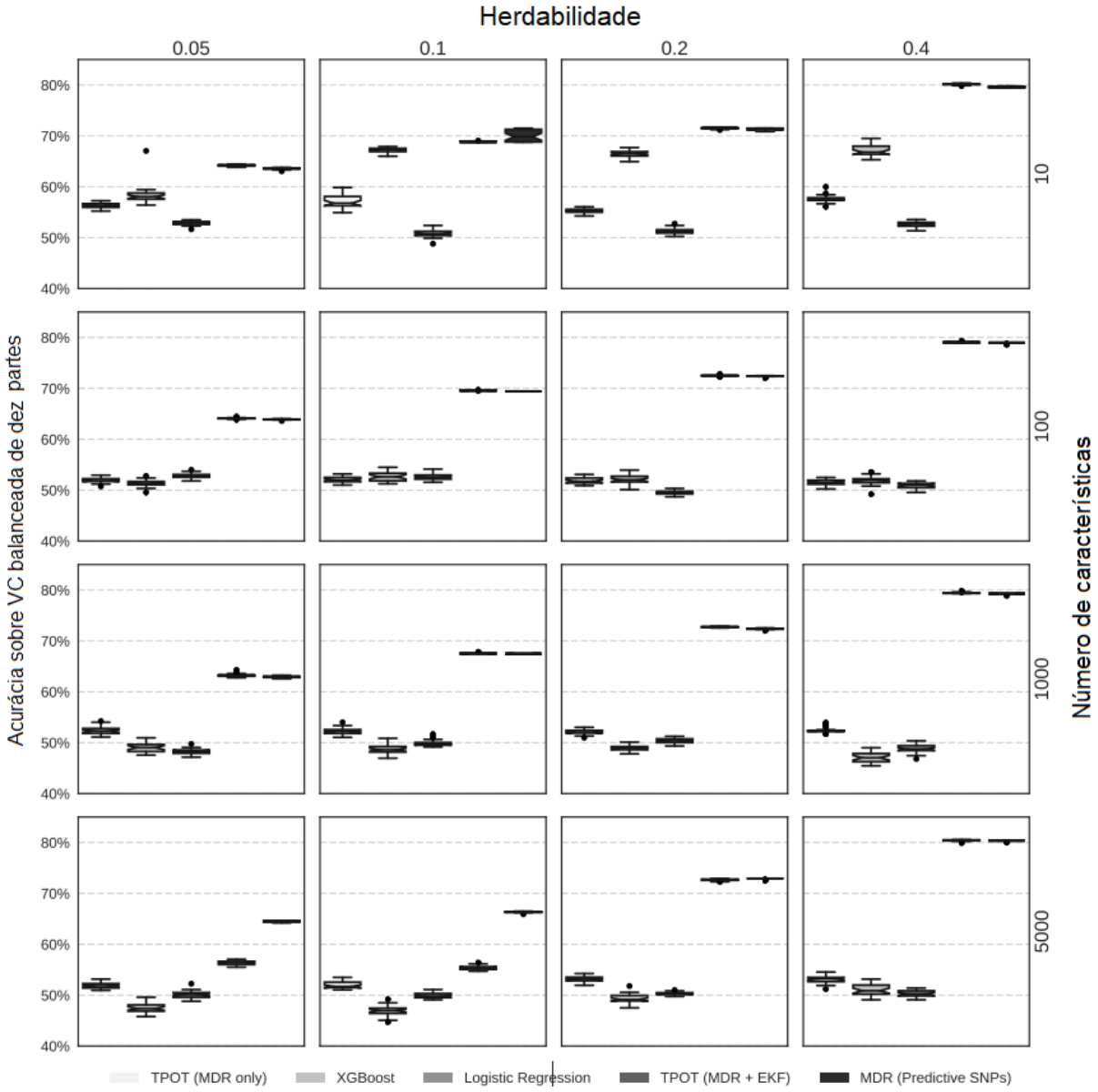


Figura 5.3: Figura extraída de (SOHN; OLSON; MOORE, 2017). Comparação dos algoritmos TPOT, *XGBoost*, Regressão logística, TPOT (MDR+EKF) e MDR preditivo. Os dados foram simulados com o GAMETES. Cada gráfico mostra distribuição da médias das acurácias sobre uma validação cruzada de 10 partes para cada experimento, utilizando intervalo de confiança de 95%. Os dados do topo da direita representam os cenários de classificação mais "fáceis", em contrapartida, os do canto inferior esquerdo são os considerados mais "difíceis".

sobre o número de atributos e herdabilidade utilizados pelo conjunto de dados simulado.

Pode-se observar que o experimento somente testou os algoritmos para classificação de indivíduos, nada foi mencionado sobre busca de interações ou de marcadores isolados. Dessa forma, apesar do algoritmo não lidar bem em problemas de classificação com um grande número de marcadores, segundo o trabalho, ele ainda é eficiente e rápido para a criação de modelos. Assim, uma ideia foi criar subgrupos de marcadores e classificá-

los separadamente. Voltando ao exemplo do conjunto de dados simulado anteriormente, onde o *XGboost* foi utilizado para classificá-lo. O conjunto foi divido em várias partes com números similares de marcadores e foi utilizado classificação binária. Foram executadas 10 iterações para cada subgrupo e a medida de avaliação da área abaixo da curva ROC (AUC) foi selecionada. Primeiramente um subconjunto contendo apenas 10 marcadores foi selecionado, dentre os 10, estão presentes os dois marcadores de interesse $SNP999$ e $SNP1000$. Então, foram classificados outros subgrupos maiores de atributos sempre contendo os marcadores de interesse em cada um deles. Assim, pela Figura 5.4, pode-se notar que quando o número de marcadores aumenta no subgrupo, a AUC diminui, sendo que, até aproximadamente 200 marcadores, a AUC fica acima dos 50%, semelhantes aos resultados discutidos em (SOHN; OLSON; MOORE, 2017).

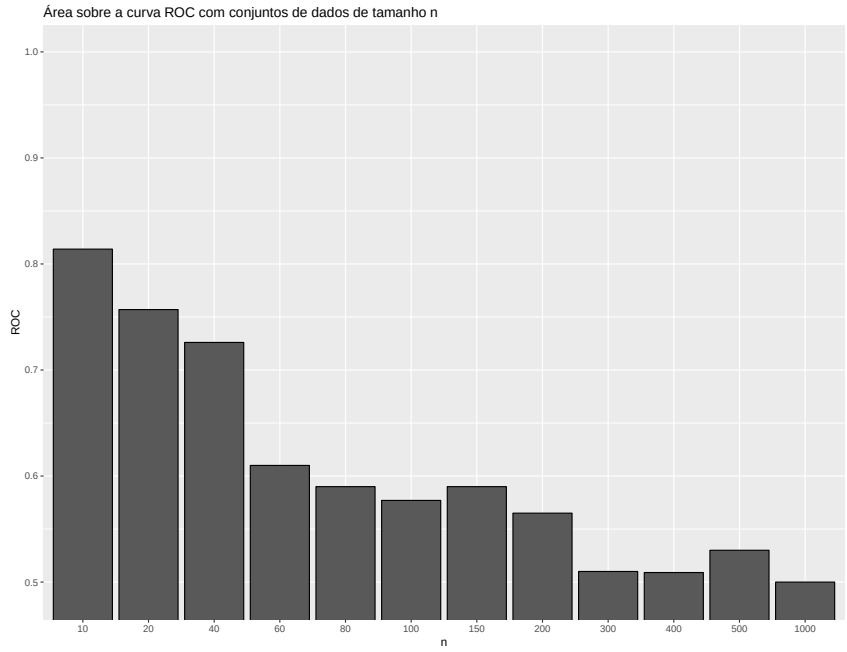


Figura 5.4: Gráfico contendo subconjuntos de tamanhos variados. Cada subconjunto contém os dois marcadores de interesse, no caso $SNP999$ e $SNP1000$. Pode-se notar que à medida em que o número de marcadores aumenta nos subconjuntos, menor é AUC.

O mesmo foi feito com subgrupos que não contém os marcadores de interesse e os seus resultados podem ser vistos na Figura 5.5. Os resultados indicam que quando os marcadores de interesse não estão presentes no subgrupo, o classificador não consegue uma boa AUC, o que implica que o mesmo não consegue modelar o problema corretamente.

Seguindo essa proposta, pode-se observar que quando os marcadores de interesse estão presentes nos subconjuntos, a área sobre a curva ROC é evidenciada. O objetivo então nessa primeira etapa é reduzir o espaço de busca original dividindo o conjunto de dados

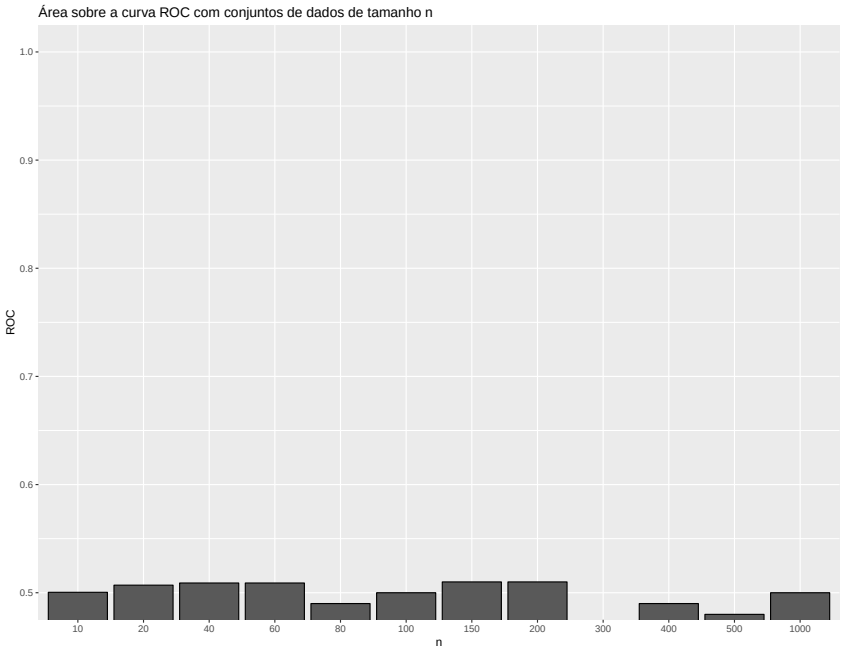


Figura 5.5: Gráfico contendo subconjuntos de tamanhos variados. Neste cenário, não existe a presença de marcadores de interesse nos subconjuntos. Pode-se notar que o número de marcadores não têm influência sobre o resultado.

em subgrupos menores para então classificá-los com o *XGBoost*. A ideia é encontrar subconjuntos com as melhores áreas sobre as curvas ROC e selecioná-los para a próxima etapa. A subseção à seguir demonstra o processo de criação e seleção de subgrupos proposto no trabalho.

Retornando ao exemplo utilizado no inicio da seção, a metodologia discutida foi aplicada sobre o conjunto de dados definindo subconjuntos de tamanho 10. Assim, foram criados 100 subgrupos, alocando todos os mil marcadores do conjunto de dados original. Antes de serem alocados, o conjunto original teve a suas colunas aleatoriamente reposicionadas, para evitar que os marcadores de interesse $SNP999$ e $SNP1000$ permanecessem alocados no mesmo subgrupo no inicio do processo. Um total de 4950 combinações foram geradas, cada uma com 20 marcadores, foram classificadas e no final das avaliações a combinação de subgrupos c_i foi apresentada, os resultados são apresentados na Tabela 5.1.

SNPs da permutação c_i
SNP853, SNP3, SNP675, SNP587, SNP954, SNP612, SNP999, SNP698, SNP951, SNP76, SNP572, SNP966, SNP709, SNP254, SNP247, SNP931, SNP470, SNP282, SNP1000, SNP577

Tabela 5.1: SNPs presentes no subgrupo c_i melhores avaliados pelo algoritmo *XGBoost*.

Pode-se observar que os marcadores de interesse estão na posição 7 e 19, respectivamente no conjunto c_i composto pela combinação de dois subgrupos g_i. Com esta parametrização para os subgrupos de marcadores e os respectivos conjuntos de combinações, conseguiu-se selecionar dos 1000 marcadores originais, apenas 20, ressaltando-se que os marcadores de interesse foram adequadamente selecionados.

O que demonstra que nesse exemplo, houve uma redução do espaço de busca original de 1000 marcadores para apenas 20, onde os marcadores de interesse estão presentes nesse espaço reduzido representado pela combinação de subconjuntos.

5.3 Experimentos com ranqueamento

Para calcular a importância das variáveis, foram utilizadas duas medidas, a pVI e a gVI apresentadas no capítulo anterior. Para exemplificar, foram considerados dois cenários distintos, um com 100 marcadores, $h^2 = 0.4$ e MAF de 0.4 e outro cenário também composto de 100 marcadores, $h^2 = 0.1$ e MAF de 0.2. As duas medidas de importância de variável foram calculadas para cada cenário e os resultados são apresentados na Figura 5.6 e Figura 5.7.

Os marcadores de interesse nesses cenários são "M0P1"e "M0P2", responsáveis pela interação epistática simulada. Pode-se notar que no primeiro cenário, ambas as medidas ranquearam corretamente os dois marcadores, que estão presentes no topo das listas. Já para o segundo cenário, que apresenta h^2 mais baixa, verifica-se que na medida pVI os marcadores de interesse ficaram na primeira e terceira posição do processo de ranqueamento. Já a medida gVI, que utiliza o índice Gini, não ranqueou corretamente os marcadores, sendo que o SNP "M0P1" ficou na vigésima quinta posição e o "M0P1" não ficou entre os melhores 30 marcadores ranqueados, resultado esperado já que pVI é calculada para as amostras OOB, sendo que se uma variável apresenta alguma importância para a classificação, a combinação realizada em pVI pode aumentar ou diminuir o erro de classificação. Como os marcadores de interesse estão presentes nos subgrupos gerados na etapa anterior que serão avaliados por essa medida, ela se torna mais eficiente em relação a gVI para ranqueamento nesses subgrupos de marcadores.

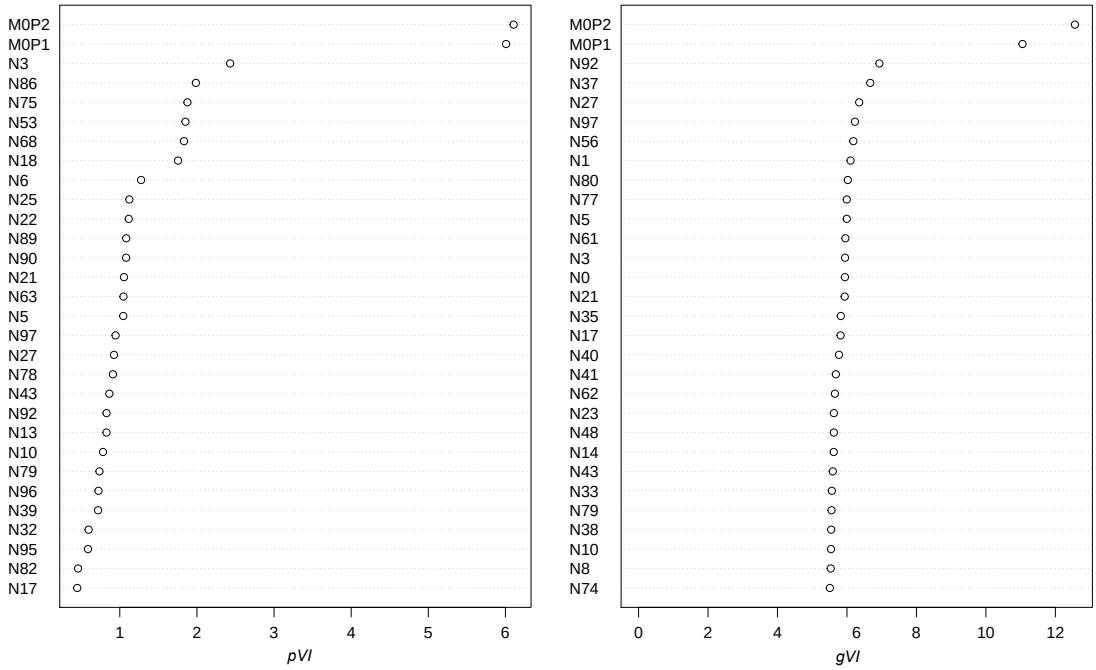


Figura 5.6: Ranqueamento realizado pelas medidas pVI (esquerda) e gVI (direita) em um conjunto de dados simulado com 100 marcadores $h^2 = 0.4$ e MAF de 0.4.

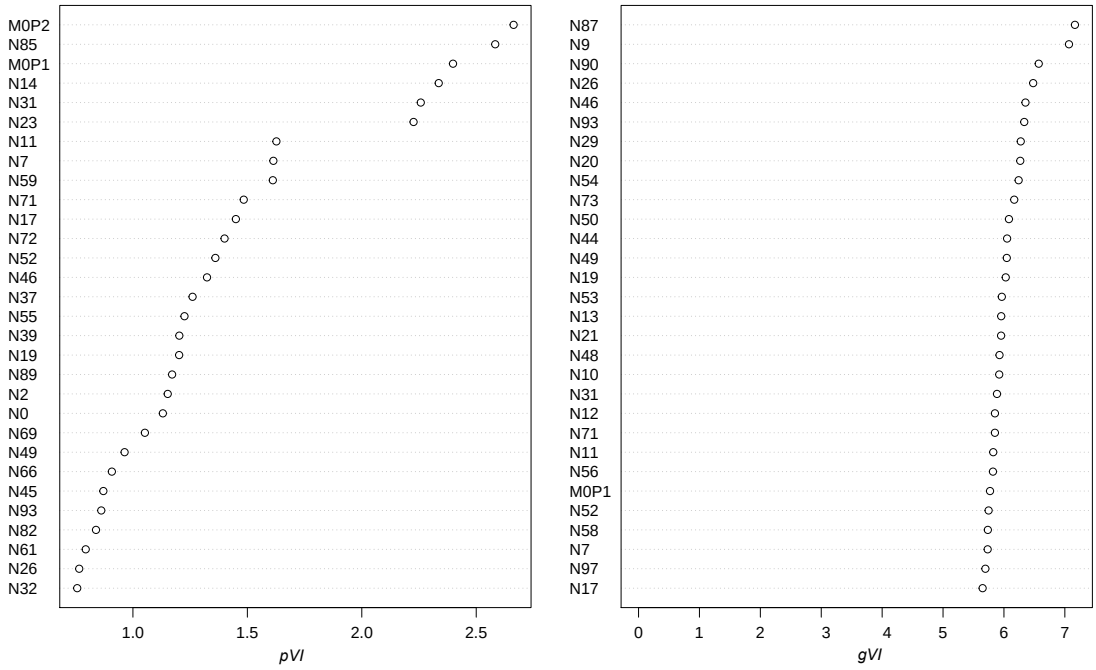


Figura 5.7: Ranqueamento realizado pelas medidas pVI (esquerda) e gVI (direita) em um conjunto de dados simulado com 100 marcadores $h^2 = 0.1$ e MAF de 0.2.

6 Experimentos computacionais

Neste capítulo, os experimentos computacionais e os seus respectivos resultados são apresentados. Para avaliar o desempenho do modelo proposto, os experimentos foram realizados sobre conjuntos de dados simulados. Infelizmente, existem poucas referencias na literatura sobre interações epistáticas que foram descobertas e replicadas, por esse motivo, utiliza-se de simulações para comparação e avaliação dos algoritmos de GWAS.

Foram definidos diversos cenários que incluem o tamanho do conjunto de dados, número de SNPs envolvidos na interação epistática, MAF e herdabilidade. O modelo proposto foi comparado com outros três algoritmos fundamentalmente conhecidos na literatura de GWAS. Dentre eles estão o GPAS, MDR e MDR + Relieff. Sendo o MDR o algoritmo referência e mais utilizado em análises desse tipo de estudo. Estes modelos foram selecionados pois geram regras de associação indicando os marcadores presentes nas interações e por serem métodos diretos, que utilizam todo o conjunto de dados para avaliação do modelo.

Os experimentos foram realizados em computadores diferentes, dentre eles: um computador com processador Intel® Core i7-4770K com 3.50GHz x 8 e 32GB de memória RAM DDR3, um computador Processador Intel® Core i5-8400 com 4.0GHz x 6 e 16GB de memória RAM DDR4, e um supercomputador AMD com 32 processadores e 64 *threads*. Este último computador mencionado é parte do núcleo de computação de alto desempenho do programa de pós-graduação em Modelagem Computacional da UFJF. Pelos experimentos terem sido executados em sistemas com diferentes configurações, os cálculos de *benchmark* ficaram imprecisos para fins de comparação.

6.1 Critérios para avaliação dos modelos

Para avaliação dos resultados, o modelo ideal deve selecionar somente os SNPs envolvidos nas interações epistáticas, ou seja, aquele que captura somente os SNPs verdadeiros positivos e elimina os SNPs falso-negativos. Entretanto, encontrar somente os SNPS que marcam regiões próximas aos genes de interesse que expressam um determinado fenótipo é um esforço árduo e complexo devido a inconsistências envolvidas em diferentes etapas

do estudo de GWAS. O que torna esse modelo ideal muito raro ou até mesmo impraticável de ser encontrado. Assim, é possível adotar critérios de avaliação mais adequados para esse tipo de estudo de associação em escala genômica.

Para isso, podemos definir outros critérios de escolha, como por exemplo, o melhor modelo como sendo aquele que consegue selecionar o maior número de SNPs verdadeiros-positivos, mesmo que para isso, o mesmo selecione mais SNPs falsos-positivos, OU aquele que identifica o maior número de SNPs causais envolvidos na interação com o menor número de SNPs não-causais.

Este último entendimento foi utilizado como método de avaliação dos algoritmos comparados neste trabalho. A opção por esse método é a existência de uma probabilidade maior de encontrar os genes que definem uma característica de interesse quando o maior número de SNPs causais é selecionado. Nas primeiras seções deste capitulo, o modelo desenvolvido foi comparado com outros três modelos presentes na literatura: GPAS, MDR e MDR + Relieff, os métodos foram apresentados no Capítulo 2. O MDR é um algoritmo largamente utilizado desde a sua primeira versão. Diversas modificações do MDR foram propostas, entre elas, o MDR + Relieff, que utiliza o algoritmo Relieff como filtro para selecionar um grupo de marcadores do conjunto de dados original para que o MDR seja então executado sobre esse grupo menor, reduzindo o seu custo computacional.

Para os primeiros experimentos, foram simulados conjuntos de dados com interação epistática entre dois SNPs funcionais, herdabilidade com valores de 0.4, 0.3, 0.2, 0.1. MAF variando de 0.2 e 0.4, número de marcadores variando de 10^2, 10^3 e 10^4. Cada conjunto de dados foi replicado 5 vezes. Assim, cada modelo foi executado 5×10 vezes com o número de vezes em que os SNPs funcionais encontrados foram contabilizados. Essa medida foi denominada de *"Power"*, ou seja, a quantidade de vezes em que cada algoritmo identifica o maior número de SNPs causais envolvidos na interação epistática, como foi definido para a medida de avaliação. Esse valor representa o poder preditivo dos modelos para estimar os fenótipos de interesse, ou seja, qual a frequência em que os modelos estão aptos a encontrarem a solução esperada. Para os modelos, considera-se como saída correta do xGPi, somente a última expressão, representando o melhor indivíduo obtido pelo processo evolutivo da execução da etapa de programação genética do modelo proposto. Para o GPAS, se os SNPs causais estão presentes nos melhores 5 indivíduos, a solução é considerada válida. Esse critério de avaliação foi adotado em

(SZE-TO et al., 2013) Para o MDR e MDR+ Relieff, a solução é válida se em alguma de suas saídas, os SNPs causais estão co-presentes.

6.2 Experimentos com dados simulados

Um modelo que apresenta uma interação epistática descreve esse efeito como sendo a combinação de genótipos (ou seja, a combinação de diferentes marcadores do tipo SNP) que influenciam no risco de uma determinada doença. Dessa forma, entre os estudos de GWAS, existe a suposição de que o risco de expressão de um fenótipo pode aumentar ou diminuir de acordo com a frequência dos alelos. Portanto, associações de doenças são frequentemente dadas em duas dimensões, a frequência do alelo determinada pela penetrância e a força do efeito que é estimado sobre a base da definição da *Odds Ratio* da doença.

Assim, um modelo epistático pode apresentar efeitos diferentes: efeito principal e efeito fraco. O primeiro é aquele que destaca SNPs com um efeito particular grande ou moderado sobre o fenótipo. O segundo descreve o marcador com pouco OU nenhum efeito individual, mas que quando atua junto com outros marcadores, gera grande influência no efeito da doença. Os conjuntos de dados foram simulados de forma aleatória, de acordo com os parâmetros de simulação definidos anteriormente, portanto, o termo epistasia utilizado aqui, pode ser interpretado como sendo de qualquer um dos tipos discutidos.

Nesta seção, serão apresentados os experimentos realizados sobre conjuntos de dados artificialmente gerados. Para a geração dos conjuntos de dados, foi utilizado o simulador GAMETES (URBANOWICZ et al., 2012). Como a maioria dos algoritmos de geração de conjunto de dados, o GAMETES leva em consideração o equilíbrio de Hardy-Weinberg (HWE). Dessa forma, as frequências alélicas de um SNPs pode ser utilizada para calcular a sua frequência genotípica na forma: freq(AA) = p^2, freq(Aa) = $2pq$ e freq(aa) = q^2, onde p é a frequência do alelo mais comum, ou seja, o homozigoto dominante representado por 'A', q é a frequência do menor alelo, ou seja, a MAF do menor alelo 'a', ou seja, o homozigoto recessivo, e $p+q = 1$. O GAMETES também assume que alelos em diferentes *loci* estão em desequilíbrio de ligação. Além disso, ele permite a geração de modelos de conjuntos de dados com níveis de herdabilidade diferentes. Assim, é possível gerar diversos tipos de funções de penetrância que definem a relação entre o genótipo e fenótipo. A

Tabela 6.1 exemplifica uma função de penetrância similar àquelas geradas pelo simulador GAMETES.

	AA (0.25)	Aa (0.50)	aa (0.25)
BB (0.25)	0.451	0.214	0.190
Bb (0.50)	0.192	0.164	0.065
bb (0.25)	0.139	0.350	0.463

Tabela 6.1: Exemplo de uma função de penetrância para um modelo que apresenta epistasia entre dois marcadores. Nesse exemplo, os marcadores envolvidos na interação são representados pelos marcadores A e B. A combinação de algumas de suas variações alélicas combinadas simulam o efeito de interação entre as mesmas.

A função de penetrância define as regras da interação epistática. Considerando a Tabela 6.1, têm-se AA=0, Aa = 1, e aa=2, analogamente o mesmo vale para BB,Ba,bb. O marcador A representa o primeiro SNP causal envolvido na interação e o marcador B o segundo. Na prática, nos dados simulados, os marcadores são codificados em números na forma: SNP1,SNP2,...,SNP$_n$ (ou N1,N2,...,NP$_n$) e a variação homozigoto dominante como 0, heterozigoto como 1 e homozigoto recessivo como 2. Assim, os resultados gerados pelos algoritmos apresentados terão formas de regras contendo os marcadores encontrados por cada modelo, assim como sua respectiva variação alélica.

Nas abordagens utilizadas, pode-se mencionar que o algoritmo do GPAS utiliza somente de duas combinações alélicas. Assim, o método difere apenas homozigotos de heterozigotos na geração de suas regras de associação. Para o MDR é necessário definir *à priori* o número de marcadores causais envolvidos na interação. Portanto, seu espaço de busca precisa ser necessariamente definido, o que pode reduzir o custo computacional na descoberta de tal associação. A metodologia aqui definida como MDR+ReliefF utiliza um filtro antes da execução do MDR. Esse filtro denominado de ReliefF, foi apresentado no Capítulo 2. Para um melhor entendimento, o algoritmo utiliza uma medida de cálculo de proximidade entre indivíduos para identificar similaridades genéticas. Logo após, é realizada uma avaliação da qualidade das variantes genéticas de acordo com o quanto elas podem ser úteis para distinguir indivíduos que estão próximos um dos outros (MOORE, 2014).

Os dados utilizados nos primeiros experimentos utilizam modelos epistáticos gerados aleatoriamente com efeito principal e sem efeito principal. Nas próximas seções os experimentos serão detalhados e os resultados discutidos.

6.2.1 Experimentos com interação epistática entre dois loci e 100 marcadores

Os resultados obtidos pelas quatro abordagens distintas nos conjuntos de dados com SNPs causais e um total de cem marcadores são demonstrados na Figura 6.1,Figura 6.2,Figura 6.3 e Figura 6.4. Cada barra nos gráficos representa uma metodologia, sendo o GPAS o primeiro a ser apresentado, seguido pelo MDR e MDR+ReliefF, e por fim, a metodologia proposta neste trabalho, denominada de xGPi.

Os resultados são apresentados de acordo com a maior herdabilidade até a menor herdabilidade utilizada. Em cada gráfico, os resultados referentes a MAF 0.2 pode ser encontrada à esquerda e a MAF 0.4 consequentemente do lado oposto.

Ao observar os resultados, conclui-se que dentre as quatro metodologias distintas, as que encontraram os SNPs causais em todas as execuções foram o algoritmo MDR e o xGPi. Como já era esperado, o MDR identificou corretamente os marcadores. O modelo testou cada par de interações dentro do espaço de busca de 100×100 marcadores e encontrou um modelo de associação em todas as 50 execuções, assim como o método proposto, que obteve sucesso ao encontrar os marcadores causais. Para o xGPi, os subgrupos foram divididos em tamanhos fixos de 10 marcadores, gerando um total de 10 subgrupos em cada execução do modelo.

Uma questão interessante observada é que à medida em que a herdabilidade e a MAF diminuem, o algoritmo GPAS oscila ao encontrar as associações, mas ainda assim, na maior parte das execuções, mesmo com a herdabilidade em 0.1 e MAF 0.2, o modelo acertou os SNPs causais em aproximadamente 60% das vezes.

A metodologia MDR+ReliefF obteve bons resultados, entretanto, para herdabilidade 0.1 e MAF 0.4, o algoritmo não obteve sucesso ao encontrar as associações em menos da metade das execuções. Uma questão surge a partir desse resultado, o MDR+ReliefF obteve mais sucesso com uma MAF menor do que com uma MAF maior, essa questão pode ser justificada da seguinte maneira. Para cada um dos cinco conjuntos de dados, a posição dos indivíduos se difere, ou seja, para cada uma das 10 execuções em cada conjunto de dados, a cada 10 de 50, a posição continua inalterada, modificando quando o outro conjunto de dados é apresentado ao modelo. Consequentemente, como o número de vizinhos próximos definidos pelo algoritmo é de 10 e o modelo não gera muda a ordem em que as colunas (SNPs) são apresentados nem as linhas (indivíduos), os valores de

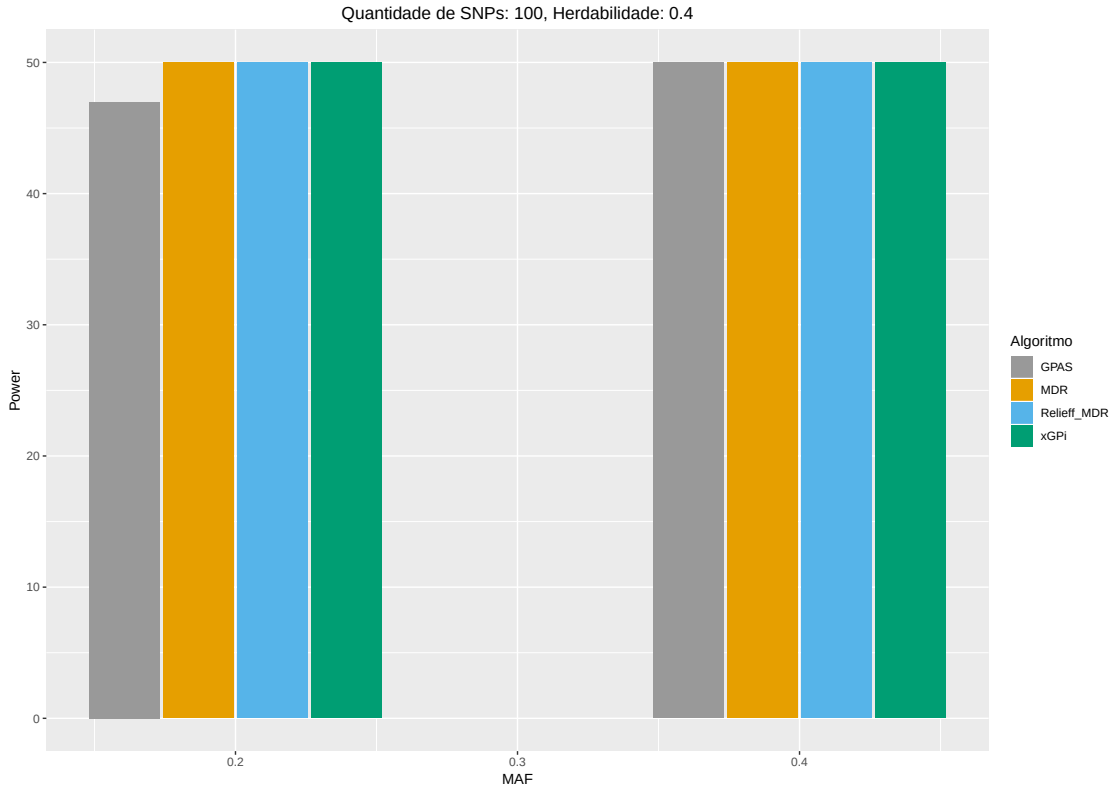


Figura 6.1: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPs causais em conjuntos de dados com 100 marcadores e $h^2 = 0.4$.

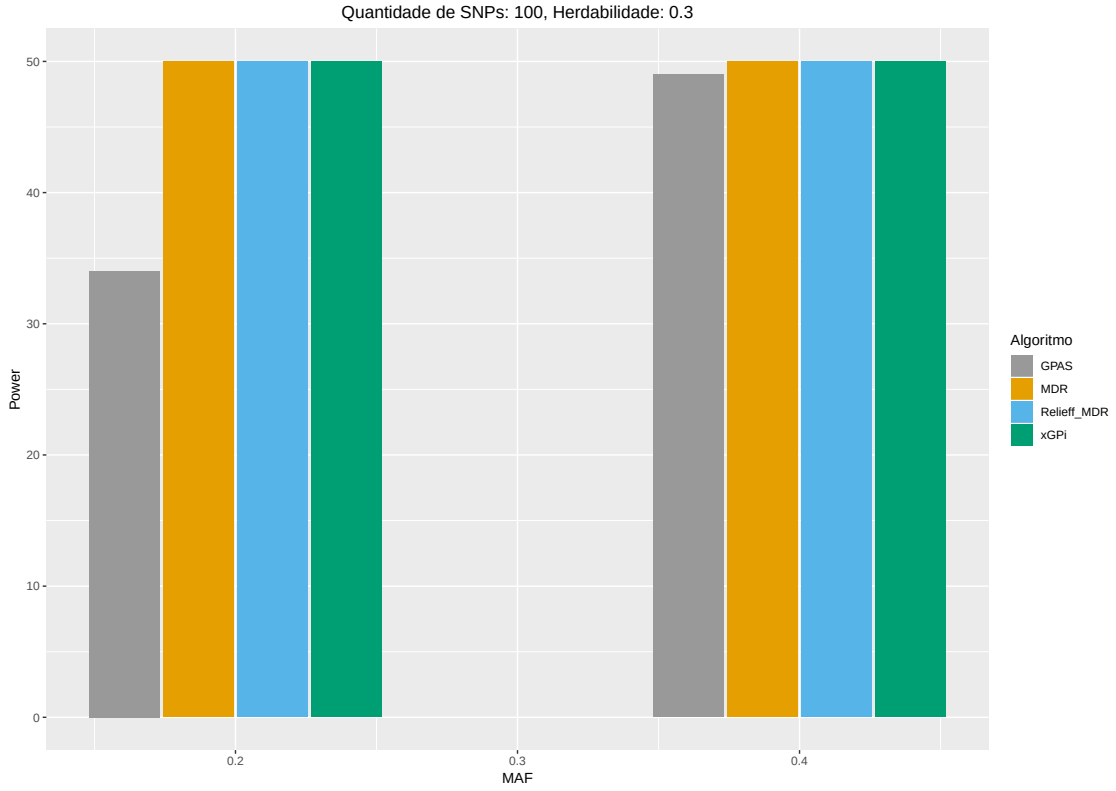


Figura 6.2: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPs causais em conjuntos de dados com 100 marcadores e $h^2 = 0.3$.

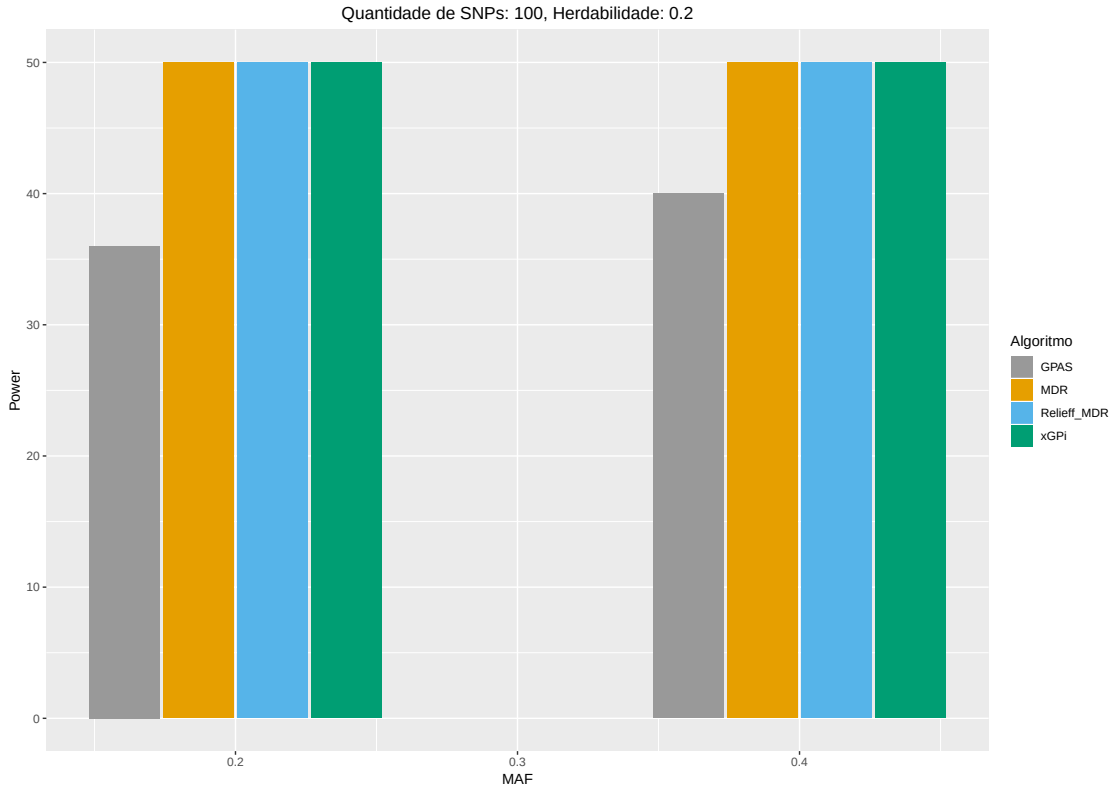


Figura 6.3: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPs causais em conjuntos de dados com 100 marcadores e $h^2 = 0.2$.

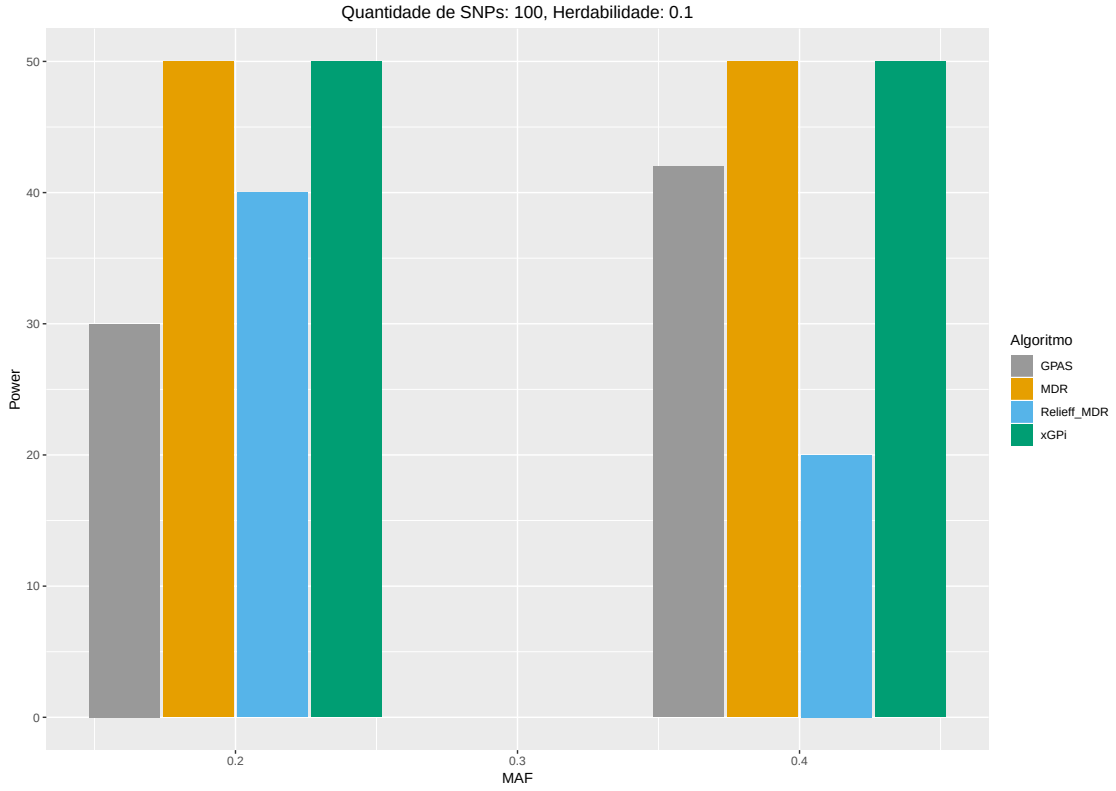


Figura 6.4: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPs causais em conjuntos de dados com 100 marcadores e $h^2 = 0.1$.

importância continuam os mesmos a cada 10 execuções, assim, o modelo seleciona os mesmos vizinhos durante esse processo, o que pode fazer com que independente da MAF, esse fator deve ser levado em consideração para a obtenção da associação correta.

A Tabela 6.3 até Tabela 6.10 apresentam as regras de associação obtidas por cada uma das quatro abordagens distintas em 10 execuções. As primeiras quatro tabelas ilustram as regras obtidas no cenário de herdabilidade 0.4 e MAF 0.4, enquanto as quatro últimas mostram o cenário de herdabilidade 0.1 e MAF 0.2. Somente os dois casos foram apresentados porque representam os casos extremos do primeiro experimento. Nas tabelas, os marcadores causais são representados pelos $SNP99$ e $SNP100$ respectivamente. Para o método MDR e MDR+ReliefF, as regras geradas se diferem das apresentadas pelo GPAS e pelo xGPi, sendo apresentadas na forma de regras "se-então"para cada um dos modelos gerados. Por exemplo: para uma execução em conjunto de dados hipotético, o MDR gerou dois modelos distintos para definir as interações. O primeiro modelo é dado somente pelo $SNP21$ e o segundo pela interação entre os marcadores $SNP1$ e $SNP2$. Os exemplos de modelos gerados pelo MDR são mostrados na Tabela 6.2. Entretanto, quando os modelos são executados via linha de comando, necessários quando os conjuntos de dados possuem uma quantidade maior de SNPs a saída é apresentada somente pelos marcadores da interação, sem as regras. Os resultados do MDR e MDR+Relief são apresentados dessa forma na seção 5.2.3.

Exemplo de uma execução do MDR	
Modelo	Regras
1	Se SNP21 = 1 então classifique como 1.
1	Se SNP21 = 2 então classifique como 0.
1	Se SNP21 = 0 então classifique como 0.
2	Se SNP1 = 2 e SNP2 = 0 então classifique como 1.
2	Se SNP1 = 2 e SNP2 = 1 então classifique como 0.
2	Se SNP1 = 2 e SNP2 = 2 então classifique como 1.
2	Se SNP1 = 1 e SNP2 = 0 então classifique como 1.
2	Se SNP1 = 1 e SNP2 = 1 então classifique como 1.
2	Se SNP1 = 1 e SNP2 = 2 então classifique como 0.
2	Se SNP1 = 0 e SNP2 = 0 então classifique como 0.
2	Se SNP1 = 0 e SNP2 = 1 então classifique como 1.
2	Se SNP1 = 0 e SNP2 = 2 então classifique como 1.

Tabela 6.2: Exemplo das regras "se-então"geradas pelos modelos obtidos pelo método MDR. Foram gerados dois modelos, o primeiro com o marcador SNP21 e o segundo pela interação entre os marcadores SNP1 e SNP2.

Para uma melhor interpretabilidade dos modelos com a finalidade de comparação com as regras geradas pelos métodos GPAS e xGPi, utiliza-se somente a parte das regras que são classificados como sendo da classe 1, ou seja, que são ditos portadores da doença. Assim, pode-se interpretar o modelo 1 e 2 da seguinte forma compacta:

$$\text{Modelo 1: SNP21} = 1 \tag{6.1}$$

$$\begin{aligned}\text{Modelo 2: (SNP1} = 2 \text{ e SNP2} = 0) \text{ OU (SNP1} = 2 \text{ e SNP2} = 2) \text{ OU (SNP1} = 1 \text{ e SNP2} = 0) \\ \text{ou (SNP1} = 1 \text{ e SNP2} = 1) \text{ OU (SNP1} = 0 \text{ e SNP2} = 1) \text{ OU (SNP1} = 0 \text{ e SNP2} = 2)\end{aligned} \tag{6.2}$$

Para as Tabelas referentes aos resultados do MDR e MDR + ReliefF, somente uma das regras será exibida para fins de comparação e ao invés de compactação das tabelas. A seguir, seguem as tabelas com as regras geradas pelos quatro métodos nos dois cenários selecionados.

GPAS- Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	((SNP99!=0) E (SNP100!=0) E (SNP100!=2)) OU((SNP99=0) E (SNP5=0) E (SNP100=0) E (SNP38!=2) E (SNP52!=2)
2	((SNP99!=0) E (SNP100!=0) E (SNP100!=2)) OU((SNP100=0) E (SNP99=0) E (SNP49!=2) E (SNP25!=2) E (SNP85!=2))
3	((SNP100!=0) E (SNP99!=0) E (SNP100!=2)) OU((SNP100=0) E (SNP99=0) E (SNP25!=2) E (SNP49!=2) E (SNP89!=2))
4	((SNP100!=0) E (SNP99!=0) E (SNP100!=2)) OU((SNP99=0) E (SNP100=0) E (SNP49!=2) E (SNP25!=2) E (SNP85!=2))
5	((SNP100!=0) E (SNP99!=0) E (SNP100!=2)) OU((SNP100=0) E (SNP99=0) E (SNP49!=2)) OU((SNP100=2) E (SNP46=0)
6	((SNP99!=0) E (SNP100!=0) E (SNP100!=2)) OU((SNP100=0) E (SNP99=0) E (SNP49!=2) E (SNP12!=2) E (SNP25!=2))
7	((SNP99!=0) E (SNP100!=0) E (SNP100!=2)) OU((SNP100=0) E (SNP99=0) E (SNP49!=2) E (SNP25!=2) E (SNP12!=2))
8	((SNP99!=0) E (SNP100!=0) E (SNP100!=2)) OU((SNP99=0) E (SNP100=0) E (SNP49!=2) E (SNP25!=2) E (SNP85!=2))
9	((SNP99!=0) E (SNP100!=0) E (SNP100!=2)) OU((SNP100=0) E (SNP99=0) E (SNP38!=2) E (SNP5=0) E (SNP26!=2)
10	((SNP99!=0) E (SNP100!=0) E (SNP100!=2)) OU((SNP100=0) E (SNP99=0) E (SNP38!=2) E (SNP5=0) E (SNP52!=2))

Tabela 6.3: Regras de associação geradas pelo método GPAS. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.2.

Ao observar com atenção as tabelas, pode-se notar que as regras geradas pelo método GPAS são mais extensas em relação aos outros métodos. Essa característica se torna mais evidente à medida em que a herdabilidade do conjunto de dados diminui. Pode-se observar também que a segunda parte das regras geradas pelos métodos MDR e MDR + ReliefF são quase idênticas. Esse comportamento se justifica porque nas regras aqui apresentadas, a cada duas regras o conjunto de dados é modificado. Como o método realiza uma busca exaustiva na análise das interações, as mesmas permanecem idênticas se o conjunto de dados for o mesmo. Pode-se observar a mesma característica no algoritmo MDR + ReliefF, isto se dá porque algoritmo RelieF gera o mesmo resultado para filtrar

MDR- Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	SNP64 = 0 OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2)
2	SNP64 = 0 OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2)
3	SNP8 = 0 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2) OU (SNP99 = 2 e SNP100 = 0)
4	SNP8 = 0 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2) OU (SNP99 = 2 e SNP100 = 0)
5	SNP83 = 0 OU SNP83=1 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
6	SNP83 = 0 OU SNP83=1 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
7	SNP9 = 1 OU SNP9=2 OU (SNP99= 1 e SNP100 = 0) OU (SNP99= 1 e SNP100 = 1) OU (SNP99= 2 e SNP100 = 0) OU (SNP99= 2 e SNP100 = 2) OU (SNP99= 0 e SNP100 = 1) OU (SNP99= 0 e SNP100 = 2)
8	SNP9 = 1 OU SNP9=2 OU (SNP99= 1 e SNP100 = 0) OU (SNP99= 1 e SNP100 = 1) OU (SNP99= 2 e SNP100 = 0) OU (SNP99= 2 e SNP100 = 2) OU (SNP99= 0 e SNP100 = 1) OU (SNP99= 0 e SNP100 = 2)
9	SNP80 = 0 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
10	SNP80 = 0 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)

Tabela 6.4: Regras de associação geradas pelo método MDR. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.4.

MDR + ReliefF - Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	SNP14 = 0 OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2)
2	SNP14 = 0 OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2)
3	SNP83 = 0 OU SNP83=1 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
4	SNP83 = 0 OU SNP83=1 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
5	SNP7 = 0 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
6	SNP7 = 0 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) ou (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
7	SNP99 = 1 OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
8	SNP99 = 1 OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
9	SNP80 = 0 OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
10	SNP80 = 0 OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)

Tabela 6.5: Regras de associação geradas pelo método MDR + ReliefF. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.2.

xGPi - Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	(SNP100 != 0 E SNP99 = 0) OU(SNP99 = 1 E SNP100 != 1)
2	(SNP99 = 1 E SNP100 != 1) OU(SNP100 = 1 E SNP99 != 1)
3	(SNP99 != 1 OUSNP100 != 1) E (SNP99 = 1 OUSNP100 = 1)
4	(SNP99 = 1 OUSNP100 = 1) E (SNP100 != 1 OUSNP99 != 1)
5	(SNP99 != 0 E SNP100 = 0) OU(SNP99 = 0 E SNP100 != 0)
6	(SNP99 != 0 E SNP100 != 1) OU(SNP100 != 0 E SNP99 != 1)
7	(SNP99 != 0 E SNP100 = 0) OU(SNP99 = 0 E SNP100 != 0)
8	(SNP99 = 0 E SNP100 != 0) OU(SNP100 != 1 E SNP99 = 1)
9	(SNP100 != 1 OUSNP99 != 1) E (SNP100 = 1 OUSNP99 = 1)
10	(SNP99 != 0 E SNP100 = 0) OU(SNP99 != 1 E SNP100 = 1)

Tabela 6.6: Regras de associação geradas pelo método xGPi. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.4.

os marcadores de um determinado conjunto dados, pela utilização de sua métrica para avaliar a semelhança entre o genótipo de vizinhos mais próximos de todos os indivíduos presentes. Assim, as regras geradas pelos pares apresentam-se idênticas.

Ao observar a Tabela 6.6 e Tabela 6.10, pode-se notar que as regras geradas pela

GPAS - Herdabilidade 0.1 e MAF 0.2	
i	((SNP100=0) E (SNP99!=0)) OU((SNP99=0) E (SNP100!=0)) OU((SNP11=2) E (SNP10!=0) E (SNP66!=0)) OU((SNP30=2) E (SNP61!=0) E (SNP
1	"((SNP100=0) E (SNP99!=0)) OU((SNP99=0) E (SNP100!=0)) OU((SNP11=2) E (SNP10!=0) E (SNP66!=0) E (SNP72!=2)) OU((SNP30=2) E (SNI (SNP29!=2))"
2	((SNP99=0) E (SNP100!=0)) OU((SNP100=0) E (SNP99!=0)) OU((SNP11=2) E (SNP26=0) E (SNP39!=2) E (SNP33=0)) OU((SNP83=2) E (SNP7 (SNP49=2))"
3	((SNP99=0) E (SNP100!=0)) OU((SNP100=0) E (SNP99!=0)) OU((SNP67=2) E (SNP37=0) E (SNP100=0) E (SNP14=0)) OU(SNP84=2) OU((SNP
4	((SNP100=0) E (SNP99!=0)) OU((SNP11=2) E (SNP10!=0) E (SNP66!=0)) OU((SNP99=0) E (SNP100!=0)) OU((SNP11=2) E (SNP89=2) E (SNP5
5	"((SNP100=0) E (SNP99!=0)) OU((SNP99=0) E (SNP100!=0)) OU((SNP85=2) E (SNP7=0) E (SNP32!=2)) OU((SNP22=2) E (SNP47=2)) OU((SN
6	(SNP100=0) E (SNP99!=0)) OU((SNP99=0) E (SNP100!=0)) OU((SNP85=2) E (SNP7=0) E (SNP27!=0)) OU((SNP22=2) E (SNP47=2)) OU((SNP8
7	((SNP100!=0) E (SNP99=0)) OU((SNP100=0) E (SNP99!=0)) OU((SNP86=2) E (SNP75!=0) E (SNP4!=2) E (SNP49=0) E (SNP9=0)) OU((SNP57=
8	((SNP99=0) E (SNP100!=0)) OU((SNP100=0) E (SNP99!=0)) OU((SNP75=2) E (SNP16=2) E (SNP28!=2) E (SNP83!=2)) OU((SNP57=2) E (SNP3
9	((SNP99=0) E (SNP100!=0)) OU((SNP100=0) E (SNP99!=0)) OU((SNP75=2) E (SNP16=2) E (SNP28!=2)) OU((SNP86=2) E (SNP75!=0) E (SNP4
10	((SNP99=0) E (SNP100!=0)) OU((SNP85=2) E (SNP7=0) E (SNP27!=0)) OU((SNP100=0) E (SNP99!=0)) OU((SNP57=2) E (SNP35!=0) E (SNP20

Tabela 6.7: Regras de associação geradas pelo método GPAS. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.1 e MAF de 0.2.

MDR- Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	SNP64 = 0 OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2)
2	SNP64 = 0 OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2)
3	SNP8 = 0 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2) OU (SNP99 = 2 e SNP100 = 0)
4	SNP8 = 0 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2) OU (SNP99 = 2 e SNP100 = 0)
5	SNP83 = 0 OU SNP83=1 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
6	SNP83 = 0 OU SNP83=1 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
7	SNP9 = 1 OU SNP9=2 OU (SNP99= 1 e SNP100 = 0) OU (SNP99= 1 e SNP100 = 1) OU (SNP99= 2 e SNP100 = 0) OU (SNP99= 2 e SNP100 = 2) OU (SNP99= 0 e SNP100 = 1) OU (SNP99= 0 e SNP100 = 2)
8	SNP9 = 1 OU SNP9=2 OU (SNP99= 1 e SNP100 = 0) OU (SNP99= 1 e SNP100 = 1) OU (SNP99= 2 e SNP100 = 0) OU (SNP99= 2 e SNP100 = 2) OU (SNP99= 0 e SNP100 = 1) OU (SNP99= 0 e SNP100 = 2)
9	SNP80 = 0 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
10	SNP80 = 0 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)

Tabela 6.8: Regras de associação geradas pelo método MDR. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.4.

MDR + ReliefF - Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	SNP14 = 0 OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2)
2	SNP14 = 0 OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2)
3	SNP83 = 0 OU SNP83=1 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
4	SNP83 = 0 OU SNP83=1 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
5	SNP7 = 0 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
6	SNP7 = 0 OU (SNP99 = 0 e SNP100 = 1) OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
7	SNP99 = 1 OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
8	SNP99 = 1 OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
9	SNP80 = 0 OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)
10	SNP80 = 0 OU (SNP99 = 0 e SNP100 = 2) OU (SNP99 = 1 e SNP100 = 1) OU (SNP99 = 1 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 0) OU (SNP99 = 2 e SNP100 = 2)

Tabela 6.9: Regras de associação geradas pelo método MDR + ReliefF. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.2.

metodologia proposta nesse trabalho foram as mais compactas entre as abordagens que foram comparadas. Esse resultado indica que para conjunto de dados com um número relativamente pequeno de marcadores e com interação entre pares de SNPs, pouco ou

xGPi - Herdabilidade 0.1 e MAF 0.2	
i	Regras
1	(SNP99 != 0 E SNP100 = 0) OU(SNP99 = 0 E SNP100 != 0)
2	(SNP99 != 1 E SNP100 = 1) OU(SNP100 = 0 E SNP99 != 0)
3	(SNP99 = 0 E SNP100 != 0) OU(SNP100 = 0 E SNP99 = 1)
4	(SNP100 = 1 E SNP99 != 1) OU(SNP100 = 0 E SNP99 != 0)
5	(SNP100 != 1 OUSNP99 != 1) E (SNP99 != 0 OUSNP100 != 0)
6	(SNP100 = 1 OUSNP99 = 1) E (SNP99 != 1 OUSNP100 != 1)
7	(SNP99 != 0 OUSNP100 != 0) E (SNP100 = 0 OUSNP99 = 0)
8	(SNP100 = 0 E SNP99 != 0) OU(SNP100 != 0 E SNP99 = 0)
9	(SNP99 != 0 E SNP100 = 0) OU(SNP99 = 0 E SNP100 != 0)
10	(SNP99 = 1 E SNP100 != 1) OU(SNP99 = 0 E SNP100 != 0)

Tabela 6.10: Regras de associação geradas pelo método xGPi. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.1 e MAF de 0.2.

nenhum ruído é adicionado aos resultados, gerando regras mais precisas e interpretáveis sobre as interações.

6.2.2 Experimentos com interação epistática entre dois loci e 1000 marcadores

Esta seção apresenta os experimentos e resultados obtidos pelas quatro abordagens em conjuntos de dados com mil marcadores. Dessa forma, o número de marcadores foi aumentado em dez vezes em relação ao experimento anterior. Os resultados são mostrados nas Figura 6.5, Figura 6.6, Figura 6.7 e Figura 6.8. Novamente, cada barra nos gráficos representa uma metodologia, sendo o GPAS o primeiro a ser apresentado, seguido pelo MDR e MDR+ReliefF, e por último, o xGPi.

Para esse grupo de conjunto de dados, os parâmetros dos algoritmos foram mantidos em relação aos experimentos anteriores, com exceção do xGPi. O tamanho dos grupos na primeira etapa de execução do algoritmo dobrou de tamanho, ou seja, cada subgrupo apresenta um subconjunto de 20 marcadores, gerando um total de 50 subgrupos. Consequentemente o algoritmo *XGBoost*, é executado sobre os subgrupos combinados, contendo um total de 40 marcadores cada. Portanto, ao final da primeira etapa de execução da metodologia, o subgrupo que obteve o melhor resultado sobre a área sobre a curva ROC a partir do processo de validação cruzada é selecionado. Esse subgrupo de tamanho 40 é então submetido à floresta randômica no passo seguinte e ao final do processo ao algoritmo de programação genética.

Ao observar os resultados obtidos nesse experimento, pode-se notar que mesmo nos casos onde a herdabilidade é 0.4, o algoritmo GPAS apresentou corretamente a combinação dos SNPs causais em no máximo 10% das execuções. Situação que é pouco alterada em relação aos conjuntos de dados com herdabilidade 0.3 e que se agrava nos conjuntos que apresentam herdabilidade 0.2 e 0.1, chegando em casos com MAF de 0.4 não ter encontrado a interação em nenhuma de suas execuções. Portanto, duas questões relevantes a serem ressaltadas sobre o GPAS nesses cenários são: o algoritmo não obtém sucesso ao determinar uma interação epistática quando o número de marcadores no conjunto de dados aumenta e sobre o quanto o modelo sofre influência da pressão da herdabilidade e da MAF utilizada.

Os resultados obtidos pelo método de referência MDR demonstram a robustez e eficiência de sua metodologia. Ele foi método que apresentou os melhores resultados e obteve sucesso ao encontrar as interações em todas as execuções dos conjuntos de dados sobre os diferentes cenários discutidos. Entretanto, o leitor atento deve notar que novamente o número de SNPs causais envolvidos na interação deva ser definido *à priori*, gerando um espaço de busca bem definido e com tamanho referente ao número de marcadores envolvidos no conjunto de dados e os SNPs causais. Sendo assim, nesse cenário especifico, o espaço de busca é de 1000^2 para os pares de marcadores, consequentemente seria de 1000^3 e 1000^4 para a busca de interações de ordem três e quatro, respectivamente.

Pode-se observar que os resultados apresentados pelo método MDR + ReliefF oscilaram nos cenários de herdabilidade 0.3 e 0.4. Mesmo assim, obtiveram sucesso em mais de 60% das execuções. A partir da Figura 6.7, nota-se que o método encontrou as interações somente nos conjuntos com $h^2 = 0.2$ e somente em 20 execuções. Esse número se deve novamente as 10 execuções em cada um dos cinco conjunto de dados gerados com essa configuração. Sendo assim, em dois conjuntos, o algoritmo selecionou corretamente os SNPs na etapa de filtro, e consequentemente como os indivíduos permanecem os mesmos, seus vizinhos mais próximos do ponto de vista do genótipo também. Pode-se inferir então, que o método é sensível à sua etapa inicial. Quando os marcadores não são encontrados nessa etapa, por conseguinte não podem ser encontrados durante a execução do MDR.

O método proposto obteve um resultado próximo ao do MDR, não encontrando as interações somente em duas execuções nos conjuntos com herdabilidade 0.2 e MAF 0.2, três vezes nos conjuntos com herdabilidade 0.1 e MAF 0.2 e somente uma vez nos con-

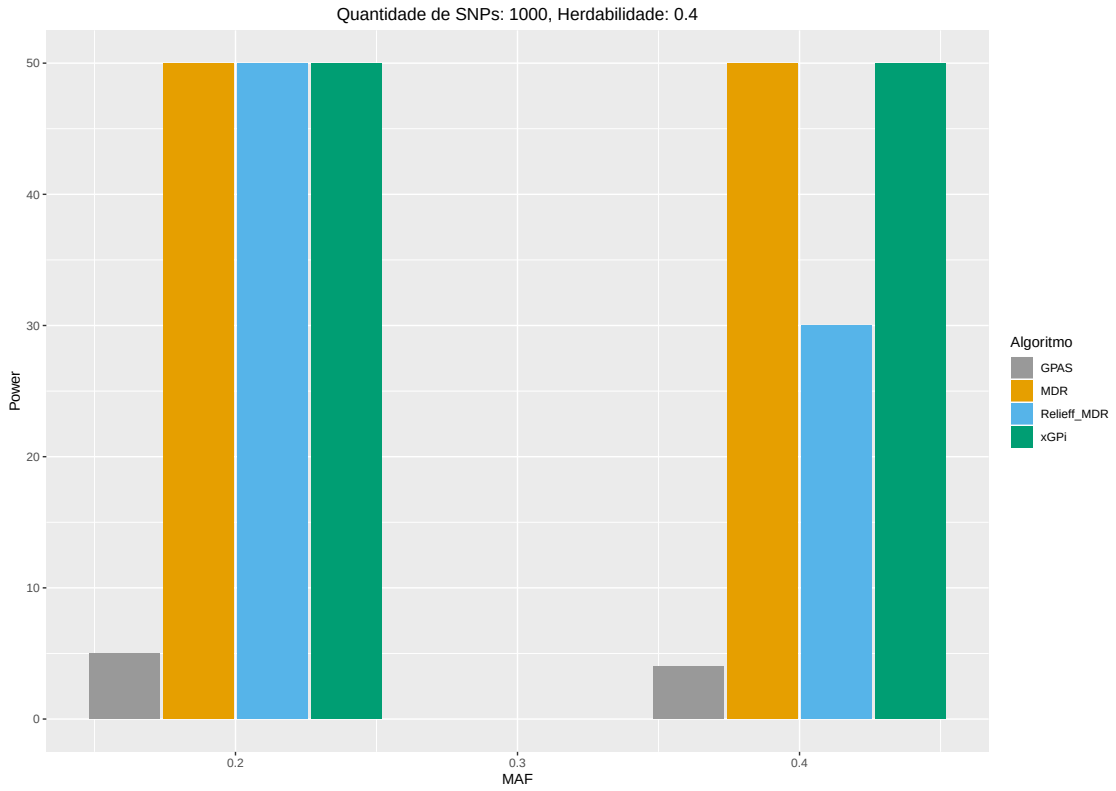


Figura 6.5: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 1000 marcadores e $h^2 = 0.4$.

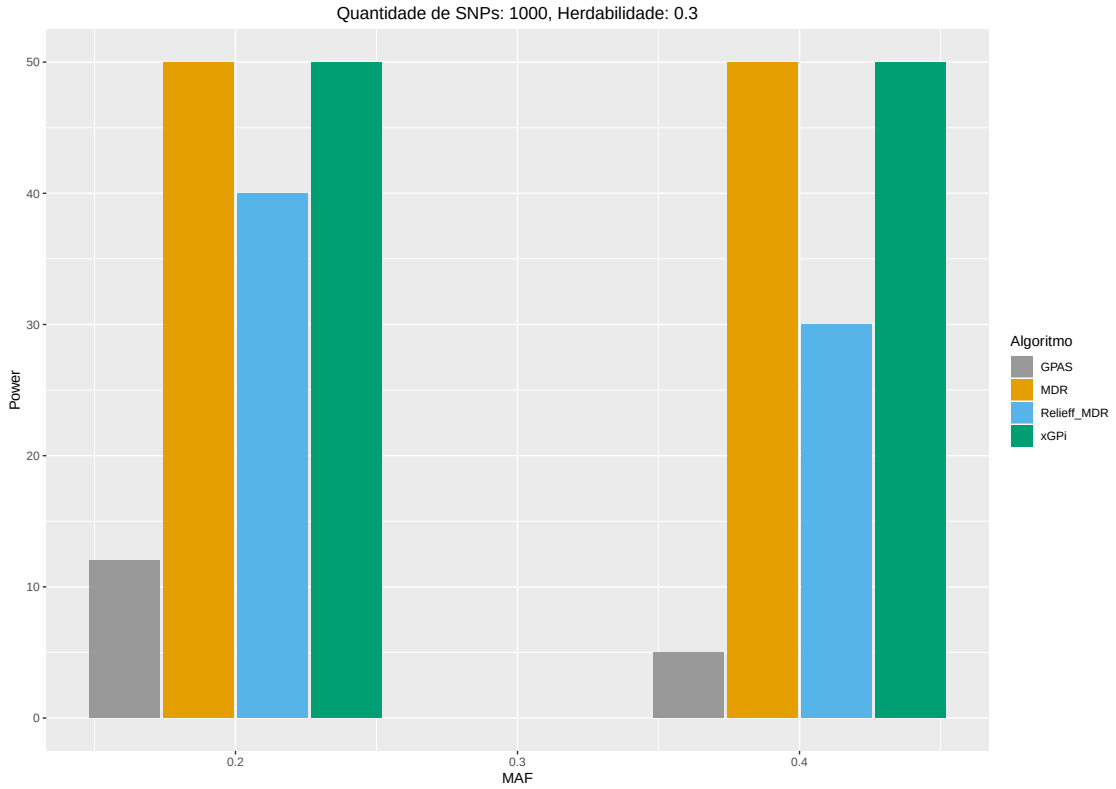


Figura 6.6: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 1000 marcadores e $h^2 = 0.3$.

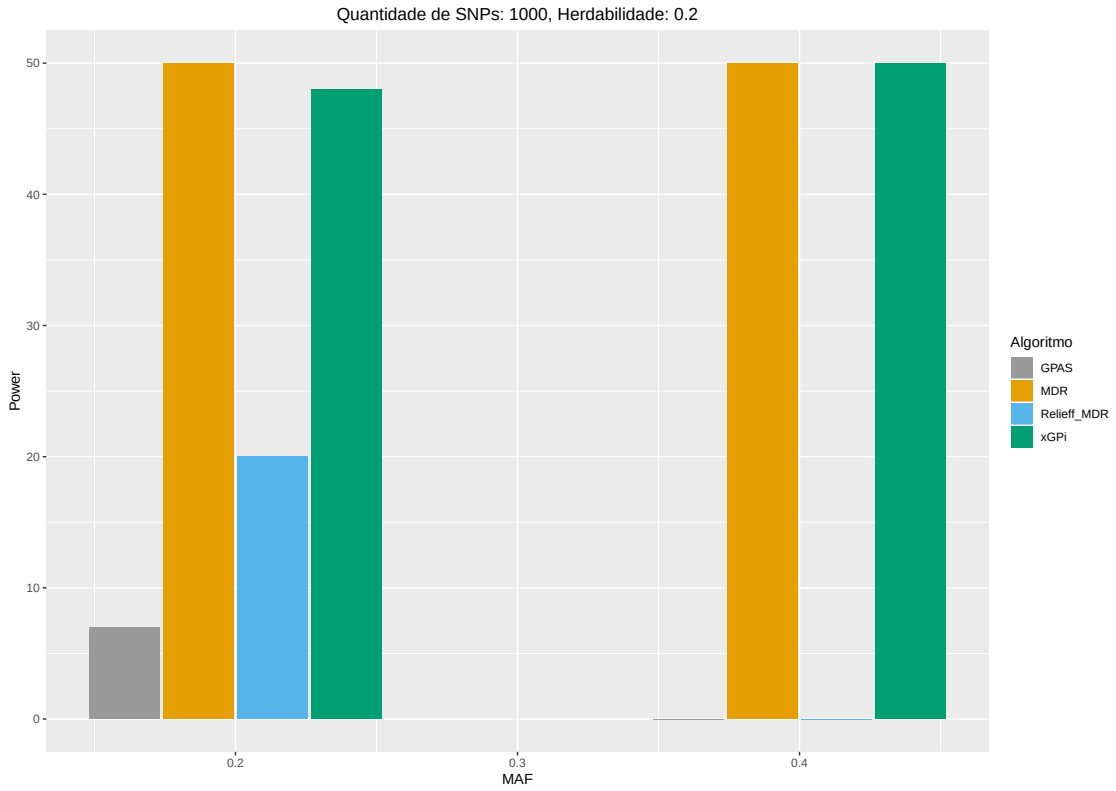


Figura 6.7: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 1000 marcadores e $h^2 = 0.2$.

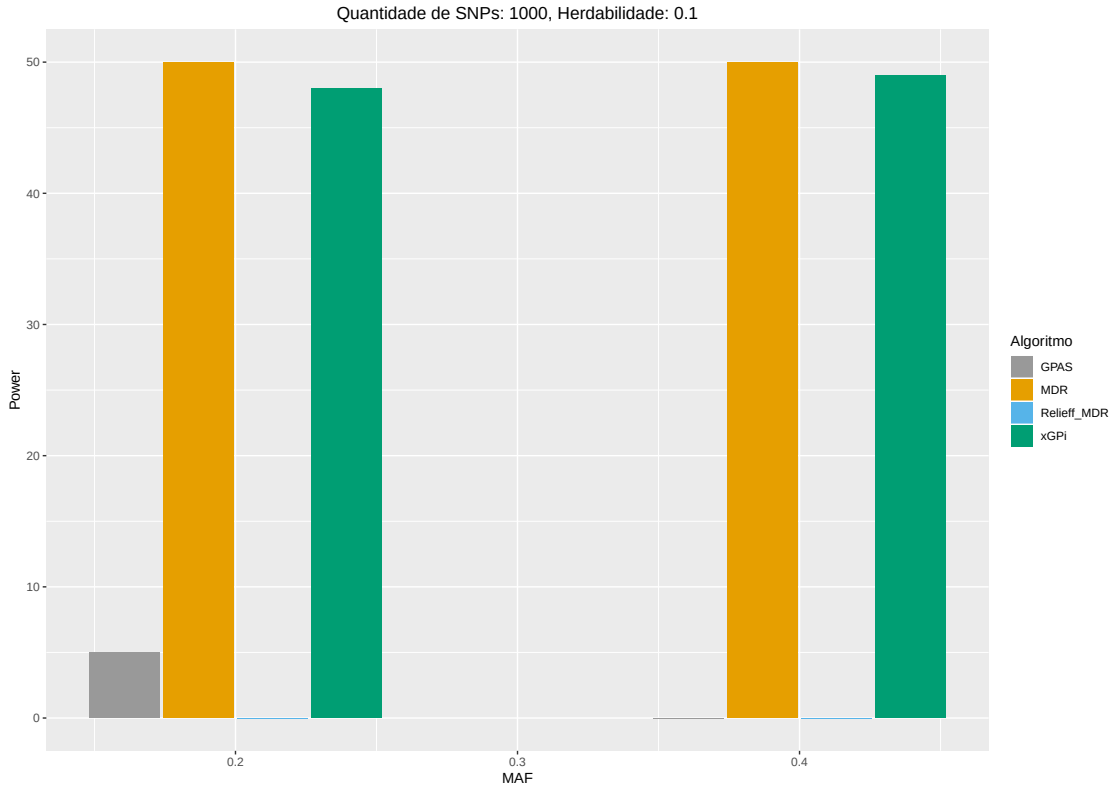


Figura 6.8: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 1000 marcadores e $h^2 = 0.1$.

juntos com herdabilidade 0.1 e MAF 0.4. Esses erros foram geradas durante a primeira etapa de execução do algoritmo. Nestes casos, o tamanho dos subgrupos combinados pode causar pequenas interferências no processo de classificação realizada pelo *XGBoost*, indicando que essa característica deve ser avaliada com mais atenção e de forma mais efetiva posteriormente.

Esse resultados ainda justificam que seguido do MDR, o xGPi foi o método que melhor apresentou resultados relevantes. Ainda em comparação com o MDR, o seu espaço de busca foi de somente $\frac{1}{2} \times \frac{50!}{(50-2)!} = 1225$, ou seja, o número total de combinações sem repetição de SNPs dos subgrupos de tamanho 20, que representa uma redução na complexidade em relação a busca exaustiva realizada pelo MDR que foi de 1000^2. Posteriormente a primeira etapa de execução do método, a ordenação produzida pela métrica pVI do algoritmo de floresta randômica mostrou-se eficiente, pois quase no total das execuções nos cenários abordados os SNPs causais foram alocados em posições superiores, o que fez com que o algoritmo de programação genética inserisse corretamente esses marcadores mais relevantes com maior efetividade no processo de geração da população inicial do algoritmo, fazendo com que as interações pudessem ser encontradas de forma correta em quase todos os cenários.

GPAS - Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	((SNP906=2) e (SNP890!=0)) OU ((SNP869=0) e (SNP990=0) e (SNP194!=0)) OU ((SNP684=2) e (SNP52=0)) OU ((SNP802!=0) e (SNP142=0) e (SNP278!=2)) OU ((SNP562=2) e (SNP551=0))
2	(SNP303!=0) e (SNP399=0)) OU ((SNP869=0) e (SNP244=0) e (SNP993!=2) e (SNP237!=2)) OU ((SNP906=2) e (SNP890!=0)) OU ((SNP784!=0) e (SNP894!=0)) OU (SNP912=2) OU (SNP683=2)
3	((SNP947=0) e (SNP29!=0) e (SNP774!=0)) OU ((SNP802!=0) e (SNP386=0) e (SNP25!=2) e (SNP26!=2)) OU ((SNP617=2) e (SNP409=0)) OU ((SNP684=2) e (SNP52=0)) OU (SNP494=2)
4	((SNP999=0) e (SNP1000=0)) OU ((SNP999!=0) e (SNP1000=2) e (SNP999!=2)) OU ((SNP1000!=0) e (SNP999=2) e (SNP1000!=2)) OU ((SNP1000!=0) e (SNP999!=0) e (SNP869=0) e (SNP909=0))
5	SNP931=2) OU ((SNP732=2) e (SNP291!=2)) OU ((SNP693!=0) e (SNP91=0)) OU ((SNP340!=0) e (SNP662!=0) e (SNP311!=2)) OU (SNP388=2) OU ((SNP534=0) e (SNP566=2) e (SNP278!=0))
6	((SNP693!=0) e (SNP91=0)) OU ((SNP732=2) e (SNP291!=2) e (SNP941=0)) OU (SNP931=2) OU ((SNP388=2) e (SNP965=0)) OU ((SNP340!=0) e (SNP662!=0)) OU (SNP630=2) OU (SNP327=2)
7	((SNP693!=0) e (SNP91=0)) OU (SNP931=2) OU ((SNP732=2) e (SNP291!=2)) OU ((SNP739=2) e (SNP687!=0)) OU ((SNP388=2) e (SNP447!=0)) OU ((SNP340!=0) e (SNP984=0)) OU (SNP239=2)
8	(SNP931=2) OU ((SNP693!=0) e (SNP234!=2)) OU ((SNP732=2) e (SNP291!=2)) OU ((SNP256!=0) e (SNP685!=0) e (SNP101!=0)) OU ((SNP325=2) e (SNP94!=0)) OU ((SNP388=2) e (SNP490!=0))
9	((SNP693!=0) e (SNP91=0)) OU (SNP931=2) OU ((SNP732=2) e (SNP291!=2) e (SNP941=0)) OU ((SNP340!=0) e (SNP275!=0)) OU ((SNP388=2) e (SNP447!=0)) OU ((SNP335=2) e (SNP11=0))
10	(SNP931=2) OU ((SNP732=2) e (SNP291!=2)) OU ((SNP693!=0) e (SNP234!=2)) OU (SNP325=2) OU ((SNP256!=0) e (SNP136!=0) e (SNP808!=0)) OU ((SNP388=2) e (SNP965=0)) OU (SNP327=2)

Tabela 6.11: Regras de associação geradas pelo método GPAS. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.4 em conjuntos de dados com 1000 marcadores.

Dessa forma, as regras geradas pelo xPGi para os cenários com $h^2 = 0.4$ e MAF de 0.4 podem ser observados na Tabela 6.14 e para $h^2 = 0.1$ e MAF de 0.1 na Tabela 6.18. Pode-se observar que o método encontrou alguns ruídos no último cenário, entretanto, em quase todas as execuções, o xGPi encontrou corretamente os marcadores envolvidos na

MDR - Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	SNP751 OU (SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 1)
2	SNP751 OU (SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 1)
3	SNP145 OU (SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 1)
4	SNP145 OU (SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 1)
5	SNP437 OU (SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 1)
6	SNP437 OU (SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 1)
7	SNP901 OU (SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 0 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 2)
8	SNP901 OU (SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 0 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 2)
9	(SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 2) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 0) OU SNP744
10	(SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 2) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 0) OU SNP744

Tabela 6.12: Regras de associação geradas pelo método MDR. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.4 em conjuntos de dados com 1000 marcadores.

MDR + ReliefF - Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	(SNP341 = 0 e SNP438 = 1) OU (SNP341 = 0 e SNP438 = 2) OU (SNP341 = 2 e SNP438 = 0) OU (SNP341 = 2 e SNP438 = 2) OU (SNP341 = 1 e SNP438 = 2) OU SNP341 = 0
2	(SNP341 = 0 e SNP438 = 1) OU (SNP341 = 0 e SNP438 = 2) OU (SNP341 = 2 e SNP438 = 0) OU (SNP341 = 2 e SNP438 = 2) OU (SNP341 = 1 e SNP438 = 2) OU SNP341 = 0
3	(SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 1) OU SNP707=0
4	(SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 1) OU SNP707=0
5	SNP842 = 0 OU (SNP41 = 0 e SNP1000 = 0) OU (SNP41 = 1 e SNP1000 = 1) OU (SNP41 = 2 e SNP1000 = 1)
6	SNP842 = 0 OU (SNP41 = 0 e SNP1000 = 0) OU (SNP41 = 1 e SNP1000 = 1) OU (SNP41 = 2 e SNP1000 = 1)
7	(SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 0 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 2) OU SNP288 = 0
8	(SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 0 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 0) ou (SNP999 = 2 e SNP1000 = 2) OU SNP288 = 0
9	(SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 2) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 2) ou (SNP999 = 2 e SNP1000 = 0) OU SNP891 = 1
10	(SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 2) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 0) OU SNP891 = 1

Tabela 6.13: Regras de associação geradas pelo método MDR + ReliefF. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.4 em conjuntos de dados com 1000 marcadores.

xGPi - Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	(SNP999 = 0 OU SNP1000 != 0) e SNP1000 != 1
2	(SNP999 = 0 e SNP1000 != 2) OU SNP1000 != 0
3	SNP1000 != 0 e SNP999 != 0
4	SNP1000 = 0 e SNP999 != 1
5	SNP1000 != 2 e (SNP999 = 0 OU SNP1000 = 1)
6	(SNP999 != 2 e SNP1000 = 0) OU SNP999 = 1
7	SNP999 != 1 OU (SNP1000 = 1 e SNP999 = 1)
8	SNP1000 != 1 OU SNP999 = 1
9	SNP1000 = 0 e SNP999 != 1
10	(SNP999 = 2 OU SNP1000 = 0) e SNP999 != 1

Tabela 6.14: Regras de associação geradas pelo método xGPi. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.4 em conjuntos de dados com 1000 marcadores.

GPAS - Herdabilidade 0.1 e MAF 0.2	
i	Regras
1	((SNP108!=0) e (SNP957!=0)) OU ((SNP994=2) e (SNP587!=2)) OU ((SNP573=0) e (SNP791=0) e (SNP353!=0)) OU ((SNP147=2) e (SNP645!=0)) OU ((SNP58=2) e (SNP616=0)) OU (SNP805=2)
2	((SNP567!=0) e (SNP575!=0)) OU ((SNP220=0) e (SNP832!=0) e (SNP695=0)) OU ((SNP108!=0) e (SNP957!=0) e (SNP526=0)) OU ((SNP58=2) e (SNP190!=0)) OU ((SNP994=2) e (SNP999=0))
3	((SNP999=0) e (SNP1000!=0) e (SNP1000!=2)) OU ((SNP1000=0) e (SNP999!=0) e (SNP999!=2) e (SNP796!=2) e (SNP497=0)) OU ((SNP986=2) e (SNP999=0)) OU ((SNP220=0) e (SNP272!=0))
4	((SNP108!=0) e (SNP957!=0) e (SNP268!=2)) OU ((SNP220=0) e (SNP205=0) e (SNP853=0) e (SNP987!=2)) OU ((SNP994=2) e (SNP674!=0)) OU ((SNP567!=0) e (SNP215!=0)) OU (SNP805=2)
5	((SNP327=2) e (SNP542=0)) OU ((SNP858!=0) e (SNP444=0) e (SNP598=0)) OU ((SNP790=0) e (SNP857=0) e (SNP294=0)) OU ((SNP812!=0) e (SNP375!=0)) OU ((SNP882=2) e (SNP614=0))
6	(SNP327=2) OU ((SNP858!=0) e (SNP444=0) e (SNP165!=2)) OU ((SNP479=2) e (SNP433=0)) OU ((SNP935=2) e (SNP201!=2)) OU ((SNP661=2) e (SNP822!=0)) OU ((SNP810=2) e (SNP568=0))
7	((SNP858!=0) e (SNP334!=0)) OU (SNP327=2) OU (SNP479=2) OU ((SNP935=2) e (SNP110=0)) OU ((SNP810=2) e (SNP543!=2)) OU ((SNP796=0) e (SNP744=0) e (SNP88!=0) e (SNP626!=2))
8	((SNP858!=0) e (SNP334!=0) e (SNP351!=2)) OU ((SNP327=2) e (SNP542=0)) OU (SNP810=2) OU (SNP479=2) ou ((SNP935=2) e (SNP110=0)) OU ((SNP796=0) e (SNP744=0) e (SNP626!=2))
9	(SNP327=2) OU ((SNP858!=0) e (SNP334!=0) e (SNP351!=2)) OU ((SNP810=2) e (SNP543!=2)) OU (SNP479=2) OU ((SNP935=2) e (SNP201!=2)) OU ((SNP796=0) e (SNP744=0) e (SNP626!=2))
10	((SNP858!=0) e (SNP444=0) e (SNP598=0)) OU (SNP327=2) OU ((SNP790=0) e (SNP294=0) e (SNP857=0)) ou ((SNP812!=0) e (SNP375!=0)) OU ((SNP796=0) e (SNP640!=0) e (SNP726=2))

Tabela 6.15: Regras de associação geradas pelo método GPAS. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.1 e MAF de 0.2 em conjuntos de dados com 1000 marcadores.

MDR - Herdabilidade 0.1 e MAF 0.2	
i	Regras
1	(SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 0 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 0) OU SNP403 = 1
2	(SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 0 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 0) OU SNP403 = 1
3	SNP446 = 1 OU (SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 0 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 0)
4	SNP446 = 1 OU (SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 0 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 0)
5	SNP494 = 1 OU (SNP999 = 0 SNP1000 = 0) OU (SNP999 = 0 SNP1000 = 2) OU (SNP999 = 1 SNP1000 = 1) OU (SNP999 = 2 SNP1000 = 0)
6	SNP494 = 1 OU (SNP999 = 0 SNP1000 = 0) OU (SNP999 = 0 SNP1000 = 2) OU (SNP999 = 1 SNP1000 = 1) OU (SNP999 = 2 SNP1000 = 0)
7	SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 2) OU SNP632 = 1
8	SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 2 e SNP1000 = 2) ou SNP632 = 1
9	(SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 2)
10	(SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 2)

Tabela 6.16: Regras de associação geradas pelo método MDR. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.1 e MAF de 0.2 em conjuntos de dados com 1000 marcadores.

MDR + ReliefF - Herdabilidade 0.1 e MAF 0.2	
i	Regras
1	(SNP303 = 0 e SNP845 = 1) OU (SNP303 = 0 e SNP845 = 2) OU (SNP303 = 1 e SNP845 = 1) OU (SNP303 = 2 e SNP845 = 0)
2	(SNP303 = 0 e SNP845 = 1) OU (SNP303 = 0 e SNP845 = 2) OU (SNP303 = 1 e SNP845 = 1) OU (SNP303 = 2 e SNP845 = 0)
3	(SNP6 = 0 e SNP741 = 0) OU (SNP6 = 1 e SNP741 = 1) OU (SNP6 = 2 e SNP741 = 0) OU SNP800 = 1 OU SNP800 = 2
4	(SNP6 = 0 e SNP741 = 0) OU (SNP6 = 1 e SNP741 = 1) OU (SNP6 = 2 e SNP741 = 0) OU SNP800 = 1 OU SNP800 = 2
5	(SNP696 = 1 e SNP939 = 0) OU (SNP696 = 0 e SNP939 = 1) OU (SNP696 = 0 e SNP939 = 2) OU (SNP696 = 2 e SNP939 = 1)
6	(SNP696 = 1 e SNP939 = 0) OU (SNP696 = 0 e SNP939 = 1) OU (SNP696 = 0 e SNP939 = 2) OU (SNP696 = 2 e SNP939 = 1)
7	(SNP571 = 0 e SNP302 = 1) OU (SNP571 = 0 e SNP302 = 1) OU (SNP571 = 2 e SNP302 = 1) OU (SNP571 = 0 e SNP302 = 1)
8	(SNP571 = 0 e SNP302 = 1) OU (SNP571 = 0 e SNP302 = 1) OU (SNP571 = 2 e SNP302 = 1) OU (SNP571 = 0 e SNP302 = 1)
9	(SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 2)
10	(SNP999 = 0 e SNP1000 = 0) OU (SNP999 = 1 e SNP1000 = 1) OU (SNP999 = 1 e SNP1000 = 2) OU (SNP999 = 2 e SNP1000 = 2)

Tabela 6.17: Regras de associação geradas pelo método MDR + ReliefF. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.1 e MAF de 0.2 em conjuntos de dados com 1000 marcadores.

interação.

xGPi - Herdabilidade 0.1 e MAF 0.2	
i	Regras
1	(SNP999 = 0 e SNP1000 != 1) OU (SNP1000 = 1 e (SNP1000 != 1 OU SNP999 = 1))
2	(SNP776 = 2 OU (SNP1000 = 1 e SNP999 = 1)) OU (SNP280 != 2 e SNP245 = 1)
3	(SNP1000 != 1 e SNP999 != 1) OU (SNP1000 = 1 e SNP999 = 1)
4	SNP152 != 2 e ((SNP246 != 0 OU (SNP902 = 0 OU SNP429 != 0)) e SNP724 = 0)
5	(SNP1000 != 0 OU SNP999 = 0) e (SNP999 = 1 OU (SNP1000 = 0 e SNP999 = 0))
6	SNP999 != 0 e SNP1000 != 0) OU ((SNP999 = 0 OU SNP1000 = 2) e SNP1000 = 0)
7	((SNP1000 != 2 e SNP999 = 0) OU SNP1000 != 0) e (SNP1000 = 0 OU SNP999 != 0)
8	(SNP1000 = 0 OU SNP999 != 0) e (SNP1000 = 1 OU (SNP999 = 0 e SNP1000 = 0))
9	(SNP1000 != 0 OU (SNP999 = 0 e SNP1000 = 0)) e (SNP1000 = 0 OU SNP999 != 0)
10	(SNP1000 = 0 e SNP999 = 0) OU (SNP999 != 0 e SNP1000 != 0)

Tabela 6.18: Regras de associação geradas pelo método xGPi. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.1 e MAF de 0.2 em conjuntos de dados com 1000 marcadores.

6.2.3 Experimentos com interação epistática entre dois loci e 10000 marcadores

Nesta seção são apresentados os experimentos e resultados com conjuntos de dados que apresentam interações epistáticas entre dois *loci* e 10000 marcadores. Foram utilizados para comparação o xGPi, o MDR e o MDR+ReliefF. Somente o modelo GPAS não foi comparado, devido a limitação do modelo de se trabalhar com conjuntos de dados dessa magnitude. Neste cenário, a interação epistática foi modelada pelos marcadores causais: $SNP9999$ e $SNP10000$. Os resultados obtidos nos experimentos são mostrados na Figura 6.9, Figura 6.10, Figura 6.11 e Figura 6.12.

Novamente, os parâmetros dos algoritmos MDR e MDR+ReliefF foram mantidos em relação aos experimentos anteriores. Já para o algoritmo xGPi, o tamanho dos subgrupos na primeira etapa do modelo foi alterado para os testes em conjuntos de dados com herdabilidade 0.1, que teve o tamanho fixado em 10. Já para os subgrupos com demais herdabilidades, o tamanho para os subgrupos foi de 20, ou seja, o dobro do tamanho. Essa fato ocorreu devido a diminuição da herdabilidade nos dados com um número grande de marcadores, assim, para os experimentos executados nesta seção, fez-se necessário esta adaptação. Foram geradas um total de 124750 combinações para cada experimento, resultantes de um total de 500 subgrupos. Já para $h^2 = 1$, foi necessário a criação de 1000 subgrupos de tamanho 10, gerando um total de 499500 combinações.

Pode-se observar através dos resultados obtidos pelo MDR que o modelo encontrou

os SNPs causais corretamente em todos os cenários avaliados. Entretanto, para essa quantidade de marcadores em um conjunto de dados, foram necessários $10000 \times 10000 = 10^8$ classificadores para testar todas as combinações entres pares de marcadores para determinar a associação de interesse. A Tabela 6.19 e Tabela 6.22 exibem as regras geradas pelo algoritmo nos cenários de $h^2 = 0.4$ e $h^2 = 0.1$.

O método MDR+ReliefF não apresentou resultados consistentes nos cenários avaliados. Em nenhuma das execuções as interações de interesse foram detectadas. As Tabela 6.20 e Tabela 6.23 mostram as regras geradas pelo modelo, pode-se observar que em todas as regras obtidas o método foi capaz de capturar apenas ruídos.

Ao observar os resultados obtidos nesse experimento, pode-se notar que o xGPi obteve sucesso em todos cenários, identificando corretamente as interações em quase todos os conjuntos de dados com os diferentes níveis de herdabilidade e de MAF. Apenas para os conjuntos de dados com $h^2 = 0.1$ o método não identificou corretamente os marcadores envolvidos na interação epistática em 4 execuções. As Tabela 6.21 e Tabela 6.24 mostram as regras geradas pelo xGPi nos conjuntos de dados com $h^2 = 0.4$ e MAF 0.4 e $h^2 = 0.1$ e MAF de 0.2.

Através desse experimento, pode-se observar que o modelo proposto foi capaz de lidar com conjuntos de dados de tamanho 10^3, encontrando resultados consistentes mesmo em cenários que apresentam baixa herdabilidade. Como discutido anteriormente no experimento anterior, que o tamanho dos subgrupos combinados pode causar pequenas interferências no processo de classificação realizada pelo *XGBoost*. A adoção de um tamanho menor de subgrupos para níveis de herdabilidade muito baixas foi necessário para a identificação correta dos marcadores causais.

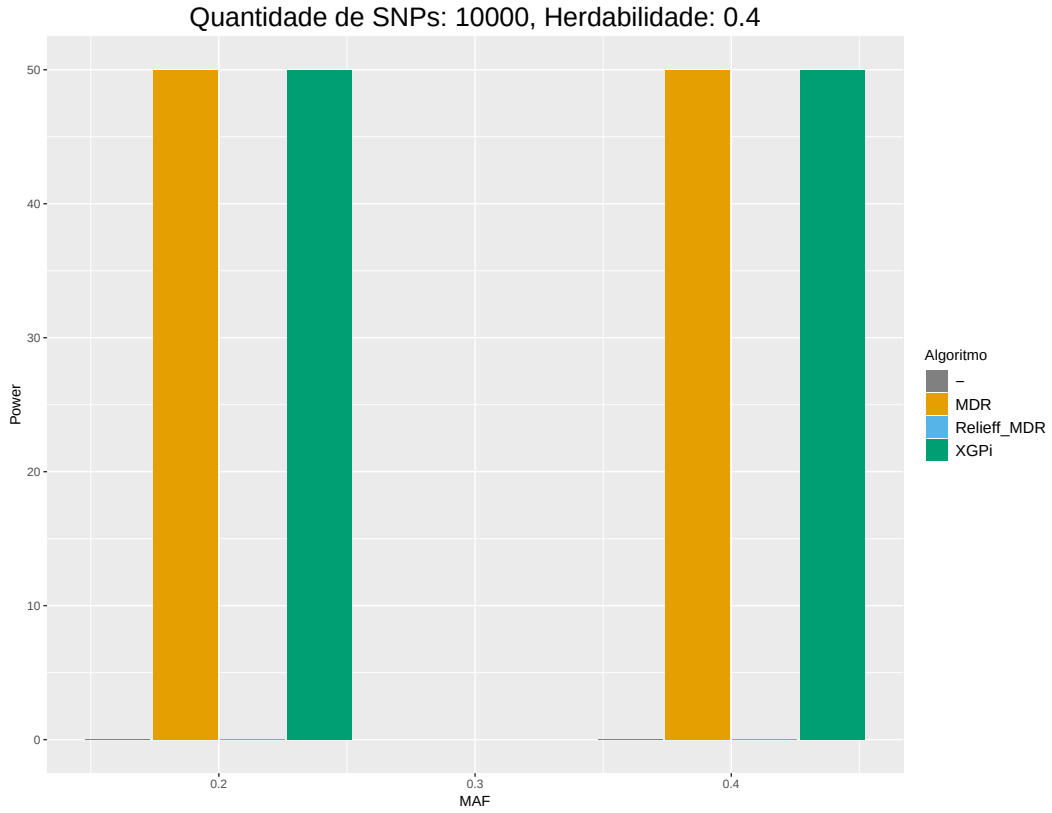


Figura 6.9: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 10000 marcadores e $h^2 = 0.4$.

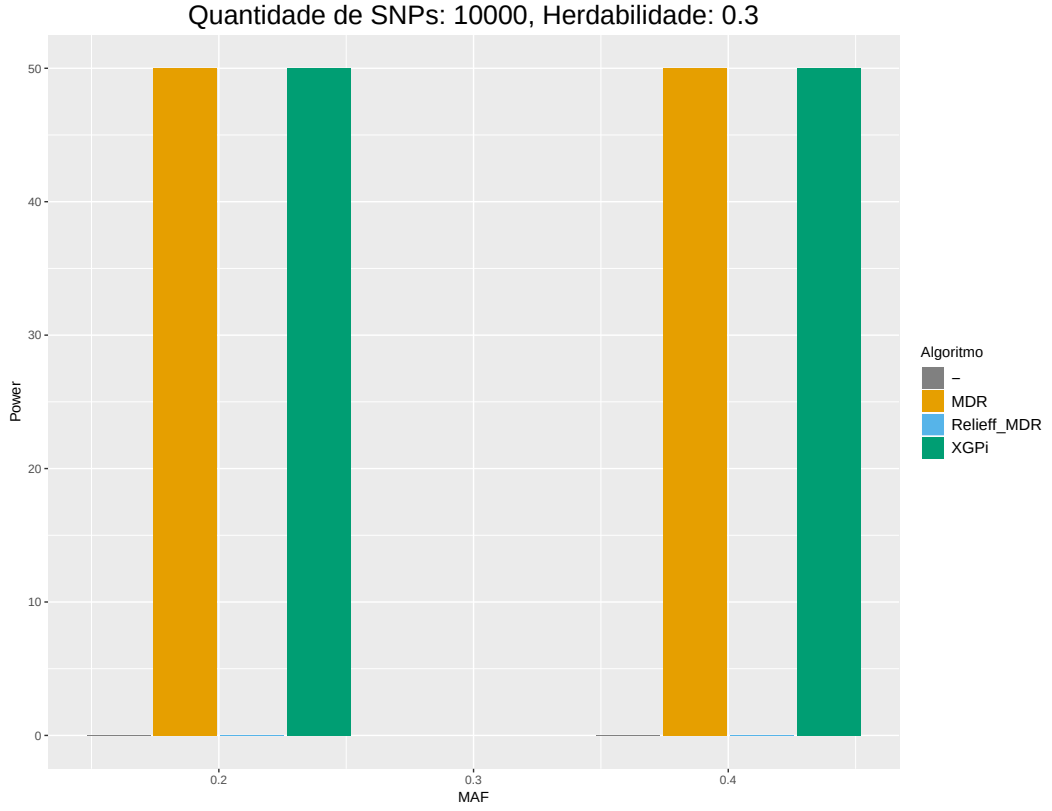


Figura 6.10: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 10000 marcadores e $h^2 = 0.3$.

Figura 6.11: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 10000 marcadores e $h^2 = 0.2$.

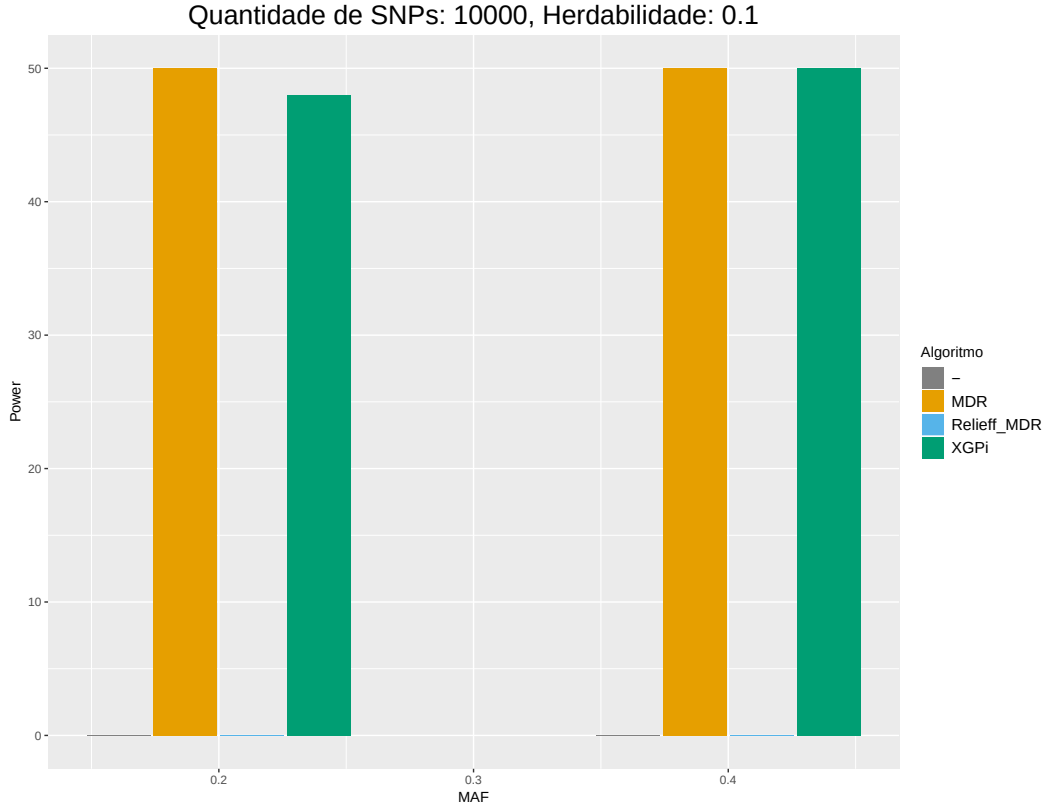


Figura 6.12: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 10000 marcadores e $h^2 = 0.1$.

MDR - Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	(N9999 = 1 e N10000=1) OU (N9999 = 1 e N10000=2)
2	(N9999 = 1 e N10000=1) OU (N9999 = 1 e N10000=2)
3	(N9999 = 0 e N10000=0) OU (N9999 = 1 e N10000=2) OU (N9999 = 1 e N10000=0)
4	(N9999 = 0 e N10000=0) OU (N9999 = 1 e N10000=2) OU (N9999 = 1 e N10000=0)
5	(N9999 = 2 e N10000=2) OU (N9999 = 2 e N10000=0) OU (N9999 = 1 e N10000=1)
6	(N9999 = 2 e N10000=2) OU (N9999 = 2 e N10000=0) OU (N9999 = 1 e N10000=1)
7	(N9999 = 1 e N10000=1)
8	(N9999 = 1 e N10000=1)
9	(N9999 = 0 e N10000=1) OU (N9999 = 2 e N10000=1)
10	(N9999 = 0 e N10000=1) OU (N9999 = 2 e N10000=1)

Tabela 6.19: Regras de associação geradas pelo método MDR. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.4 em conjuntos de dados com 10000 marcadores.

MDR+ReliefF - Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	(N7885 = 0 e N3700 = 0) OU N7885 = 1
2	(N7885 = 0 e N3700 = 0) OU N7885 = 1
3	N2548 = 1 OU (N7505 = 0 e N435 = 1)
4	N2548 = 1 OU (N7505 = 0 e N435 = 1)
5	N7505 = 0 OU (N435 = 0 e N2548 = 0)
6	N7505 = 0 OU (N435 = 0 e N2548 = 0)
7	(N8460 = 2 e N262 = 1) OU N8460 = 0
8	(N8460 = 2 e N262 = 1) OU N8460 = 0
9	(N4744 = 2 e N1169 = 0) OU N4744 = 2
10	(N4744 = 2 e N1169 = 0) OU N4744 = 2

Tabela 6.20: Regras de associação geradas pelo método MDR+ReliefF. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.4 em conjuntos de dados com 10000 marcadores.

xGPi - Herdabilidade 0.4 e MAF 0.4	
i	Regras
1	(N9999 != 1 E (N10000 != 1 OUN9999 = 0)) E N10000 != 0
2	(N9999 = 1 E N10000 != 2) OU(N5918 != 1 E (N10000 = 1 E N9999 != 2))
3	((N10000 = 0 OUN9999 != 1) E N10000 != 1) OU(N10000 = 1 E N9999 != 2)
4	(N9999 = 0 E N10000 = 1) OU(N10000 != 1 E N9999 = 1)
5	(N9999 = 0 E N10000 != 0) OU(N10000 != 1 E N9999 = 1)
6	((N10000 != 1 OUN10000 = 0) E N9999 != 0) E (N10000 = 0 OUN9999 != 2)
7	(N9999 != 2 E ((N9999 != 2 E N10000 != 1) OUN9999 != 1)) E N10000 != 0
8	(N10000 = 1 E N9999 != 1) OU(N9999 = 1 E N10000 = 0)
9	N9999 != 1 E ((N10000 != 0 E N9999 = 0) OUN10000 = 2)
10	(N9999 != 0 E N10000 != 2) OU(N9999 != 1 E N10000 != 0)

Tabela 6.21: Regras de associação geradas pelo método xGPi. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.4 e MAF de 0.4 em conjuntos de dados com 10000 marcadores.

MDR - Herdabilidade 0.1 e MAF 0.2	
i	Regras
1	(N9999 = 1 e N10000= 2) OU (N9999 = 0 e N10000= 1) OU (N9999 = 0 e N10000 = 0) OU (N9999 = 2 e N10000 =1)
2	(N9999 = 1 e N10000= 2) OU (N9999 = 0 e N10000= 1) OU (N9999 = 0 e N10000 = 0) OU (N9999 = 2 e N10000 =1)
3	(N9999 = 1 e N10000=1) OU (N9999 = 0 e N10000=1)
4	(N9999 = 1 e N10000=1) OU (N9999 = 0 e N10000=1)
5	(N9999 = 0 e N10000= 1) OU (N9999 = 1 e N10000= 2) OU (N9999 = 0 e N10000 = 0) OU (N9999 = 1 e N10000 =1)
6	(N9999 = 0 e N10000= 1) OU (N9999 = 1 e N10000= 2) OU (N9999 = 0 e N10000 = 0) OU (N9999 = 1 e N10000 =1)
7	(N9999 = 2 e N10000= 1) OU (N9999 = 0 e N10000= 0) OU (N9999 = 0 e N10000 = 1)
8	(N9999 = 2 e N10000= 1) OU (N9999 = 0 e N10000= 0) OU (N9999 = 0 e N10000 = 1)
9	(N9999 = 2 e N10000=1) OU (N9999 = 0 e N10000=1)
10	(N9999 = 2 e N10000=1) OU (N9999 = 0 e N10000=1)

Tabela 6.22: Regras de associação geradas pelo método MDR. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.1 e MAF de 0.2 em conjuntos de dados com 10000 marcadores.

MDR+ReliefF - Herdabilidade 0.1 e MAF 0.2	
i	Regras
1	N639=0 OU (N639=0 e N7990=1)
2	N639=0 OU (N639=0 e N7990=1)
3	(N7799 = 1 e N9875 = 0) OU N1822 = 1
4	(N7799 = 1 e N9875 = 0) OU N1822 = 1
5	(N1720 = 0 e N1720 = 0) OU (N8035 =1) OU N5846 = 1 OU (N5846 = 0 e N7747=1)
6	(N1720 = 0 e N1720 = 0) OU (N8035 =1) OU N5846 = 1 OU (N5846 = 0 e N7747=1)
7	(N4098 = 1 e N284 = 2) OU N4098 = 1
8	(N4098 = 1 e N284 = 2) OU N4098 = 1
9	(N3993 = 0 e N3508 = 1) OU N3993 = 0
10	(N3993 = 0 e N3508 = 1) OU N3993 = 0

Tabela 6.23: Regras de associação geradas pelo método MDR+ReliefF. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.1 e MAF de 0.2 em conjuntos de dados com 10000 marcadores.

xGPi - Herdabilidade 0.1 e MAF 0.2	
i	Regras
1	(N10000 != 0 E N9999 = 1) OU(N2748 = 1 E (N9999 = 2 OUN5011 = 1))
2	(N9999 = 1 E N10000 = 1) OU((N9999 != 0 OUN10000 = 0) E N9999 != 1)
3	((N10000 = 1 E N10000 = 1) OUN9999 != 1) E N10000 != 2
4	(N9999 != 1 E N10000 != 1) OU(N9999 != 0 E N10000 != 0)
5	(N9999 = 0 E (N252 = 2 OUN10000 = 0)) OUN10000 = 1
6	(N10000 = 0 E N599 != 1) OU(N9999 != 0 E N10000 = 1)
7	(N10000 = 0 OUN10000 = 2) E (N10000 != 2 E N9999 = 0)
8	(N9999 != 1 E N10000 != 1) OU(N9999 != 0 E N10000 != 0)
9	(N9999 != 0 OUN10000 != 1) E (N9999 != 1 OUN10000 != 0)
10	(N10000 != 0 OUN9999 = 0) E (N10000 = 0 OUN491 = 1)

Tabela 6.24: Regras de associação geradas pelo método xGPi. Cada regra representa uma de dez execuções nos conjuntos de dados com herdabilidade de 0.1 e MAF de 0.2 em conjuntos de dados com 10000 marcadores.

6.2.4 Experimentos com interação epistática entre três loci

Nesta seção são discutidos experimentos com conjuntos de dados que apresentam interações epistáticas entre 3 marcadores. Ou seja, o objetivo do experimento é identificar interação entre alelos presentes em 3 *loci* distintos que estão significantemente associados com o fenótipo. Diferentemente de simular interações entre 2 *loci*, o GAMETES não permite uma customização completa na geração de conjuntos de dados de interações entre 3 marcadores. Dessa forma, o conjunto de dados gerado apresenta os seguintes parâmetros: $h^2 = 0.22$ e MAF de 0.2. Foram gerados conjuntos de dados de 100 marcadores para os experimentos. As funções de penetrância do modelo são apresentadas na Tabela 6.25, Tabela 6.26 e Tabela 6.27.

$h^2 = 0.22$, MAF=0.2			
	AA	Aa	aa
BB	0.273	0.538	0.292
Bb	0.511	0.036	0.489
bb	0.504	0.064	0.383

Tabela 6.25: Função de penetrância do modelo de ordem 3. Interação entre o SNP de interesse 1 e e SNP de interesse 2.

$h^2 = 0.22$, MAF=0.2			
	AA	Aa	aa
CC	0.522	0.010	0.531
Cc	0.070	0.977	0.023
cc	0.052	0.981	0.285

Tabela 6.26: Função de penetrância do modelo de ordem 3. Interação entre o SNP de interesse 1 e e SNP de interesse 3.

No experimento, os marcadores de interesse são os $SNP98$, $SNP99$ e $SNP100$. Representados nas funções de penetrância por A, B e C, respectivamente. Os algoritmos GPAS, MDR, MDR+ReliefF foram comparados com o modelo proposto, de forma similar aos experimentos apresentados nas seções anteriores. Para avaliar uma interação entre 3 marcadores a seleção de subgrupos do método xGPi foi modificada. Assume-se então, que as combinações são realizadas em 3 subgrupos. Por exemplo, considere o conjunto de

$h^2 = 0.22$, MAF=0.2			
	BB	Bb	bb
CC	0.421	0.271	0.047
Cc	0.218	0.565	0.957
cc	0.478	0.096	0.546

Tabela 6.27: Função de penetrância do modelo de ordem 3. Interação entre o SNP de interesse 2 e e SNP de interesse 3.

subgrupos $P(B) = \{1, 2, 3, 4, 5\}$, observando que $P(B)$ possui um total de 5 elementos. Nesta caso, utiliza-se somente os subgrupos de tamanho três. Dessa forma o mesmo conjunto B seria composto por: $P(B) = \{\{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 5\}, \{1, 3, 4\}, ..., \{3, 4, 5\}\}$. Para esse experimento, cada subgrupo teve o seu tamanho fixado em 5. Os resultados obtidos pelos métodos avaliados serão demonstrados na Figura 6.13.

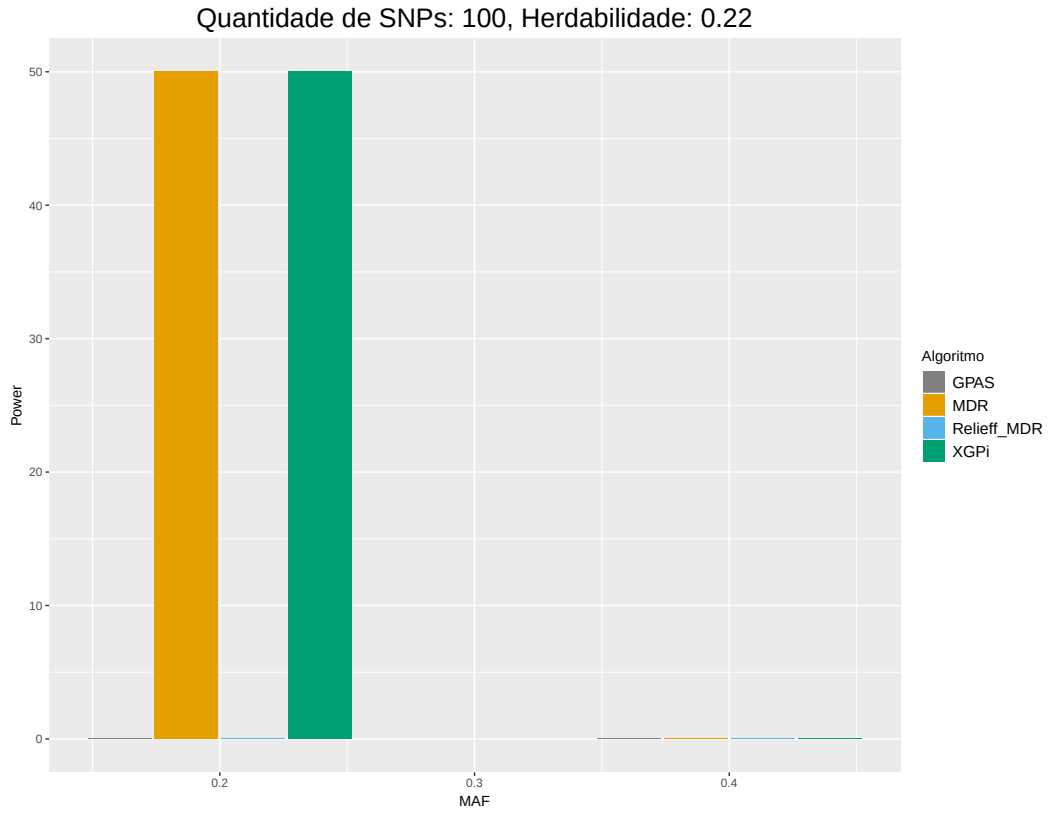


Figura 6.13: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os três SNPS causais em conjuntos de dados com 100 marcadores, $h^2 = 0.22$ e MAF=0.2.

Pode-se observar que os modelos GPAS e MDR+ReliefF não foram capazes de identificar corretamente as interações de interesse em nenhuma execução. Em contrapartida o modelo proposto e o MDR identificaram as interações em todas as execuções. Nota-

se que o algoritmo GPAS não foi capaz de lidar com interações de ordem superior a 2, limitando-se a identificação de pares de SNPs. O MDR+ReliefF é fortemente dependente do inicialização do algoritmo ReliefF, que nesse cenário não conseguiu identificar vizinhos próximos que auxiliariam na descoberta das associações.

Para o modelo proposto, depois da etapa de seleção de subgrupos, foi realizado o processo de ranqueamento. A Figura 6.14 mostra o resultado da avaliação de cada marcador pela métrica pVI da floresta randômica. Pode-se observar que os marcadores de interesse $SNP98$, $SNP99$ e $SNP100$ estão localizados no topo da lista.

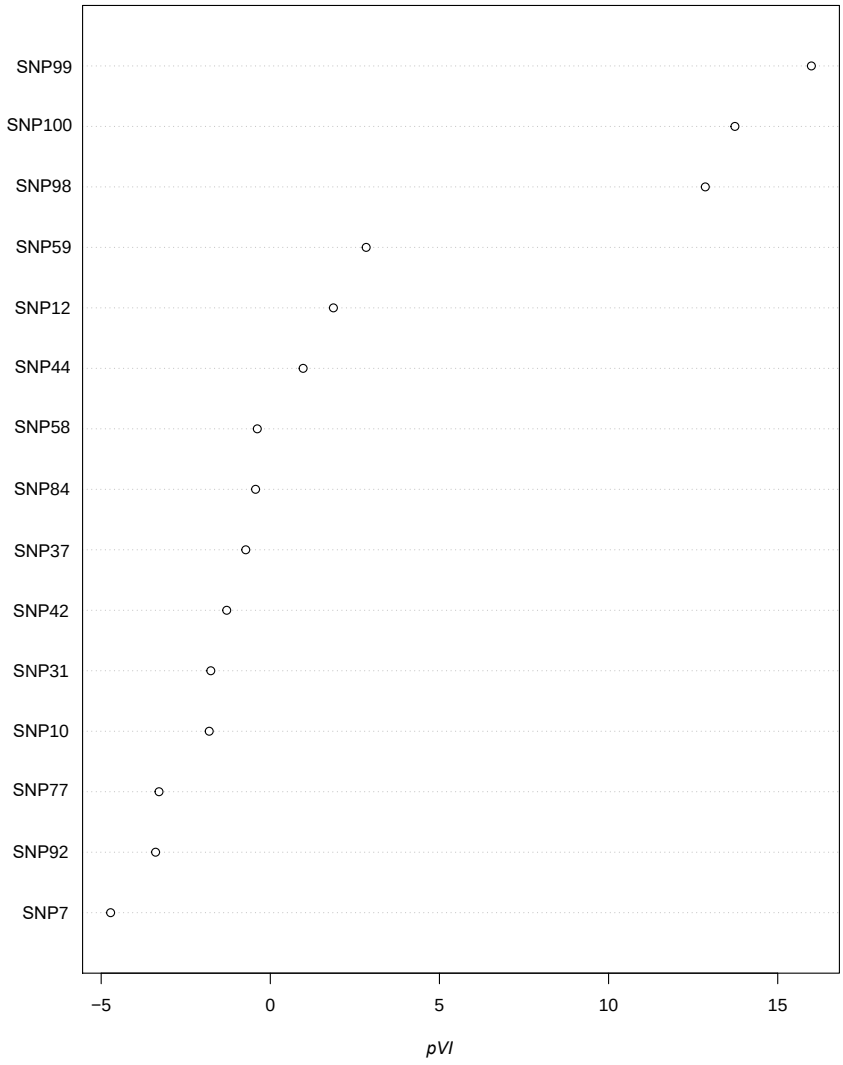


Figura 6.14: Resultado do processo de ranqueamento utilizando a métrica pVI.

As regras geradas pelo modelo proposto podem ser observadas na Tabela 6.28. Nota-se que as regras demostram a interação de forma consistente. Avaliando o processo de

execução do xGPi, os resultados mostram que o método pode ser aplicado em cenários com interações entre um número maior de marcadores.

	xGPi - Regras
1	N98 = 1 E (N100 != 1 E (N100 != 0 OUN99 != 0))
2	(N99 = 1 E N100 != 1) OU((N100 != 0 E N99 = 0) OUN98 = 1)
3	(N100 != 1 OUN99 != 0) E ((N99 != 0 E N100 != 1) E N98 = 1)
4	N100 = 2 OU((N99 != 0 E N100 != 1) OU(N98 = 1 E N100 != 2))
5	(N100 != 1 OUN98 = 1) E ((N98 = 1 E N100 = 0) OUN99 != 0)

Tabela 6.28: Regras de associação geradas pelo algoritmo xGPi.

	GPAS - Regras
1	(SNP71=0) E (SNP2=0) E (SNP83!=0)) OU((SNP19=0) E (SNP75!=0))
2	((SNP89=2) E (SNP10=0)) OU((SNP19=0) E (SNP75!=0)) OU((SNP27=2) E (SNP29!=2))
3	(SNP19=0) E (SNP79=0)) OU((SNP19=0) E (SNP75!=0)) OU((SNP64=0) E (SNP73!=0) E (SNP65=0))
4	(SNP19=0) E (SNP79=0)) OU((SNP19=0) E (SNP75!=0))
5	((SNP89=2) E (SNP10=0)) OU((SNP26=2) E (SNP88=0))

Tabela 6.29: Regras de associação geradas pelo algoritmo GPAS.

	MDR - Regras
1	(SNP98 = 0 E SNP99 = 0 E SNP100 = 0) OU (SNP98 = 0 E SNP99 = 0 E SNP100 = 2)
2	(SNP98 = 0 E SNP99 = 2) OU (SNP98 = 0 E SNP99 = 2 E SNP100 = 1)
3	(SNP98 = 2 E SNP99 = 0) E SNP100 = 1 OU (SNP98 = 2 E SNP99 = 2 E SNP100 = 0)
4	(SNP98 = 2 E SNP99 = 1 E SNP100 = 1) OU (SNP98 = 1 E SNP99 = 2 E SNP100 = 0)
5	(SNP98 = 0 E SNP99 = 1 E SNP100 = 1)

Tabela 6.30: Regras de associação geradas pelo algoritmo MDR.

	MDR+ReliefF - Regras
1	(SNP18 = 0 E SNP45 = 2 E SNP57 = 2) OU (SNP18 = 0 E SNP45 = 2 E SNP57 = 1)
2	SNP18 = 2 E SNP45 = 0 E SNP57 = 0
3	SNP18 = 1 E SNP45 = 2 E SNP57 = 1
4	(SNP18 = 1 E SNP45 = 1 E SNP57 = 2) OU (SNP18 = 0 E SNP45 = 1 E SNP57 = 2)
5	(SNP18 = 0 E SNP45 = 2 E SNP57 = 2)

Tabela 6.31: Regras de associação geradas pelo algoritmo MDR + ReliefF.

O tamanho de cada subgrupo deve ser melhor avaliado para esse tipo de interação, podendo influenciar os resultados obtidos conforme mostrado na Figura 4.4 do Capítulo anterior. Os resultados indicam que o modelo pode ser adaptado para esse pdrão de cenário. As regras geradas pelos outros métodos estão disponíveis nas tabelas: Tabela 6.29, Tabela 6.30 e Tabela 6.31.

6.2.5 Conjuntos de dados com loci sem efeito principal

Nesta seção são apresentados os resultados em conjuntos de dados que apresentam interações epistáticas sem efeito principal. Para isso, foram replicados os dados utilizados em (VELEZ et al., 2007). Os conjuntos de dados são compostos de 70 modelos com funções de penetrância específicas e definidas no trabalho citado. Além disso, as taxas de herdabilidade e MAF variam para cada grupo de modelo. As taxas de herdabilidade variam de 0.4 até 0.01 com MAF de 0.4 e 0.2. Os parâmetros de cada função de penetrância e cada grupo de modelos podem ser vistos no Apêndice B. Para cada conjunto, foram geradas 5 replicações e cada uma foi executada 10 vezes por cada algoritmo (MDR e o xGPi), onde cada conjunto de dados possui 1000 marcadores e composta de 1600 indivíduos, sendo metade pertencente ao grupo de casos e metade pertencente ao grupo de controle. A Tabela 6.32 descreve as características de todos os 70 modelos utilizados no experimento.

Modelo	Conjuntos	h^2	MAF
1	00-04	0.4	0.2
2	05-09	0.4	0.4
3	10-14	0.3	0.2
4	15-19	0.3	0.4
5	20-24	0.2	0.2
6	25-29	0.2	0.4
7	30-34	0.1	0.2
8	35-39	0.1	0.4
9	40-44	0.05	0.2
10	45-49	0.05	0.4
11	50-59	0.025	0.2
12	55-59	0.025	0.4
13	60-64	0.01	0.2
14	65-69	0.01	0.4

Tabela 6.32: Descrição dos modelos que apresentam epistasia sem efeito principal utilizados no experimentos. Os modelos foram desenvolvidos em (VELEZ et al., 2007).

A Figura 6.15, Figura 6.16, Figura 6.17, Figura 6.18 e Figura 6.19 apresentam os resultados comparativos entre os métodos. Para esse experimento foram considerados somente os métodos MDR e xGPi, os outros modelos não foram utilizados devido as suas limitações. O MDR e o xGPi apresentam resultados consistentes até o modelo 8, como observado até a Figura 6.17. A partir do conjunto 40 é possível observar que o xGPi não

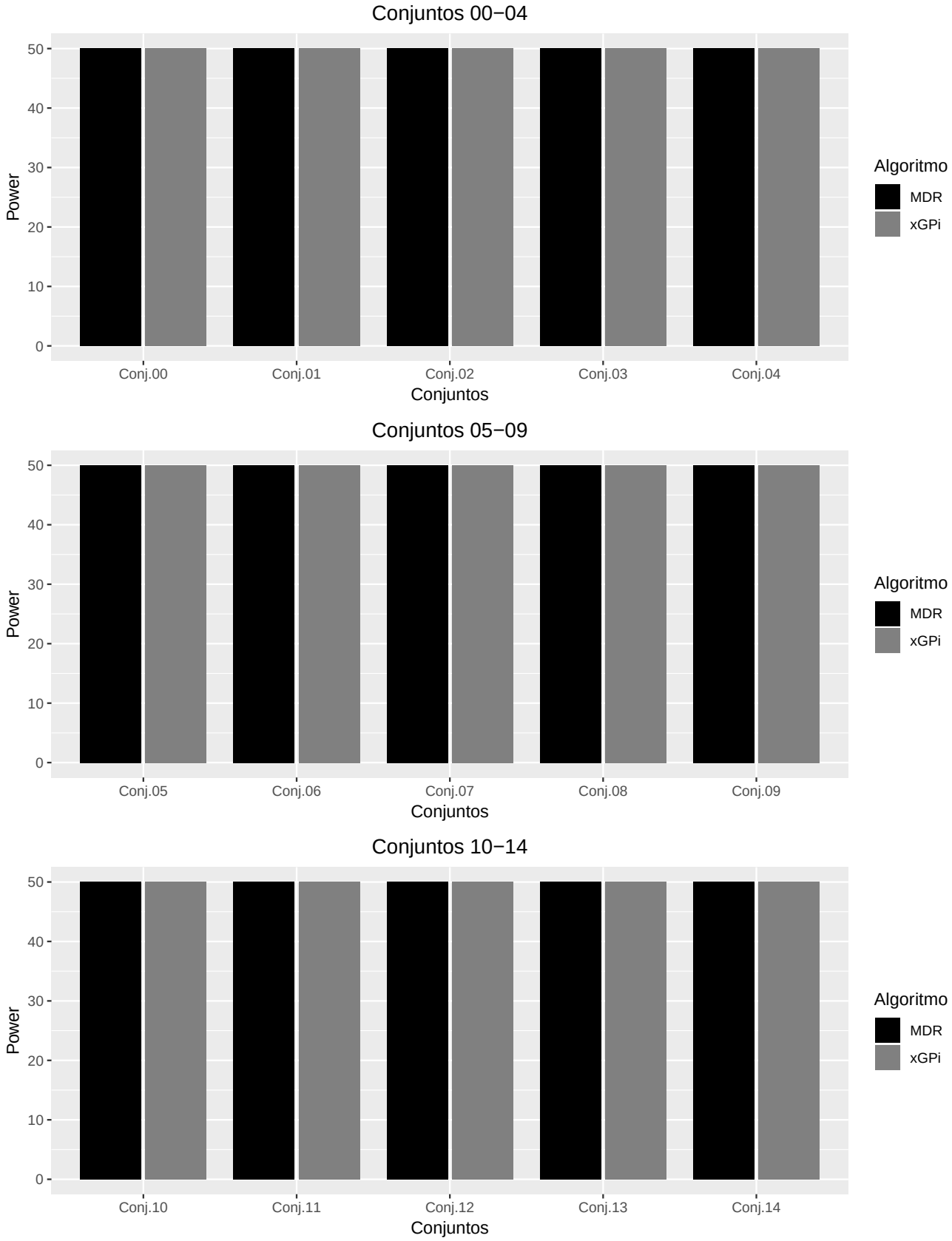


Figura 6.15: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 1000 marcadores e $h^2 = 0.2$.

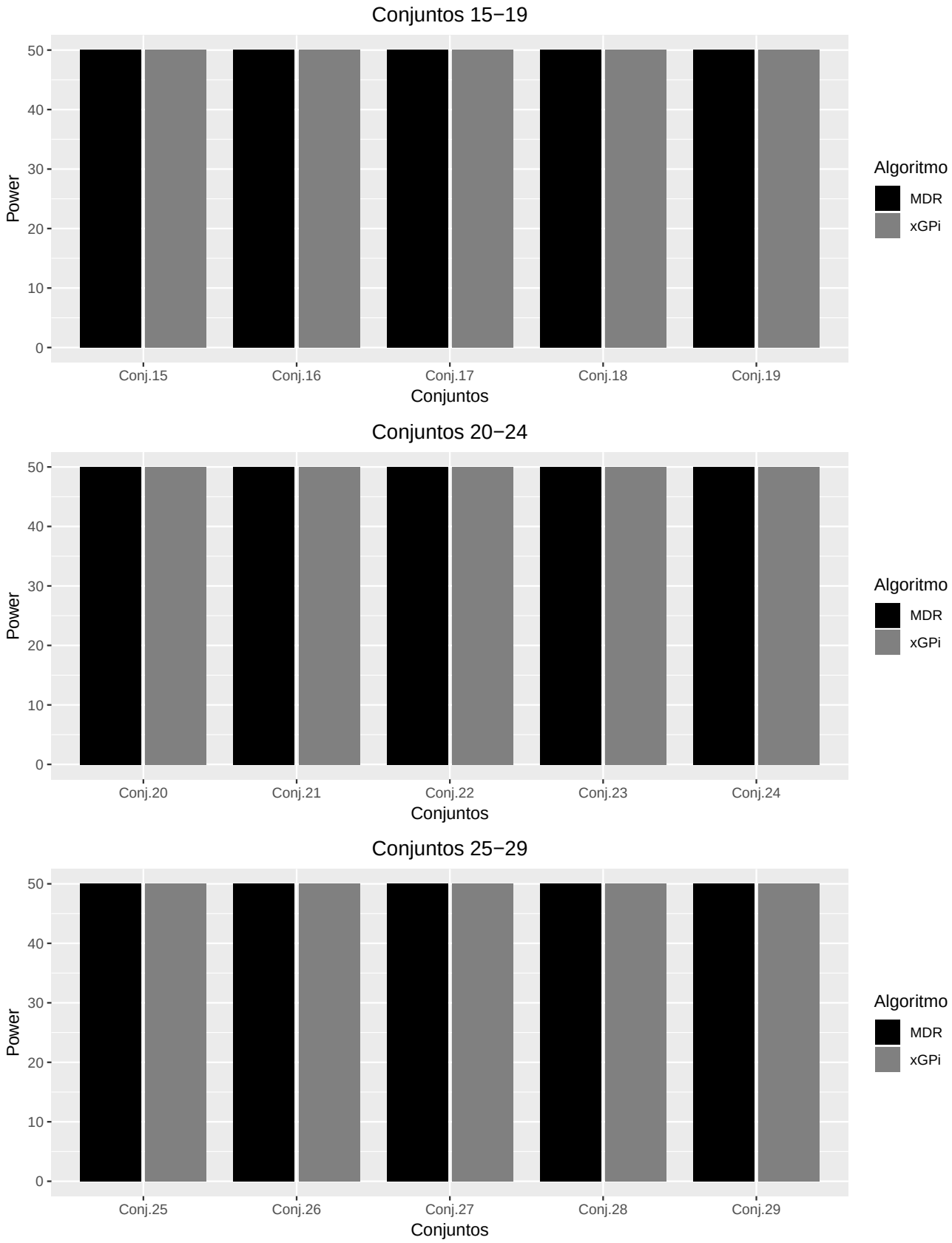


Figura 6.16: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 1000 marcadores e $h^2 = 0.2$.

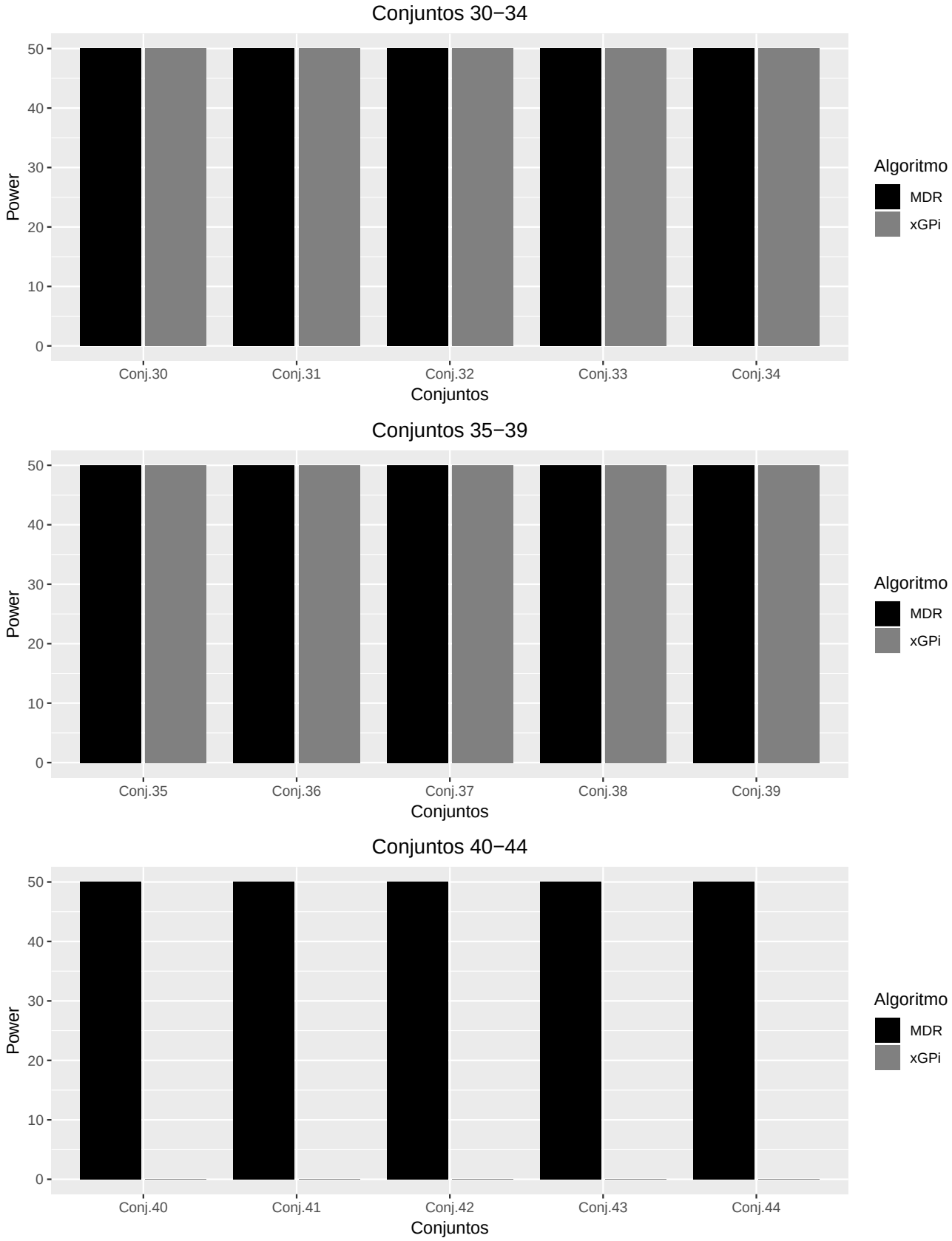


Figura 6.17: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 1000 marcadores e $h^2 = 0.2$.

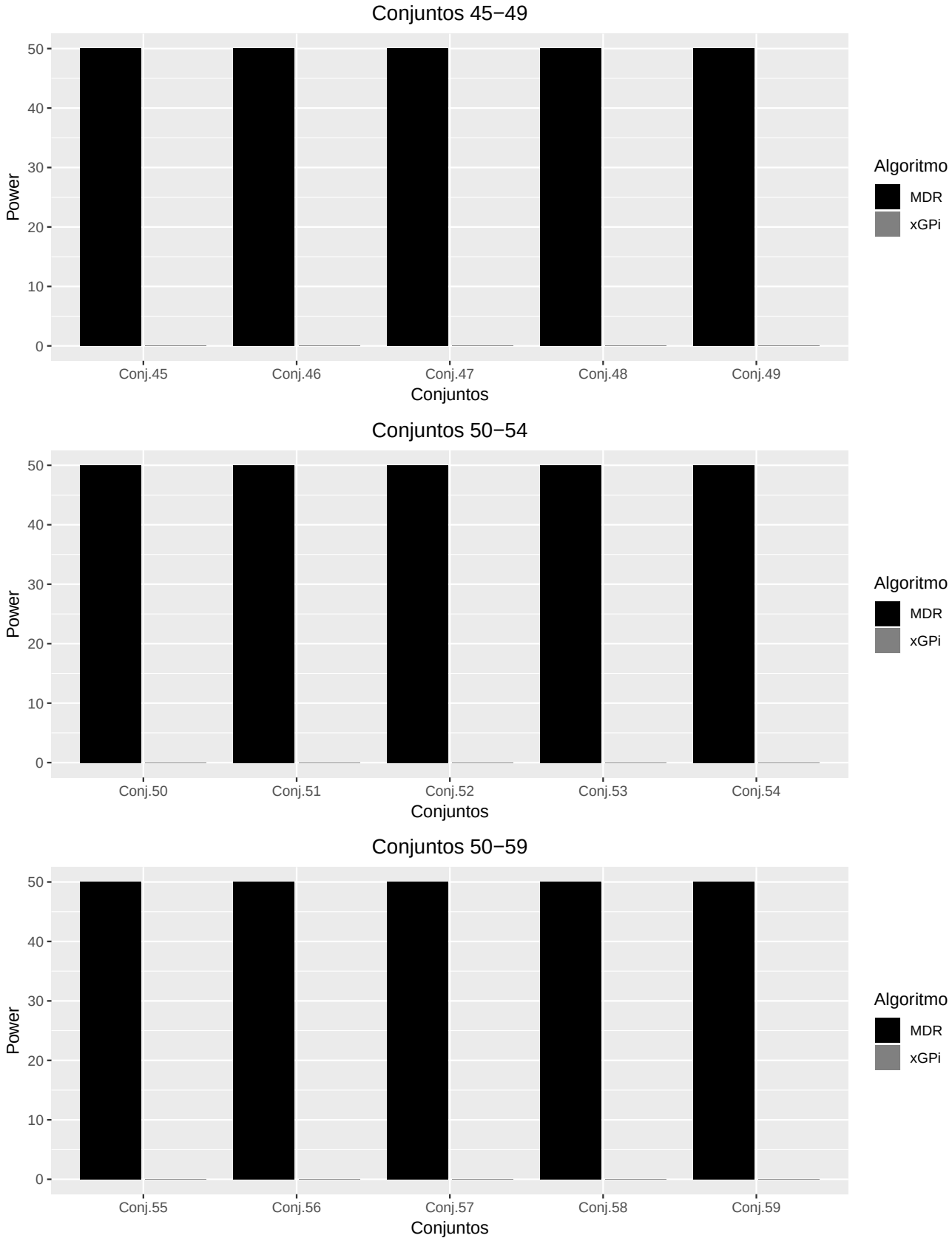


Figura 6.18: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 1000 marcadores e $h^2 = 0.2$.

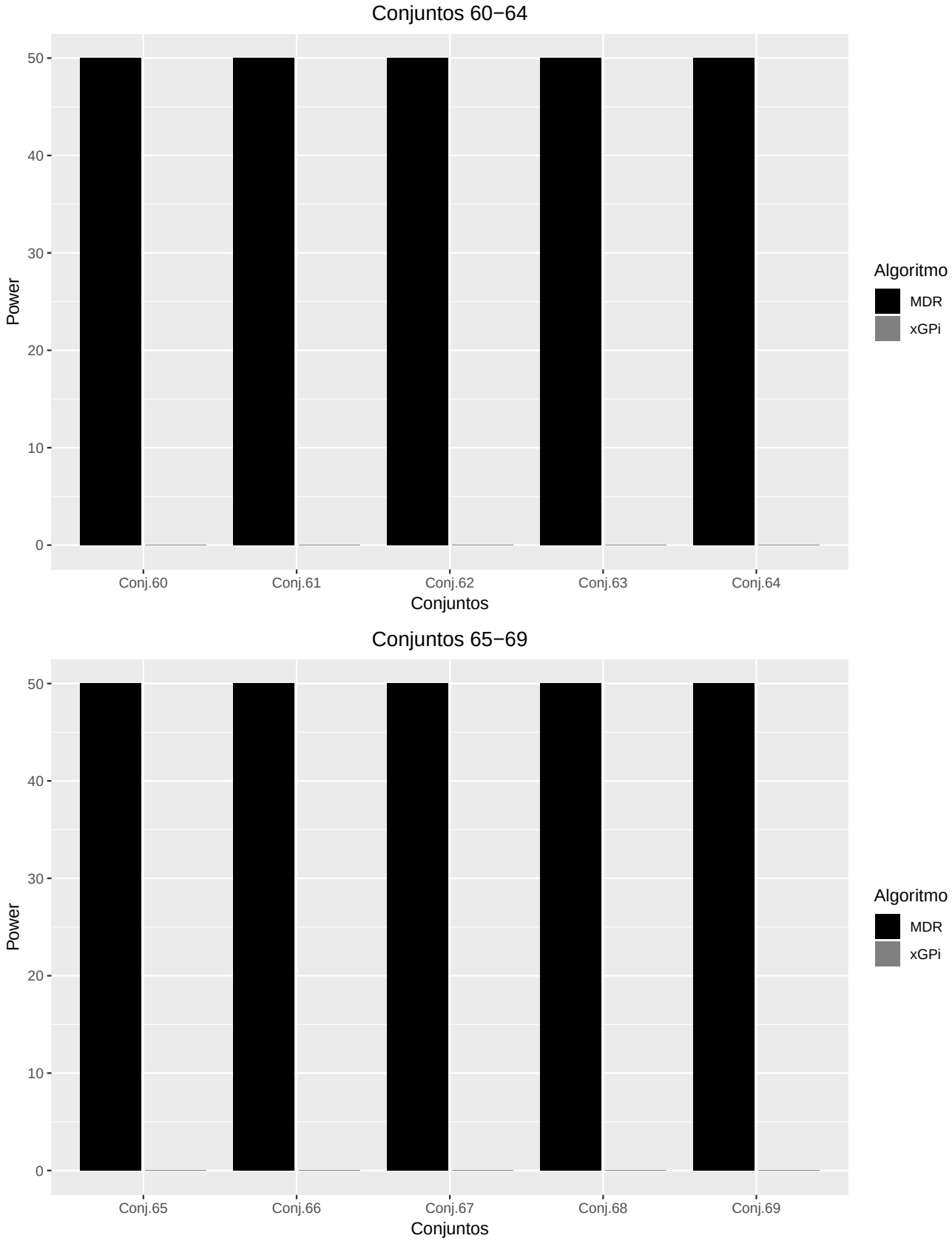


Figura 6.19: Gráfico representando o número de vezes em que os algoritmos identificaram corretamente os SNPS causais em conjuntos de dados com 1000 marcadores e $h^2 = 0.2$.

têm nenhum acerto. Isso indica que o método possui uma limitação para se trabalhar com dados de herdabilidade muito baixas. Pelos resultados o método limita-se a herdabilidade de 0.1, ao menos em cenários que apresentam epistasia sem efeito principal. Esse fato deve-se a influência de h^2 na taxa de penetrância em baixos valores, o que torna o processo de identificação da interação mais complexo.

O MDR obteve os melhores resultados independentemente dos níveis de h^2 e MAF apresentadas pelos modelos. O que já era esperado devido a construção do modelo que permite que cada par de marcadores presentes no conjunto de dados sejam avaliados.

6.2.6 Experimento com conjuntos de dados com 100 mil marcadores

Nesta seção são apresentados experimentos com conjunto de dados na ordem de 10^5 marcadores. O objetivo é analisar o desempenho do método proposto na detecção de interações epistáticas em cenários que apresentam uma grande quantidade de variáveis. Nesta seção foi utilizado somente o modelo proposto porque os demais métodos apresentam limitações para manipular essa quantidade de dados, com exceção do MDR. Entretanto, a utilização dele demandaria um esforço computacional inviável para fins de comparação. Inicialmente foi criado uma tabela de contingência da função de penetrância, dada pela Tabela 6.33. O conjunto de dados utilizado no experimento foi novamente gerado pelo simulador GAMETES. A herdabilidade do conjunto de dados é estimada em 0.4 e a MAF 0.4. Dessa forma, os alelos dos marcadores já estão em equilíbrio de Hardy-Weinberg (HWE) e SNPs em diferentes *loci* estão em desequilíbrio de ligação. A interação epistática é definida por dois marcadores. No caso pelos $SNP99999$ e $SNP100000$.

h2 = 0.4, MAF=0.4			
	AA	Aa	aa
BB	0.462	0.895	0.910
Bb	0.953	0.019	0.054
bb	0.443	0.988	0.469

Tabela 6.33: Função de penetrância dos marcadores de interesse do conjunto de dados com 100 mil marcadores.

Os marcadores foram separados em 5 cromossomos hipotéticos. Os valores-p de cada

marcador em cada cromossomo foram calculados e apresentados no gráfico de Manhattan na Figura 6.20. Pode-se observar que ambos os marcadores de interesse não apresentaram valores-p significativos. O marcador $SNP99999$ apresenta um valor-p de 0.02056 e o marcador $SNP100000$ um valor-p de 0.2722183. Isso indica que esses marcadores isolados apresentam baixa correlação com o fenótipo. É importante mencionar que depois da geração do conjunto dos dados e cálculo dos valores-p, todas as colunas foram aleatoriamente permutadas, dessa forma, permitindo que o $SNP99999$ e $SNP100000$ estejam em posições distantes um do outro. Dessa forma, o marcador $SNP99999$ ficou na posição 8.650 e o marcador $SNP100000$ na posição 93.232, o que define o primeiro marcador de interesse no primeiro cromossomo e o segundo no último cromossomo hipotético.

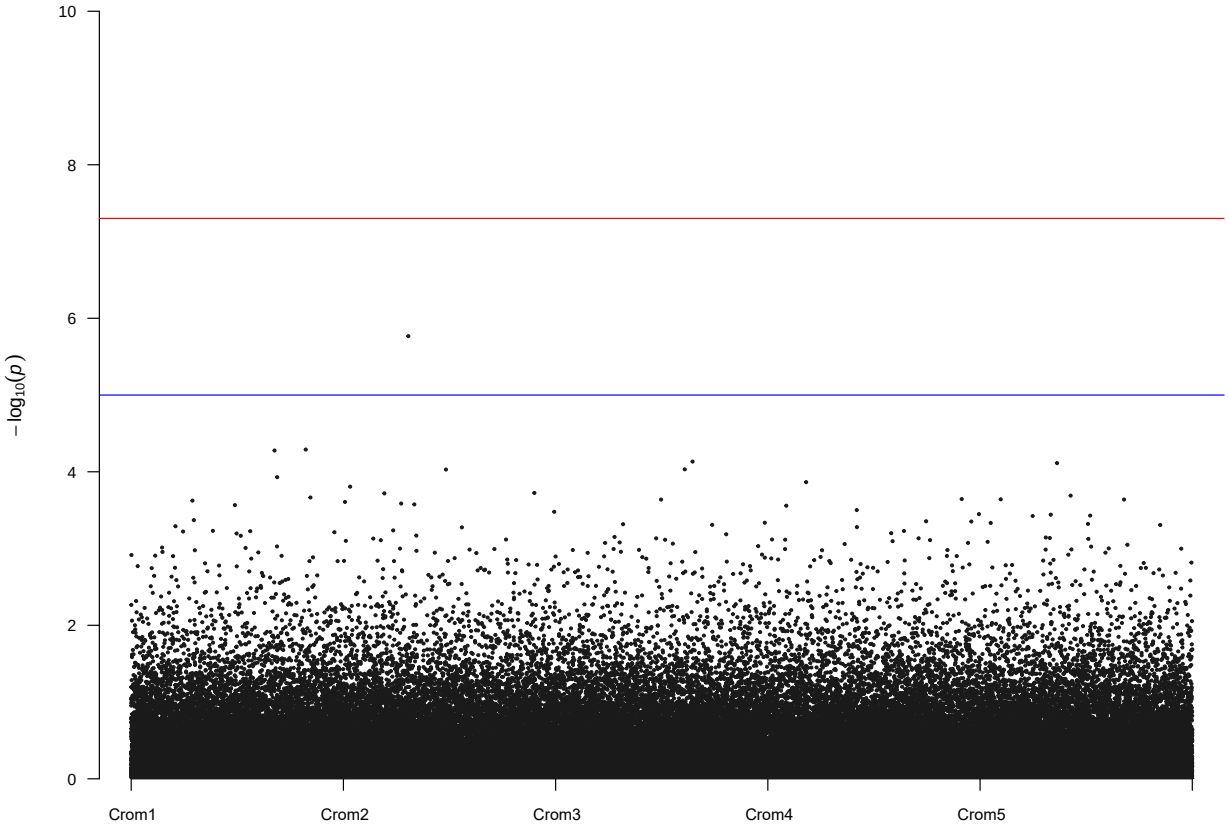


Figura 6.20: Gráfico de Manhattan do conjunto de dados. Os marcadores foram divididos e alocados em 5 diferentes cromossomos hipotéticos.

O conjunto de dados de 100 mil marcadores foi particionado em subconjuntos de 50 SNPs cada. Foram gerados um total de 2000 subgrupos. Após a geração de subgrupos foi calculado o número total de combinações, no caso 1999000. O algoritmo *XGBoost* foi exe-

cutado sobre cada combinação. Ao final do processo, a combinação que apresenta a maior área sobre a curva ROC, resultado da combinação dos seguintes subgrupos de marcadores foi selecionada. O modelo é capaz de selecionar diversas combinações, entretanto, nesse caso, foi verificado que a melhor combinação apresentava os SNPs de interesse, então somente essa combinação foi selecionada. Os marcadores são demonstrados na Tabela 6.34 e Tabela 6.35. Pode-se observar que o marcador de interesse $SNP99999$ está na posição 10 do subgrupo 1 e o marcador de interesse $SNP100000$ está na posição 15 do segundo subgrupo.

SNPs subgrupo 1
SNP34406,SNP67377,SNP21121,SNP52984,SNP93105,SNP65294,SNP21347
SNP93891,SNP23318,SNP99999,SNP97453,SNP57579,SNP37503,SNP38139,
SNP99949,SNP74194,SNP87799,SNP8990,SNP74172,SNP69374,SNP71481,
SNP31964,SNP82261,SNP76797,SNP67838,SNP65104,SNP39113,SNP17292,
SNP70398,SNP96570,SNP12453,SNP23250,SNP88958,SNP80679,SNP28909,
SNP21262,SNP97113,SNP3061,SNP31567,SNP37499,SNP96077,SNP30765,
SNP74166,SNP59924,SNP11326,SNP27033,SNP11459,SNP4049,SNP69659,SNP37364

Tabela 6.34: SNPs presentes no primeiro subgrupo da combinação selecionada que obteve maior área sobre a curva ROC. O marcador de interesse $SNP99999$ é dado em vermelho e está na posição 10 do subgrupo.

SNPs subgrupo 2
SNP98181,SNP95417,SNP38924,SNP28365,SNP91795,SNP14582,SNP61411
SNP80598,SNP54980,SNP27547,SNP47828,SNP16910,SNP97206,SNP35879
SNP100000,SNP20176,SNP49867,SNP23436,SNP74600,SNP67295,SNP7771
SNP84931,SNP96294,SNP87793,SNP91179,SNP75551,SNP73387,SNP68685
SNP21138,SNP95096,SNP56829,SNP46763,SNP56320,SNP89156,SNP29047
SNP40299,SNP91393,SNP31798,SNP65213,SNP63216,SNP25243,SNP87246
SNP19035,SNP70640,SNP60967,SNP23051,SNP84120,SNP40734,SNP86273,SNP30237

Tabela 6.35: SNPs presentes no segundo subgrupo da combinação selecionada que obteve maior área sobre a curva ROC. O marcador de interesse $SNP100000$ é dado em vermelho e está na posição 15 do subgrupo.

Após o processo de seleção de marcadores, é calculada a importância de variável sobre a combinação constituída dos dois subgrupos apresentados, ou seja, o pVI de cada marcador é calculado. A Figura 6.21 mostra um gráfico com o comparativo das 30 melhores variáveis ranqueadas de acordo com a métrica. Pode-se observar que as variáveis de interesse foram as melhores ranqueadas no processo.

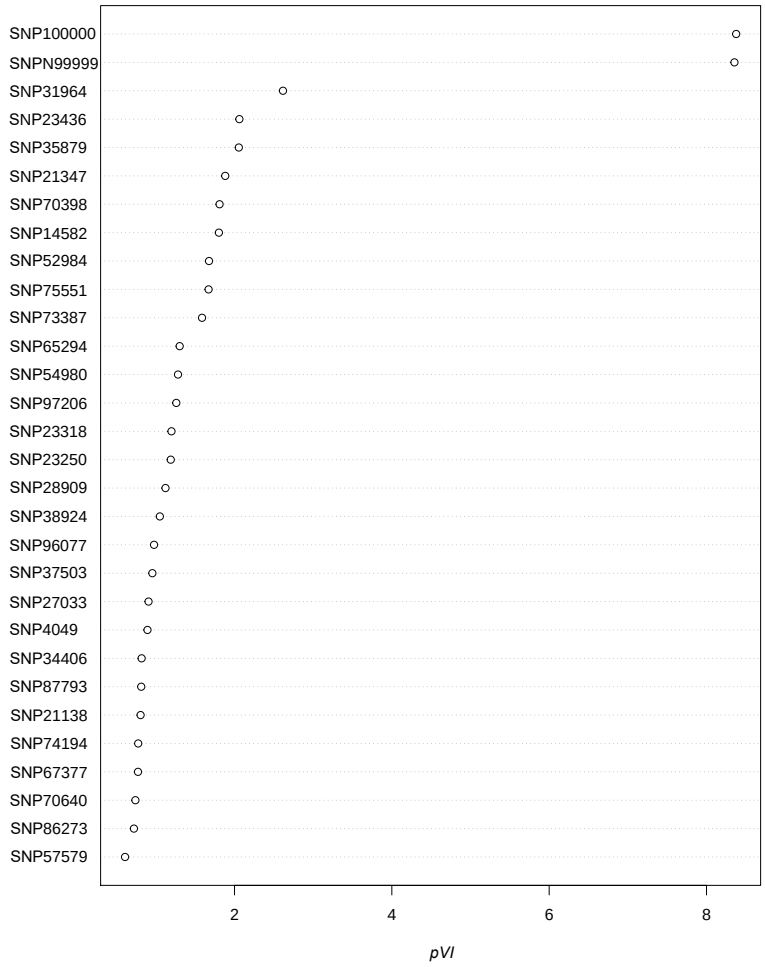


Figura 6.21: Representação da importância de variável pVI calculado pela floresta randômica. A Figura mostra o pVI de cada marcador presente na combinação que obteve a maior área sobre a curva ROC pelo algoritmo *XGBoost*.

Em seguida, baseado nos resultados obtidos na etapa de ranqueamento, a população inicial do algoritmo de GP foi gerada. Os parâmetros do algoritmo de GP permaneceram os mesmos apresentados no capítulo anterior. O algoritmo de GP foi executado 10 vezes e os resultados das regras obtidas das interações podem ser encontrados na Tabela 6.36.

Pode-se observar pelos resultados obtidos que o modelo proposto conseguiu encontrar a interação entre os marcadores causais em todas as execuções, indicando que o algoritmo pode ser utilizado em conjuntos de dados com uma grande quantidade de marcadores. Pode-se observar que etapa de maior custo computacional do modelo é a primeira etapa, responsável pela de seleção de subgrupos, que neste experimento exigiu uma demanda de $\frac{2000!}{2!(2000-2)!} = 1999000$ classificadores *XGBoost* para avaliar todas as combinações possíveis dos 20000 subgrupos de 50 SNPs cada. Entretanto, se a mesma avaliação fosse

Resultados do xGPi
(SNP100000 = 0 OU SNP99999 = 0) E (SNP100000 = 1 OU SNP99999 != 0)
SNP99999 = 1 OU (SNP100000 = 1 E (SNP99999 = 0 E SNP100000 != 0))
((SNP99999 != 1 E SNP100000 = 1) OU (SNP100000 != 1 E SNP99999 = 1)) E SNP99999 != 2
(SNP69374 = 1 OU SNP99999 != 1) E (SNP99999 = 1 OU (SNP99999 != 2 E SNP100000 = 1))
SNP100000 = 1 OU ((SNP99999 = 1 E SNP100000 != 1) OU SNP99999 != 0)
(((SNP99999 != 0 OU SNP100000 = 1) E SNP100000 != 0) OU SNP99999 != 0) E SNP99999 != 2
(SNP100000 != 1 E SNP99999 = 1) OU ((SNP99999 != 2 E SNP100000 = 1) E SNP99999 = 0)
((SNP99999 = 0 E SNP100000 = 1) OU (SNP99999 != 0 E SNP30237 != 2)) E SNP99999 != 2
((SNP100000 = 1 OU SNP99999 != 0) OU SNP27547 = 2) E (SNP99999 = 0 OU SNP100000 != 1)
(SNP100000 != 1 E SNP99999 != 0) OU ((SNP100000 != 0 E SNP99999 = 0) E SNP100000 = 1)

Tabela 6.36: Regras de associação geradas como resultado do algoritmo de GP sobre o conjunto de dados com 100 mil marcadores.

realizada pelo MDR, seriam necessários $100000^2 \times 100000^2 = 10000000000$ classificadores para a avaliar e determinar a interação, ou seja, o custo computacional seria quadrático em relação tamanho no número de marcadores presentes no conjunto de dados, o que torna o processo de execução inviável, para o equipamento computacional disponível.

Essa seção foi apresentada para avaliar se a metodologia proposta pode ser promissora e efetiva na análise de dados em larga escala. O objetivo foi simular *chips* de densidade capazes de cobrirem uma grande quantidade de SNPs, assemelhando aos cenários encontrados em dados biológicos reais.

7 Conclusões e trabalhos futuros

A identificação e caracterização de marcadores envolvidos em interações epistáticas consiste de uma das principais tarefas em problemas de GWAS. A descoberta de marcadores relevantes enfrenta obstáculos por diversos motivos, tais como a herdabilidade, MAF, quantidade de marcadores e cardinalidade do conjunto de dados.

O principal objetivo deste trabalho foi o desenvolvimento de um modelo capaz de identificar as interações e relações entre marcadores relevantes envolvidos em interações que dão origem a doenças complexas. A metodologia desenvolvida é baseada em um *workflow* construído por meio de etapas complementares visando contornar em cada uma delas as diversas complexidades inerentes ao problema visando oferecer uma forma compreensível de interpretação dos resultados ao final do processo de execução. Foram apresentados resultados obtidos sobre conjuntos de dados simulados com a devida avaliação e discussão do desempenho em relação aos principais algoritmos disponíveis e mais utilizados no âmbito das pesquisas em GWAS, principalmente em relação ao método MDR, tido como referência e estado da arte na área, o que proporcionou que a aplicabilidade do modelo proposto fosse avaliada em diversos cenários.

7.1 Contribuições do método proposto

O foco da tese foi em estudos de GWAs. Dessa forma, a contribuição do trabalho foi o desenvolvimento de uma metodologia capaz de lidar com conjuntos de dados com quantidades significativas de marcadores para a análise de interações epistáticas de segunda ordem, podendo ser estendida para ordens maiores, com diversos níveis de herdabilidade e MAF.

A metodologia proposta neste trabalho, em sua primeira etapa, apresenta uma abordagem inovadora que difere-se conceitualmente dos demais métodos na literatura, principalmente pela implementação de um procedimento baseado em subgrupos que viabilizou a escalabilidade do modelo, bem como o aumento do potencial de detecção de marcadores causais, epistáticos ou não, através da construção de subconjuntos de marcadores que apresentam alta probabilidade dos marcadores de interesse pertencerem, não só permitindo

como viabilizando o processo de refinamento das etapas posteriores, mais específicas no processo de detecção das interações. Na sequência, a segunda etapa da metodologia inicia o processo de refinamento dos marcadores selecionados na etapa anterior construindo um procedimento hierárquico baseado em florestas randômicas, onde a utilização da medida de importância de variável "pVI"se mostrou eficaz para ranquear marcadores de interesse presentes nos subgrupos selecionados na primeira etapa. Essas duas primeiras etapas contemplam grandes desafios em estudos de GWAS, descritos no Capítulo 1. Finalmente, o terceiro grande desafio é a interpretabilidade. Nesse contexto, baseou-se a construção da etapa em um modelo de computação evolucionista, ou seja, populacional. Esta possibilidade tornou-se viável graças aos processos de redução e ranqueamento executados pelas etapas anteriores, visto que algoritmos evolutivos, apesar de serem bastante eficientes em otimização global, demandam alto custo computacional. Visando obter um elevado nível de interpretabilidade das interações detectadas, o modelo a etapa foi habilmente construído baseando-se em GP, que apresenta potencial para busca de marcadores associada ao problema caso-controle, permitindo um elevado nível de interpretabilidade das soluções obtidas. As regras geradas pelo algoritmo de GP podem ser facilmente interpretadas, mesmo sem a necessidade de conhecimentos específicos em matemática e métodos estatísticos.

Todas as três etapas são críticas para o correto funcionamento da metodologia. Entretanto, como verificado nos experimentos realizados, alguns parâmetros devem ser adequadamente definidos para que o modelo, em suas etapas, obtenha sucesso na identificação das interações de interesse. Um parâmetro que requer bastante atenção, devendo ser cuidadosamente avaliado diz respeito a relação entre herdabilidade e o tamanho dos subgrupos na primeira etapa da metodologia. Além disso, foi possível verificar que o método limita-se a identificação de interações epistáticas sem efeito principal com $h^2 \geq 0.1$, sendo sua utilização limitada à conjuntos de dados do tipo caso-controle.

Finalizando, é importante ressaltar que um dos indicativos de maior relevância do modelo desenvolvido, com indícios promissores nos experimentos realizados é o potencial da metodologia para utilização em larga escala.

7.2 Trabalhos futuros

Como foi mencionado, a metodologia proposta foi desenvolvida para trabalhar com dados do tipo caso-controle, ou seja, para fenótipos discretos, dicotômicos e binários. Entretanto o método pode ser modificado para trabalhar com problemas que possuem um número maior de classes. Essa adaptação é viável de ser realizada porque em todas as etapas, os algoritmos utilizados permitem essa modificação.

A abordagem poderia ser estendida para trabalhar com diferentes tipos de variáveis discretas, como por exemplo variáveis ambientais. Assim, as regras definidas no conjunto função e terminal da GP podem ser modificadas e modeladas de acordo com problemas específicos.

Através dos experimentos com *loci* sem efeito principal, foi possível examinar o método em cenários de herdabilidade muito baixas. O método apresentou-se eficiente em casos onde os níveis de herdabilidade são próximos de 0.1. Neste contexto, faz-se necessário identificar questões relativas ao limite da predição em relação ao nível de herdabilidade, bem como a relação de dependência entre o tamanho das permutações e o poder preditivo do modelo na etapa de seleção de subgrupos.

O método ainda precisa ser validado em conjuntos de dados do mundo real para que se comprove a sua eficiência na descoberta de associações. Além disso, como cada etapa é desacoplada, elas podem ser utilizados para otimizar a busca por marcadores em outras metodologias propostas na literatura, sendo uma abordagem flexível nesse sentido.

E importante ressaltar que, todas as três etapas, incluindo interfaces, que compõem a abordagem desenvolvida precisam de aperfeiçoamentos para que se tenha uma ferramenta mais robusta e confiável, principalmente devido a complexidade do problema tratado.

Porém, talvez a questão mais relevante e de maior impacto a ser tratada no modelo esteja relacionada a verificação de seu limite no que tange a dimensão da base avaliada, em relação ao número de amostras mas, principalmente, em relação ao número de marcadores disponibilizados. Entende-se que um método que não considera o desafio de avaliar bases de dados em escala genômica não atenderá as demandas futuras. No modelo apresentado, esta tarefa complexa é basicamente responsabilidade da primeira etapa. A experiência no desenvolvimento desta etapa indicou que a eficiência obtida com o algoritmo *XGBoost*, base da etapa, não foi conquistada de forma simples. Cada componente renova-se a expectativa de evolução do modelo baseando-se no conhecimento refinado da ferramenta

base e, principalmente do problema de interesse.

REFERÊNCIAS

AJF, G. et al. *An Introduction to Genetic Analysis*. 7nd. ed. [S.l.]: New York:W. H. Freeman, 2000.

ALMGREN, P. et al. *Statistic in Genetics*. [S.l.: s.n.], 2003.

ARBEX, W. A. *Modelos Computacionais para Identificação de Informação Genômica Associada à Resistência ao Carrapato Bovino*. Tese (Doutorado) — UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2009.

ARDLIE, K.; KRUGLYAK, L.; SEIELSTAD, M. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, v. 3, n. 4, p. 299–309, 2002.

BALA, A.; CHANA, D. Article: A survey of various workflow scheduling algorithms in cloud environment. *IJCA Proceedings on 2nd National Conference on Information and Communication Technology*, NCICT, n. 4, p. 26–30, November 2011.

BATERSON, W. The progress of genetics since the rediscovery of mendel?s paper. *Progressus Rei Botanicae*, n. 1, p. 368–382, 1907.

BLEULER, S. et al. Multiobjective Genetic Programming: Reducing Bloat by Using SPEA2. In: *Congress on Evolutionary Computation (CEC 2001)*. Piscataway, NJ: IEEE, 2001. p. 536–543.

BREIMAN, L. Bagging predictors. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 24, n. 2, p. 123–140, ago. 1996. ISSN 0885-6125.

BREIMAN, L. Random forests. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 45, n. 1, p. 5–32, out. 2001. ISSN 0885-6125.

BROOKES, A. J. The essence of snps. *Gene*, v. 234, n. 2, p. 177 – 186, 1999.

BROWN, T. A. *Genomes*. [S.l.]: Garland science, 2006.

BUSH, W. S.; MOORE, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*, Public Library of Science, v. 8, n. 12, p. e1002822, 12 2012.

CAETANO, A. Marcadores snp: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. *Revista Brasileira de Zootecnia*, v. 32, n. 8, p. 64–71, 2009.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016. (KDD '16), p. 785–794.

CHEN, T.; HE, T. Higgs Boson Discovery with Boosted Trees. In: COWAN, G. et al. (Ed.). *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning*. Montreal, Canada: PMLR, 2015. (Proceedings of Machine Learning Research, v. 42), p. 69–80.

CHESSA, S. et al. Development of a single nucleotide polymorphism genotyping microarray platform for the identification of bovine milk protein genetic polymorphisms. *Journal of Dairy Science*, American Dairy Science Association, v. 90, n. 1, p. 451–464, jan. 2007.

CLARK, P. J.; KEMPTHORNE, O. An introduction to genetic statistics. *Journal of Mammalogy*, Oxford University Press (OUP), v. 39, n. 2, p. 313, maio 1958.

CORDELL, H. J. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, Springer Science and Business Media LLC, v. 10, n. 6, p. 392–404, jun. 2009.

CZEN, K.; LICHTENSTEIN, P.; HEMMINKI, K. Environmental and heritable causes of cancer among 9.6 million individuals in the swedish family-cancer database. *Int J Cancer*, Maio 2002.

CZENE, K.; LICHTENSTEIN, P.; HEMMINKI, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int. J. Cancer*, v. 99, n. 2, p. 260–266, May 2002.

DORIGO, M.; GAMBARDELLA, L. M. Ant colonies for the travelling salesman problem. *Biosystems*, Elsevier BV, v. 43, n. 2, p. 73–81, jul 1997.

EASTON, D. F.; EELES, R. A. Genome-wide association studies in cancer. *Human Molecular Genetics*, Oxford University Press (OUP), v. 17, n. R2, p. R109–R115, oct 2008.

ESTRADA-GIL, J. K. et al. GPDTI: a Genetic Programming Decision Tree induction method to find epistatic effects in common complex diseases. *Bioinformatics (Oxford, England)*, v. 23, n. 13, p. i167–74, 2007. ISSN 1367-4811.

EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, v. 8, n. 3, p. 186–194, Mar 1998.

FACELI, K. et al. Inteligência artificial: Uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, 2011.

FERREIRA, C. Programação de expressão genética: um novo algoritmo adaptável para resolver problemas. In: . [S.l.: s.n.], 2001.

FOULKES, A. *Applied Statistical Genetics with R: For Population-based Association Studies*. [S.l.]: Springer New York, 2009. (Use R!). ISBN 9780387895543.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, Academic Press, Inc., Orlando, FL, USA, v. 55, n. 1, p. 119–139, ago. 1997. ISSN 0022-0000.

FREUND, Y.; SCHAPIRE, R. E. A short introduction to boosting. In: *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. [S.l.]: Morgan Kaufmann, 1999. p. 1401–1406.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, v. 29, p. 1189–1232, 2000.

FRIEDMAN, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 38, n. 4, p. 367–378, fev. 2002. ISSN 0167-9473.

GOLDSTEIN, B. A. et al. An application of random forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genetics*, Springer Nature, v. 11, n. 1, p. 49, 2010.

GOLDSTEIN, B. A.; POLLEY, E. C.; BRIGGS, F. B. S. Random forests for genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, Walter de Gruyter GmbH, v. 10, n. 1, jan 2011.

GONDRO, C.; WERF, J. van der; HAYES, B. (Ed.). *Genome-Wide Association Studies and Genomic Prediction*. [S.l.]: Humana Press, 2013.

GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, Springer Science and Business Media LLC, v. 17, n. 6, p. 333–351, maio 2016.

GORDON, D.; ABAJIAN, C.; GREEN, P. Consed:a graphical tool for sequence finishing. *Genome Research*, Cold Spring Harbor Laboratory, v. 8, n. 3, p. 195–202, mar. 1998.

GOUDEY, B. et al. GWIS - model-free, fast and exhaustive search for epistatic interactions in case-control GWAS. *BMC Genomics*, Springer Nature, v. 14, n. Suppl 3, p. S10, 2013.

GREEN, P. Phrap, version 1.090518. v. 38, n. 6, p. 1767–1771, 2009.

GRIFFITHS, A. J. *Introdução à genética*. [S.l.]: Guanabara Koogan, 2008.

GU, J.; WU, X. Genetic susceptibility to bladder cancer risk and outcome. *Per Med*, v. 8, n. 3, p. 365–374, May 2011.

GUIMARÃES, P.; COSTA, M. Snps: sutis diferenças de um código. *Biotecnol. Cienc. Desenvolv*, v. 26, p. 24–27, 2002.

HAN, B. et al. Genetic studies of complex human diseases: Characterizing SNP-disease associations using bayesian networks. *BMC Systems Biology*, Springer Nature, v. 6, n. Suppl 3, p. S14, 2012.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. [S.l.]: Springer, 2016. ISBN 0387848576.

HEATON, M. P. et al. Use of bovine single nucleotide polymorphism markers to verify sample tracking in beef processing. *Journal of the American Veterinary Medical Association*, American Veterinary Medical Association (AVMA), v. 226, n. 8, p. 1311–1314, abr. 2005.

HOFFEE, P. *Genética Médica Molecular*. [S.l.]: Guanabara Koogan Edição, Oxford, Blackwell Science Limited, 2000.

IOANNIDIS, J. P. et al. Replication validity of genetic association studies. *Nature genetics*, Nature Publishing Group, v. 29, n. 3, p. 306–309, 2001.

KAMISKI, S. et al. MilkProtChip–a microarray of SNPs in candidate genes associated with milk protein biosynthesis–development and validation. *J. Appl. Genet.*, v. 46, n. 1, p. 45–58, 2005.

KINGSMORE, S. F. et al. Genome-wide association studies: progress and potential for drug discovery and development. *Nat Rev Drug Discov*, Nature Publishing Group, v. 7, n. 3, p. 221–230, February 2008.

KONONENKO, I. Estimating attributes: Analysis and extensions of relief. In: . [S.l.]: Springer Verlag, 1994. p. 171–182.

KOZA, J. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. [S.l.]: Bradford, 1992. (A Bradford book). ISBN 9780262111706.

KOZA, J. R. *Genetic Programming II: Automatic Discovery of Reusable Programs*. Cambridge Massachusetts: MIT Press, 1994. ISBN 0-262-11189-6.

LIAROKAPIS, M. V. et al. A learning scheme for reach to grasp movements: On EMG-based interfaces using task specific motion decoding models. *IEEE Journal of Biomedical and Health Informatics*, Institute of Electrical and Electronics Engineers (IEEE), v. 17, n. 5, p. 915–921, set. 2013.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. *R News*, v. 2, n. 3, p. 18–22, 2002.

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2, n. 3, p. 18–22, 2002.

LISTED, N. authors. A haplotype map of the human genome. *Nature*, v. 437, n. 7063, p. 1299–1320, Oct 2005.

LOUDEN, K. *Compiler construction : principles and practice*. Boston: PWS Pub. Co, 1997. ISBN 0-534-93972-4.

LUKE, S. et al. *ECJ 16: A Java-based Evolutionary Computation Research System*. http://cs.gmu.edu/~eclab/projects/ecj/: [s.n.], 2007.

MACARTHUR, J. et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Research*, Oxford University Press (OUP), v. 45, n. D1, p. D896–D901, nov 2016.

MANOLIO, T. A. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, New England Journal of Medicine (NEJM/MMS), v. 363, n. 2, p. 166–176, jul 2010.

MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072.

MITTAG, F. et al. Use of support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities. *Human Mutation*, Wiley, v. 33, n. 12, p. 1708–1718, aug 2012.

MOORE, J.; WHITE, B. Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge. *Genetic Programming Theory and Practice IV*, p. 11–28, 2007.

MOORE, J. H. The challenges of whole-genome approaches to common diseases. *JAMA: The Journal of the American Medical Association*, American Medical Association (AMA), v. 291, n. 13, p. 1642–1643, apr 2004.

MOORE, J. H. Epistasis analysis using ReliefF. In: *Methods in Molecular Biology*. [S.l.]: Springer New York, 2014. p. 315–325.

MOORE, J. H.; ANDREWS, P. C. Epistasis analysis using multifactor dimensionality reduction. In: *Methods in Molecular Biology*. [S.l.]: Springer New York, 2014. p. 301–314.

MOORE, J. H.; ASSELBERGS, F. W.; WILLIAMS, S. M. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, Oxford University Press (OUP), v. 26, n. 4, p. 445–455, jan 2010.

MOORE, J. H. et al. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*, Elsevier BV, v. 241, n. 2, p. 252–261, jul 2006.

MOORE, J. H.; WHITE, B. C. Tuning ReliefF for genome-wide genetic analysis. In: *Lecture Notes in Computer Science*. [S.l.]: Springer Berlin Heidelberg, 2007. p. 166–175.

MOORE, J. H.; WILLIAMS, S. M. Epistasis and its implications for personal genetics. *The American Journal of Human Genetics*, Elsevier BV, v. 85, n. 3, p. 309–320, sep 2009.

MORELAND, K.; TRUEMPER, K. The needles-in-haystack problem. In: *Machine Learning and Data Mining in Pattern Recognition*. [S.l.]: Springer Berlin Heidelberg, 2009. p. 516–524.

MüHLENBEIN, H.; SCHLIERKAMP-VOOSEN, D. The science of breeding and its application to the breeder genetic algorithm (BGA). *Evolutionary Computation*, v. 1, n. 4, p. 335–360, 1993.

NEALE, B. M. et al. Genome-wide association scan of attention deficit hyperactivity disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, Wiley, v. 147B, n. 8, p. 1337–1344, dez. 2008.

NIEL, C. et al. A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, Frontiers Media SA, v. 6, set. 2015.

NUNKESSER, R. et al. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics*, v. 23, n. 24, p. 3280–3288, 2007. ISSN 1367-4803.

OLAZAR, M. R. R. *Uma Metodologia para a Descoberta de Marcadores Genéticos em Estudos de Associação*. Tese (Doutorado) — UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, Maio 2013.

OLIVEIRA, C. F. *Um método para seleção de atributos em dados genômicos*. Tese (Doutorado) — Universidade Federal de Juiz de Fora, 2015.

OLSON, R. S.; MOORE, J. H. Tpot: A tree-based pipeline optimization tool for automating machine learning. In: HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. (Ed.). *Proceedings of the Workshop on Automatic Machine Learning*. New York, New York, USA: PMLR, 2016. (Proceedings of Machine Learning Research, v. 64), p. 66–74.

OLTEAN, M. Evolving evolutionary algorithms using linear genetic programming. *Evolutionary Computation*, v. 13, p. 2005, 2005.

PEARSON, H. What is a gene? *Nature*, Springer Nature, v. 441, n. 7092, p. 398–401, may 2006.

PFEIFER, S. P. From next-generation resequencing reads to a high-quality variant data set. *Heredity*, Springer Nature, v. 118, n. 2, p. 111–124, out. 2016.

PIERCE, B. *Genetics : a conceptual approach*. New York: W.H. Freeman, 2012. ISBN 1429232528.

PIERCE, B. A. *Genética. Um Enfoque Conceitual*. [S.l.]: Guanabara Koogan, 2011. ISBN 852771664X.

POLI, R.; LANGDON, W. B.; MCPHEE, N. F. *A Field Guide to Genetic Programming*. [S.l.]: Lulu Enterprises, UK Ltd, 2008. ISBN 1409200736, 9781409200734.

POLIKAR, R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, v. 6, n. 3, p. 21–45, 2006.

POULSEN, P. et al. Heritability of type ii (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance – a population-based twin study. *Diabetologia*, v. 42, n. 2, p. 139–145, 1999. ISSN 1432-0428.

RITCHIE, M. D. et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, Elsevier BV, v. 69, n. 1, p. 138–147, jul 2001.

ROBNIK-ŠIKONJA, M.; KONONENKO, I. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, Springer Nature, v. 53, n. 1/2, p. 23–69, 2003.

ROKACH, L.; MAIMON, O. *Data Mining with Decision Trees: Theroy and Applications*. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2008. ISBN 9789812771711, 9812771719.

SANGER, F.; NICKLEN, S.; COULSON, A. R. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, v. 74, n. 12, p. 5463–5467, 1977.

SCHAPIRE, R. E. *The Strength of Weak Learnability*. 1990.

SCHAPIRE, R. E. A brief introduction to boosting. In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. (IJCAI'99), p. 1401–1406.

SCHAPIRE, R. E.; FREUND, Y. *Boosting: Foundations and Algorithms*. [S.l.]: The MIT Press, 2012. ISBN 0262017180, 9780262017183.

SILVA, F. F. Seleção genômica no r. [s.I], 2013.

SOHN, A.; OLSON, R. S.; MOORE, J. H. Toward the automated analysis of complex diseases in genome-wide association studies using genetic programming. In: *Proceedings of the Genetic and Evolutionary Computation Conference*. New York, NY, USA: ACM, 2017. (GECCO '17), p. 489–496. ISBN 978-1-4503-4920-8.

STEEN, K. V. Travelling the world of gene-gene interactions. *Briefings in Bioinformatics*, Oxford University Press (OUP), v. 13, n. 1, p. 1–19, mar. 2011.

SZE-TO, H.-Y. et al. {GP-Pi}: Using Genetic Programming with Penalization and Initialization on Genome-Wide Association Study. *Artificial Intelligence and Soft Computing - 12th International Conference, {ICAISC} 2013, Zakopane, Poland, June 9-13, 2013, Proceedings, Part {II}*, v. 7895, p. 330–341, 2013.

TAN, H. et al. The estimation of heritability for twin data based on concordances of sex and disease. *Chronic Dis Can*, v. 26, n. 1, p. 9–12, 2005.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367.

Team. R.C.R,. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2008. ISBN 3-900051-07-0.

URBANOWICZ, R. J. et al. METHODOLOGY GAMETES : a fast , direct algorithm for generating pure , strict , epistatic models with random architectures. *BioData Mining 2012, 5:16*, v. 5, n. 16, p. 1–14, 2012. ISSN 1756-0381.

VALENTE, J. e. a. *Melhoramento genético de bovinos de leite*. [S.l.: s.n.], 2001.

VELEZ, D. R. et al. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetic Epidemiology*, Wiley, v. 31, n. 4, p. 306–315, maio 2007.

WAN, X. et al. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics*, Oxford University Press (OUP), v. 26, n. 1, p. 30–37, 2009.

WAN, X. et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, Elsevier BV, v. 87, n. 3, p. 325–340, set. 2010.

WANG, Y. et al. AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. *BMC Research Notes*, Springer Nature, v. 3, n. 1, p. 117, 2010.

WELTER, D. et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, Oxford Univ Press, v. 42, n. D1, p. D1001–D1006, 2014.

ZHANG, Y.; LIU, J. S. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, Springer Nature, v. 39, n. 9, p. 1167–1173, ago. 2007.

8 Apêndice A

Regras de associação geradas pelas diferentes propostas de função de avaliação apresentadas no Capítulo 4. Os resultados apresentados foram executados em conjuntos de dados com 100 marcadores. A interação epistática é dada pela combinação dos marcadores $N98$ e $N99$. A herdabilidade do conjunto de dados no experimento é de $h^2 = 0.1$ e a MAF 0.1.

i	Função adotada
1	(N99 != 1 Or N98 != 0) And (N99 != 0 Or N98 = 0)
2	((N99 != 1 Or N98 != 0) And N11 != 2) And (N99 = 1 Or N98 = 0)
3	((N98 = 0 Or N99 = 1) And (N98 != 0 Or N99 != 1))
4	(N98 != 0 And N99 != 0) Or (N99 != 1 And N98 = 0)
5	((N99 != 1 Or N98 != 0) And N99 = 1) Or (N99 != 1 And N98 = 0)
6	(N98 != 0 Or (N99 != 1 And N99 != 1)) And (N99 != 0 Or N98 != 1)
7	(N99 != 1 Or (N98 != 0 Or N99 = 2)) And (N99 = 1 Or N98 = 0)
8	((N99 != 1 Or N98 != 0) And N11 != 2) And (N98 = 0 Or N99 = 1)
9	(N98 = 0 Or N99 = 1) And (N99 != 1 Or N98 != 0)
10	(N98 != 1 And N99 != 1) Or (N99 != 2 And N98 != 0)

Tabela 8.1: Regras de associação geradas pela função de avaliação adotada pelo algoritmo de GP.

i	Precision
1	(N98 = 0 And N99 = 1) And (N62 = 1 And N65 != 1)
2	(N98 != 2 And N69 = 2) And (N99 = 2 And N65 = 0)
3	((N36 = 1 And N16 != 0) And (N99 = 1 Or N98 = 2)) And N99 = 2
4	(N65 = 1 Or N98 = 1) And (N99 = 2 And N98 = 2)
5	(N65 = 2 And N69 = 2) Or (N98 != 2 And N98 = 2)
6	(N98 != 0 And N62 = 2) And N65 = 1
7	(N98 = 1 And N99 = 2) And (N36 = 2 Or N99 = 1)
8	(N98 != 0 And N98 != 1) And (N16 = 2 And N1 = 0)
9	(N69 = 2 And N65 = 1) And N98 = 1
10	(N99 != 0 And N65 != 0) And (N36 = 2 And N69 = 2)

Tabela 8.2: Regras de associação geradas pela função de avaliação *Precision*.

i	Recall
1	N77 != 1 And ((N99 = 1 And N98 != 1) And N98 = 0)
2	(N4 = 1 Or N99 != 1) And (N99 = 1 And N98 = 0)
3	((N98 = 1 And N99 != 1) And N98 != 0) And N21 = 2
4	((N98 = 0 And N99 != 0) And (N99 != 0 And N99 = 1)) And N4 != 0
5	N4 != 0 And ((N99 != 1 Or N98 = 0) And N99 = 1)
6	N31 != 1 And ((N77 != 1 And (N98 = 0 And N99 = 1)) And N6 != 1)
7	(N98 = 0 And N99 = 1) And (N99 != 2 And N31 = 0)
8	N98 = 0 And ((N0 != 1 And N99 = 1) And (N77 = 0 And N98 = 0))
9	N99 = 0 And ((N77 = 2 And N98 = 1) And (N6 != 0 And N99 != 1))
10	(N77 = 2 And N99 = 1) And ((N99 != 0 Or N98 = 2) And N98 = 0)

Tabela 8.3: Regras de associação geradas pela função de avaliação *Recall*.

i	F1
1	(N98 = 0 And N99 = 1) And (N85 != 0 Or N98 = 1)
2	(N98 = 1 And N99 != 1) And N64 = 2
3	((N98 = 1 And N99 != 1) And N98 != 0) And (N64 != 0 And N64 = 2)
4	N98 != 0 And N99 = 0) And (N98 != 2 And N64 = 2)
5	(N98 = 1 And N99 != 1) And N64 = 2
6	(N64 = 2 And N38 = 0) And (N98 != 0 And N99 != 1)
7	((N98 = 2 Or N99 = 2) And N99 != 1) And (N38 != 1 And N64 = 2)
8	N99 != 1 And (N64 = 2 And ((N98 != 0 Or N83 != 1) And N98 = 1))
9	(N89 = 2 And N67 != 0) And (N99 = 1 And N98 != 1)
10	(N98 = 0 And ((N64 != 1 Or N99 = 1) And N64 = 1)) And N83 = 0

Tabela 8.4: Regras de associação geradas pela função de avaliação *F1*.

9 Apêndice B

Os modelos epistáticos sem efeito principal apresentados neste apêndice foram discutidos e disponibilizados em (CULVERHOUSE et al. 2002; VELEZ et al, 2007). No total, representam 70 modelos e são apresentados nas tabelas a seguir. A herdabilidade dos conjuntos de dados variam entre 0.01 até 0.4. A MAF varia entre 0.2 e 0.4.

$h^2 = 0.4$, MAF=0.2				$h^2 = 0.4$, MAF=0.4			
Conjunto 0	AA	Aa	aa	Conjunto 5	AA	Aa	aa
BB	0.486	0.960	0.538	BB	0.077	0.656	0.880
Bb	0.947	0.004	0.811	Bb	0.892	0.235	0.312
bb	0.640	0.606	0.909	bb	0.174	0.842	0.106
$h^2 = 0.4$, MAF=0.2				$h^2 = 0.4$, MAF=0.4			
Conjunto 1	AA	Aa	aa	Conjunto 6	AA	Aa	aa
BB	0.469	0.956	0.697	BB	0.895	0.323	0.161
Bb	0.978	0.019	0.585	Bb	0.068	0.728	0.806
bb	0.786	0.407	0.013	bb	0.925	0.233	0.362
$h^2 = 0.4$, MAF=0.2				$h^2 = 0.4$, MAF=0.4			
Conjunto 2	AA	Aa	aa	Conjunto 7	AA	Aa	aa
BB	0.498	0.954	0.786	BB	0.805	0.251	0.085
Bb	0.978	0.038	0.428	Bb	0.002	0.668	0.638
bb	0.590	0.821	0.380	bb	0.830	0.079	0.542
$h^2 = 0.4$, MAF=0.2				$h^2 = 0.4$, MAF=0.4			
Conjunto 3	AA	Aa	aa	Conjunto 8	AA	Aa	aa
BB	0.505	0.988	0.624	BB	0.307	0.682	0.958
Bb	0.945	0.085	0.807	Bb	0.997	0.390	0.281
bb	0.969	0.116	0.159	bb	0.012	0.990	0.698
$h^2 = 0.4$, MAF=0.2				$h^2 = 0.4$, MAF=0.4			
Conjunto 4	AA	Aa	aa	Conjunto 9	AA	Aa	aa
BB	0.486	0.963	0.512	BB	0.083	0.891	0.037
Bb	0.941	0.006	0.899	Bb	0.619	0.271	0.691
bb	0.691	0.541	0.614	bb	0.853	0.079	0.742

Tabela 9.1: Funções de penetrância dos 70 modelos epistáticos sem efeito principal. Conjuntos 0 até 9.

$h^2 = 0.3$, MAF=0.2				$h^2 = 0.3$, MAF=0.4			
Conjunto 10	AA	Aa	aa	Conjunto 15	AA	Aa	aa
BB	0.500	0.926	0.615	BB	0.891	0.362	0.480
Bb	0.895	0.131	0.647	Bb	0.213	0.829	0.601
bb	0.858	0.160	0.999	bb	0.925	0.267	0.685
$h^2 = 0.3$, MAF=0.2				$h^2 = 0.3$, MAF=0.4			
Conjunto 11	AA	Aa	aa	Conjunto 16	AA	Aa	aa
BB	0.413	0.851	0.535	BB	0.077	0.689	0.417
Bb	0.831	0.008	0.580	Bb	0.763	0.150	0.491
bb	0.692	0.268	0.736	bb	0.255	0.528	0.793
$h^2 = 0.3$, MAF=0.2				$h^2 = 0.3$, MAF=0.4			
Conjunto 12	AA	Aa	aa	Conjunto 17	AA	Aa	aa
BB	0.455	0.848	0.897	BB	0.132	0.793	0.274
Bb	0.890	0.088	0.016	Bb	0.799	0.213	0.514
bb	0.562	0.686	0.467	bb	0.255	0.528	0.793
$h^2 = 0.3$, MAF=0.2				$h^2 = 0.3$, MAF=0.4			
Conjunto 13	AA	Aa	aa	Conjunto 18	AA	Aa	aa
BB	0.609	0.980	0.980	BB	0.611	0.104	0.759
Bb	0.993	0.300	0.275	Bb	0.180	0.674	0.019
bb	0.876	0.483	0.683	bb	0.532	0.189	0.681
$h^2 = 0.3$, MAF=0.2				$h^2 = 0.3$, MAF=0.4			
Conjunto 14	AA	Aa	aa	Conjunto 19	AA	Aa	aa
BB	0.446	0.844	0.774	BB	0.091	0.827	0.863
Bb	0.879	0.044	0.233	Bb	0.869	0.393	0.415
bb	0.492	0.796	0.410	bb	0.738	0.508	0.363

Tabela 9.2: Funções de penetrância dos 70 modelos epistáticos sem efeito principal. Conjuntos 10 até 19.

$h^2 = 0.2$, MAF=0.2				$h^2 = 0.2$, MAF=0.4			
Conjunto 20	AA	Aa	aa	Conjunto 25	AA	Aa	aa
BB	0.428	0.757	0.812	BB	0.356	0.891	0.809
Bb	0.788	0.132	0.044	Bb	0.955	0.508	0.611
bb	0.559	0.548	0.373	bb	0.617	0.755	0.630
$h^2 = 0.2$, MAF=0.2				$h^2 = 0.2$, MAF=0.4			
Conjunto 21	AA	Aa	aa	Conjunto 26	AA	Aa	aa
BB	0.507	0.842	0.605	BB	0.086	0.536	0.641
Bb	0.845	0.162	0.629	Bb	0.677	0.275	0.096
bb	0.581	0.678	0.729	bb	0.219	0.413	0.712
$h^2 = 0.2$, MAF=0.2				$h^2 = 0.2$, MAF=0.4			
Conjunto 22	AA	Aa	aa	Conjunto 27	AA	Aa	aa
BB	0.577	0.247	0.428	BB	0.855	0.339	0.772
Bb	0.227	0.928	0.578	Bb	0.513	0.651	0.607
bb	0.586	0.262	0.158	bb	0.250	0.999	0.154
$h^2 = 0.2$, MAF=0.2				$h^2 = 0.2$, MAF=0.4			
Conjunto 23	AA	Aa	aa	Conjunto 28	AA	Aa	aa
BB	0.340	0.637	0.654	BB	0.506	0.838	0.024
Bb	0.689	0.017	0.041	Bb	0.603	0.454	0.957
bb	0.242	0.866	0.403	bb	0.729	0.427	0.753
$h^2 = 0.2$, MAF=0.2				$h^2 = 0.2$, MAF=0.4			
Conjunto 24	AA	Aa	aa	Conjunto 29	AA	Aa	aa
BB	0.387	0.726	0.734	BB	0.393	0.764	0.664
Bb	0.749	0.090	0.034	Bb	0.850	0.398	0.733
bb	0.551	0.401	0.724	bb	0.406	0.927	0.147

Tabela 9.3: Funções de penetrância dos 70 modelos epistáticos sem efeito principal. Conjuntos 20 até 29.

$h^2 = 0.1$, MAF=0.2				$h^2 = 0.1$,MAF=0.4			
Conjunto 30	AA	Aa	aa	Conjunto 35	AA	Aa	aa
BB	0.463	0.703	0.431	BB	0.137	0.484	0.187
Bb	0.653	0.277	0.806	Bb	0.482	0.166	0.365
bb	0.830	0.008	0.129	bb	0.193	0.361	0.430
$h^2 = 0.1$, MAF=0.2				$h^2 = 0.1$,MAF=0.4			
Conjunto 31	AA	Aa	aa	Conjunto 36	AA	Aa	aa
BB	0.319	0.507	0.569	BB	0.469	0.198	0.754
Bb	0.553	0.105	0.045	Bb	0.337	0.502	0.141
bb	0.203	0.777	0.280	bb	0.339	0.453	0.285
$h^2 = 0.1$, MAF=0.2				$h^2 = 0.1$,MAF=0.4			
Conjunto 32	AA	Aa	aa	Conjunto 37	AA	Aa	aa
BB	0.627	0.393	0.335	BB	0.478	0.311	0.864
Bb	0.396	0.779	0.953	Bb	0.387	0.579	0.263
bb	0.174	0.842	0.106	bb	0.634	0.436	0.138
$h^2 = 0.1$, MAF=0.2				$h^2 = 0.1$,MAF=0.4			
Conjunto 33	AA	Aa	aa	Conjunto 38	AA	Aa	aa
BB	0.297	0.540	0.441	BB	0.068	0.299	0.017
Bb	0.541	0.072	0.278	Bb	0.289	0.044	0.285
bb	0.434	0.293	0.228	bb	0.048	0.262	0.174
$h^2 = 0.1$, MAF=0.2				$h^2 = 0.1$,MAF=0.4			
Conjunto 34	AA	Aa	aa	Conjunto 39	AA	Aa	aa
BB	0.332	0.562	0.573	BB	0.539	0.120	0.258
Bb	0.583	0.112	0.147	Bb	0.165	0.378	0.325
bb	0.399	0.496	0.033	bb	0.123	0.426	0.276

Tabela 9.4: Funções de penetrância dos 70 modelos epistáticos sem efeito principal. Conjuntos 30 até 39.

$h^2 = 0.05$, MAF=0.2				$h^2 = 0.05$,MAF=0.4			
Conjunto 40	AA	Aa	aa	Conjunto 45	AA	Aa	aa
BB	0.492	0.664	0.481	BB	0.002	0.155	0.214
Bb	0.642	0.330	0.746	Bb	0.199	0.071	0.022
bb	0.656	0.396	0.000	bb	0.081	0.122	0.135
$h^2 = 0.05$, MAF=0.2				$h^2 = 0.05$,MAF=0.4			
Conjunto 41	AA	Aa	aa	Conjunto 46	AA	Aa	aa
BB	0.499	0.639	0.765	BB	0.188	0.020	0.171
Bb	0.666	0.389	0.083	Bb	0.032	0.174	0.059
bb	0.543	527	0.953	bb	0.134	0.087	0.092
$h^2 = 0.05$, MAF=0.2				$h^2 = 0.05$,MAF=0.4			
Conjunto 42	AA	Aa	aa	Conjunto 47	AA	Aa	aa
BB	0.212	0.350	0.116	BB	0.005	0.179	0.251
Bb	0.336	0.054	0.495	Bb	0.211	0.100	0.026
bb	0.227	0.273	0.495	bb	0.156	0.098	0.156
$h^2 = 0.05$, MAF=0.2				$h^2 = 0.05$,MAF=0.4			
Conjunto 43	AA	Aa	aa	Conjunto 48	AA	Aa	aa
BB	0.805	0.683	0.638	BB	0.174	0.321	0.154
Bb	0.657	0.936	0.989	Bb	0.223	0.254	0.245
bb	0.850	0.564	0.866	bb	0.448	0.025	0.424
$h^2 = 0.05$, MAF=0.2				$h^2 = 0.05$,MAF=0.4			
Conjunto 44	AA	Aa	aa	Conjunto 49	AA	Aa	aa
BB	0.638	0.488	0.383	BB	0.098	0.219	0.302
Bb	0.464	0.765	0.957	Bb	0.302	0.126	0.121
bb	0.580	0.562	0.719	bb	0.053	0.308	0.136

Tabela 9.5: Funções de penetrância dos 70 modelos epistáticos sem efeito principal. Conjuntos 40 até 49.

$h^2 = 0.025$, MAF=0.2				$h^2 = 0.025$,MAF=0.4			
Conjunto 50	AA	Aa	aa	Conjunto 55	AA	Aa	aa
BB	0.495	0.415	0.657	BB	0.166	0.165	0.128
Bb	0.429	0.616	0.121	Bb	0.114	0.199	0.143
bb	0.552	0.0331	0.419	bb	0.281	0.028	0.281
$h^2 = 0.025$, MAF=0.2				$h^2 = 0.025$,MAF=0.4			
Conjunto 51	AA	Aa	aa	Conjunto 56	AA	Aa	aa
BB	0.592	0.691	0.743	BB	0.108	0.006	0.080
Bb	0.712	0.493	0.419	Bb	0.026	0.079	0.046
bb	0.580	0.746	0.504	bb	0.021	0.090	0.025
$h^2 = 0.025$, MAF=0.2				$h^2 = 0.025$,MAF=0.4			
Conjunto 52	AA	Aa	aa	Conjunto 57	AA	Aa	aa
BB	0.108	0.194	0.186	BB	0.006	0.094	0.008
Bb	0.196	0.037	0.045	Bb	0.079	0.016	0.076
bb	0.172	0.073	0.130	bb	0.052	0.043	0.057
$h^2 = 0.025$, MAF=0.2				$h^2 = 0.025$,MAF=0.4			
Conjunto 53	AA	Aa	aa	Conjunto 58	AA	Aa	aa
BB	0.112	0.186	0.128	BB	0.199	0.072	0.168
Bb	0.193	0.024	0.138	Bb	0.086	0.187	0.076
bb	0.079	0.236	0.251	bb	0.125	0.108	0.226
$h^2 = 0.025$, MAF=0.2				$h^2 = 0.025$,MAF=0.4			
Conjunto 54	AA	Aa	aa	Conjunto 59	AA	Aa	aa
BB	0.272	0.192	0.185	BB	0.165	0.096	0.262
Bb	0.172	0.367	0.390	Bb	0.166	0.151	0.091
bb	0.345	0.069	0.005	bb	0.050	0.250	0.056

Tabela 9.6: Funções de penetrância dos 70 modelos epistáticos sem efeito principal. Conjuntos 50 até 59.

$h^2 = 0.01$, MAF=0.2				$h^2 = 0.01$, MAF=0.4			
Conjunto 60	AA	Aa	aa	Conjunto 65	AA	Aa	aa
BB	0.247	0.301	0.205	BB	0.103	0.063	0.124
Bb	0.300	0.173	0.378	Bb	0.098	0.086	0.069
bb	0.215	0.357	0.268	bb	0.021	0.147	0.059
$h^2 = 0.01$, MAF=0.2				$h^2 = 0.01$, MAF=0.4			
Conjunto 61	AA	Aa	aa	Conjunto 66	AA	Aa	aa
BB	0.222	0.276	0.141	BB	0.185	0.291	0.234
Bb	0.259	0.169	0.401	Bb	0.286	0.201	0.277
bb	0.278	0.128	0.420	bb	0.249	0.266	0.166
$h^2 = 0.01$, MAF=0.2				$h^2 = 0.01$, MAF=0.4			
Conjunto 62	AA	Aa	aa	Conjunto 67	AA	Aa	aa
BB	0.260	0.221	0.201	BB	0.073	0.042	0.015
Bb	0.204	0.315	0.348	Bb	0.024	0.064	0.059
bb	0.339	0.074	0.128	bb	0.068	0.019	0.095
$h^2 = 0.01$, MAF=0.2				$h^2 = 0.01$, MAF=0.4			
Conjunto 63	AA	Aa	aa	Conjunto 68	AA	Aa	aa
BB	0.139	0.188	0.221	BB	0.046	0.127	0.069
Bb	0.190	0.111	0.020	Bb	0.115	0.067	0.097
bb	0.206	0.051	0.253	bb	0.107	0.069	0.108
$h^2 = 0.01$, MAF=0.2				$h^2 = 0.01$, MAF=0.4			
Conjunto 64	AA	Aa	aa	Conjunto 69	AA	Aa	aa
BB	0.558	0.616	0.674	BB	0.095	0.122	0.127
Bb	0.632	0.499	0.418	Bb	0.097	0.129	0.010
bb	0.546	0.674	0.395	bb	0.201	0.044	0.122

Tabela 9.7: Funções de penetrância dos 70 modelos epistáticos sem efeito principal. Conjuntos 60 até 69.

Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 25, Num. 2 (2018) 85-92

RESEARCH ARTICLE

A Genetic Programming Model for Association Studies to Detect Epistasis in Low Heritability Data

Um Modelo de Programação Genética para Estudos de Associação para Detectção de Epistasia em Dados de Baixa Herdabilidade

Igor Magalhães Ribeiro[1],Carlos Cristiano Hasenclever Borges[2], Bruno Zonovelli da Silva[1], Wagner Arbex[3]*

Abstract: The genome-wide associations studies (GWAS) aims to identify the most influential markers in relation to the phenotype values. One of the substantial challenges is to find a non-linear mapping between genotype and phenotype, also known as epistasis, that usually becomes the process of searching and identifying functional SNPs more complex. Some diseases such as cervical cancer, leukemia and type 2 diabetes have low heritability. The heritability of the sample is directly related to the explanation defined by the genotype, so the lower the heritability the greater the influence of the environmental factors and the less the genotypic explanation. In this work, an algorithm capable of identifying epistatic associations at different levels of heritability is proposed. The developing model is a aplication of genetic programming with a specialized initialization for the initial population consisting of a random forest strategy. The initialization process aims to rank the most important SNPs increasing the probability of their insertion in the initial population of the genetic programming model. The expected behavior of the presented model for the obtainment of the causal markers intends to be robust in relation to the heritability level. The simulated experiments are case-control type with heritability level of 0.4, 0.3, 0.2 and 0.1 considering scenarios with 100 and 1000 markers. Our approach was compared with the GPAS software and a genetic programming algorithm without the initialization step. The results show that the use of an efficient population initialization method based on ranking strategy is very promising compared to other models.

Keywords: Bioinformatics — GWAS — SNP — Genetic Programming — Random Forest — Computational Modeling — Mathematical Modeling

Resumo: Os estudos de associação genômica ampla (genome-wide associations studies - GWAS) visam identificar os marcadores mais influentes em relação aos valores fenotípicos. Um dos desafios substanciais é encontrar um mapeamento não linear entre genótipo e fenótipo, também conhecido como epistasia, que geralmente se torna o processo de busca e identificação de SNPs funcionais mais complexos. Algumas doenças como o câncer do colo do útero, leucemia e diabetes tipo 2 têm baixa herdabilidade. A herdabilidade da amostra está diretamente relacionada à explicação definida pelo genótipo, portanto, quanto menor a herdabilidade, maior a influência dos fatores ambientais e menor a explicação genotípica. Neste trabalho, é proposto um algoritmo capaz de identificar associações epistáticas em diferentes níveis de herdabilidade. O modelo em desenvolvimento é uma aplicação de programação genética com uma inicialização especializada para a população inicial, consistindo de uma estratégia de florestal aleatória. O processo de inicialização visa classificar os SNPs mais importantes aumentando a probabilidade de sua inserção na população inicial do modelo de programação genética. O comportamento esperado do modelo apresentado para a obtenção dos marcadores causais pretende ser robusto em relação ao nível de herdabilidade. Os experimentos simulados são do tipo caso-controle, com nível de herdabilidade de 0,4, 0,3, 0,2 e 0,1, considerando cenários com marcadores de 100 e 1000. Nossa abordagem foi comparada com o software GPAS e um algoritmo de programação genética sem a etapa de inicialização. Os resultados mostram que o uso de um método eficiente de inicialização da população baseado na estratégia de ranking é muito promissor em comparação com outros modelos. .

Palavras-Chave: Bionformática — GWAS — SNP — Programação Genética — Floresta Aleatória — Modelagem Computacional — Modelagem Matemática

[1] *Postgraduate Program in Computational Modeling, Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil*
[2] *Department of Computer Science, Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil*
[3] *Federal University of Juiz de Fora and Brazilian Agricultural Research Corporation, Juiz de Fora, MG, Brazil*
*__Corresponding author__: wagner.arbex@{ufjf.edu.br,embrapa.br}
DOI: http://dx.doi.org/10.22456/2175-2745.79333 • **Received:** 02/01/2018 • **Accepted:** 20/04/2018

1. Introduction

Over the last decade, the studies about the human genome has generated a large amount of information, largely due to the emergence of density-based "chips" technology that has facilitated the measurement of hundreds of thousands of variations of DNA sequences throughout the human genome [1] [2]. The most common form of genomic variation or marker is known as single nucleotide polymorphisms (SNPs). These variations correspond to the alternation (substitution, deletion or insertion) of nucleotides A, T, C and G in a single position of the genome.

The genome-wide association study (GWAS) allows the finding out molecular markers that indicate the risk or predispose to complex diseases. The identification of these markers can help directly or indirectly understand the mechanisms of a particular disease. Directly finding the marker and indirectly indicating the gene, metabolic pathway among other biological characteristics, and the biggest challenge is to interpret and understand the large number of information obtained in the genotyping process, when molecular markers are identified [3][2].

In [4], the authors indicates that one of the types of gene interaction that gives rise to complex diseases is called epistasis. This type of interaction makes the mapping between genotype and non-linear phenotype, that is, one marker can mask or completely alter the behavior of the other generating a completely new characteristic. In this way, the interaction becomes more difficult to detect. Heritability can be estimated by the ratio between the variances of the genotype and phenotype. This ratio measures the proportionality of how much the genetic factor influences the phenotype [5]. The heritability directly interferes with the ability to correctly select markers of interest for the study. The lower the heritability, the less the explanation obtained through the genotype, and the greater the influence of environmental factors.

Several medical conditions or diseases have low heritability, for instance: asthma (0.3) [6], bladder cancer (0.07-0.31) [7], cervical cancer (0.22) [8], leukemia (0.01) [9]; type-2 diabetes (0.26) [10], and so on. Therefore, it is necessary to develop algorithms capable of identifying risk factors at different levels of heritability.

In addition to heritability, there is a complexity in the genotype-phenotype relationship due to distinct gene actions. The works [1] and [11] explain that the linear modeling (linear regression) used in GWAS problems considers only one SNP at a time, in this context, the gene-gene and environment-gene interactions of each marker are ignored. For the identification of more complex genetic actions such as epistasis and dominance, machine learning models that consider multiple markers in classification and regression problems have been presented to identify non-linear interactions between SNPs.

Initialization approaches were used in related works [1, 12, 13], and according [14] the use of expert knowledge can significantly improve the performance in detection SNP-SNP interactions in genetic programming algorithms. These proposed models used a feature selection algorithm called Relief [15] and their variants. The idea of Relief is estimate the feature weight according to their ability to discriminate between individuals and their neighbors . However these algorithms can identify possible SNP interactions, they are susceptible to noise. They may capture marginal effects (single SNP interaction with phenotype) rather than epistatic interactions.

The objective of this work is to develop a model to identify non-linear interaction of functional SNPs, i.e., epistasis, across different levels of heritability. The model proposed is an algorithm of evolution of solutions based on genetic programming (GP) with initialization through random forest. The idea behind the random forest choice is to use a most informative and robust measure in this context. The increase in mean square error (MSE) of predictions can rank SNPs with low noise and more informative to be part of epistatic interactions than Gini index for example, taking between 5%-25% of extra computing time.

2. Proposed Model

We propose a model combining evolutionary computation and machine learning techniques, more precisely, genetic programming to analysis to genotype/phenotype and random forest as expert knowledge guiding the search for SNPs interactions that could lead to a complex disease risk. The expert knowledge algorithm is used to measure of attribute quality and allows SNPs of interest to be inserted into the initial population of GP algorithm.

2.1 Individuals

The structure of individuals is based on a tree – or a tree representation of solutions – was suggested by [16] , where the authors proposed to use multi-valued logic expressions in disjunctive normal form (DNF). A DNF logic expression is disjunctive of one or more monomials, where one monomial consists of a single or a set of literals. In Figure 1, an example of generic tree with DNF logic expression representing a GP individual is shown. The GP grammar adopted is simple and the function set is given by "AND" and "OR" expressions. The terminal set consists of SNPs and their respective alleles, for instance "SNP1 = 0".

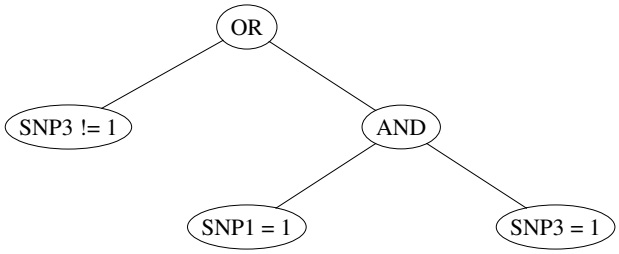


Figure 1. Example of an individual used in GPi. The individuals are expression trees that represents SNP-SNP high order interactions.

2.2 Fitness function

The evaluation of individuals is given by the fitness function f_i, shown in (1):

$$f_i = \frac{T}{VP+VN} + \frac{N_i}{\alpha} \quad (1)$$

where i is an index for an individual, T represents the total of case-control individuals. VP is the true-positives, VN is the true-negatives correctly classified. N_i is the numbers of nodes and α is a parsimony constant (introduced in [17]).

2.3 Operations to generate new individuals

To create a new generation, the genetic operators including crossover and mutation are applied. An overview on the crossover operator is given in Figure 2, where two individuals are selected to crossing. A random node are selected in each individual, then two offspring are created from their combination. The first offspring is a combination of the individual for the cut-off point with the second individual after the cut-off, respectively. The second individual is generated from the inverse composition of the first offspring. In the end, these two individuals are inserted into the new generation.

2.4 Expert knowledge to generate initial population

The initialization mechanism that generate the initial population is based on the importance of the variables according to the random forest algorithm. It is used to capture an isolated effect or a possible genotype-phenotype attribute interaction, generating a ranking of SNPs that predict the phenotype. The measure adopted is most informative than Gini index in this context. The measure represents the increase in MSE of predictions from a sample estimated with an out-of-bag cross validation method.

Usually in GWAS, the parameters are optimized, so, the number of variables to choose from the decision tree nodes and the number of trees that make up the forest need to be defined. The values were defined from empirical tests that presented significant or satisfactory results for the problem in question.

Thereby, the number of variables used in each training subset was the same number of markers used in the simulation and the amount of trees defined by the forest was 1500 for the experiments with 100 markers and 3000 for the experiments with 1000 markers.

To generate the initial population, each terminal node of each individual is submitted to a tournament process in which a marker is selected from among the markers present in the population at random. A comparison of the value assigned to each marker by the random forest algorithm is performed, the one with the highest value is selected to generate the terminal node. Each individual can only have one copy of marker, so if a terminal is populated by a given SNP, it can no longer appear in the solution tree and another tournament is performed until a previously uninserted SNP is found.

2.5 Parameter setting

Table 1 shows the parameter setting used by the algorithms that are part of the model proposed in this work. The parameters such as Population size, Generations, Crossover and Mutation frequency was based on [12] and the functions and terminal sets on [16]. The GP algorithm proposed here has been implement in ECJ [18], and R [19] [20].

Table 1. Parameter setting

Item	Parameter
Population size	4096
Generations	50
Crossover	Single-point
Mutation frequency	0.05
Selection	Tournament

3. Experiments and Results

The following experiments and analyses are conducted on Intel®Core™i7-4770K CPU with 3.50GHz × 8 and 32 GB of RAM. A simulation study was performed to evaluate our model in a GWAS problem. The objective of this simulation is to generate artificial databases capable of capturing the epistatic effects that give rise to phenotypes in cases of low heritability commonly found in genetics. Using GAMETES [21], we could selected heritabilities ranges and created penetrance functions that defines a relationship between the genotype and phenotype. Table 2 exemplifies a penetrance function used to generate a template with epistasis.

Table 2. Example of a penetrance function for a model presenting epistasis.

	AA (0.25)	Aa (0.50)	aa (0.25)
BB (0.25)	0.451	0.214	0.190
Bb (0.50)	0.192	0.164	0.065
bb (0.25)	0.139	0.350	0.463

To the experiments, we developed four penetrance functions where each model has two functional SNPs represents an epistatic interaction and a heritability range between 0.1, 0.2, 0.3, 0.4 respectively. In all scenarios, the minor allele frequency (MAF) was 0.2 and the SNPs represents three alleles (0, 1 or 2).

For each database, functional SNPs were added to other randomly generated markers, defining bases of 100 and 1000 attributes. Each algorithm was run 30 times and counted the number of times that the functional SNPs were correctly selected as the best model of the genetic programming algorithm – we call ”power” the percentage of each algorithm identifies the functional SNPs.

This value represents the predictive power estimation of the proposed method for the phenotype, that is, which frequent

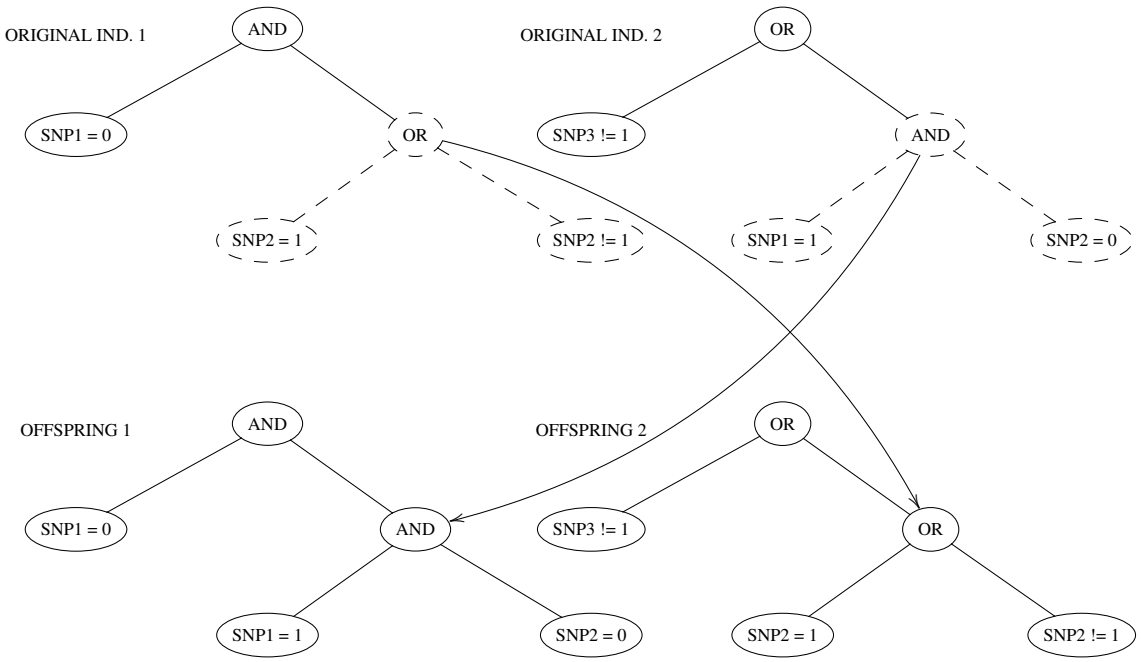


Figure 2. Examples for the crossover used in GPi

the method is able to find the expected solution. The parameter settings used in the simulations were based on [12].

We compared the power GPi algorithm, developed within the scope of this work, against GPi algorithm without initialization step – referred to as GP – and GPAS [16] on estimation of power. We consider the output of each run of GPAS as correct if the best 5 individuals contain the two functional SNPs. This evaluation criterion was used in [12].

As written previously, the simulation data was generated by GAMETES, with all the parameter settings are shown in Table 3:

Table 3. Parameters of the GAMETES simulator

Item	Parameter
MAF	0.2
Population size	2000
SNPs	100 and 1000
Heritability	0.1,0.2,0.3,0.4

For each experiment, a different penetrance function was automatically generated. For example, a possible solution is given by the syntactic tree in Figure 3 for 100 markers and heritability equal to 0.4. The solution tree is generated from the penetrance function given by Table 4.

The results obtained for 100 and 1000 SNPs can be seen respectively in Figures 4 and 5. For the datasets with 100 SNPs, we can observe that GPi – actually, the proposed model – found the correct rules for all heritabilities – even when the heritability dropped to 0.1. The GP algorithm, i.e., without the initialization step, presented satisfactory results. The results achieved by the power of GPAS showed that the algorithm is still satisfactory, presenting a variation in the results only for the case of heritability is equal to 0.1.

Table 4. Penetrance function simulating epistasis effect (database with 100 markers, heritability = 0.4).

	AA (0.25)	Aa (0.50)	aa (0.25)
BB (0.25)	0.535	0.991	0.930
Bb (0.50)	0.998	0.140	0.315
bb (0.25)	0.871	0.433	0.003

Experiments with 100 SNPs indicate that regardless of the methods, functional SNPs can be found. The proposed model obtained a small advantage than the other methods. However, in the databases with 1000 SNPs , the results obtained by each algorithm differ greatly between them. In this scenario, the complexity in finding the functional SNPs has increased. The proposed method obtained significant results even when heritability drops to 0.1. Figure 6 shows the ranking of the SNPs performed by the random forest algorithm. We can note the functional SNPs appear at the top of the all ranking list.

4. Discussion

The identification of SNPs involved directly or indirectly in the gene interactions in scenarios that present low heritability is a fundamental step for the understanding of several complex diseases. The discovery of the biological mechanisms involved in the process can help research directed to the de-

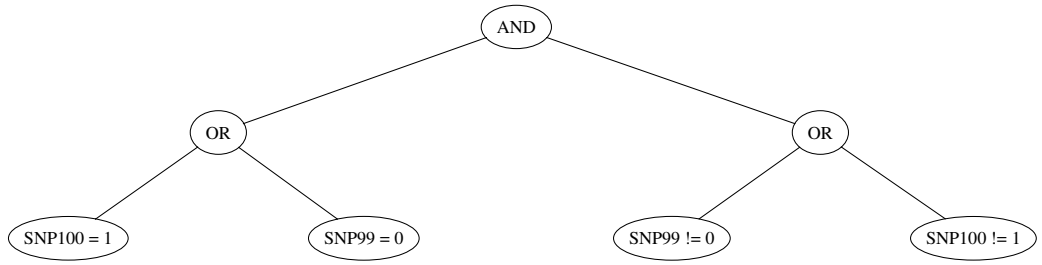


Figure 3. Individual representing a solutions for Table 4.

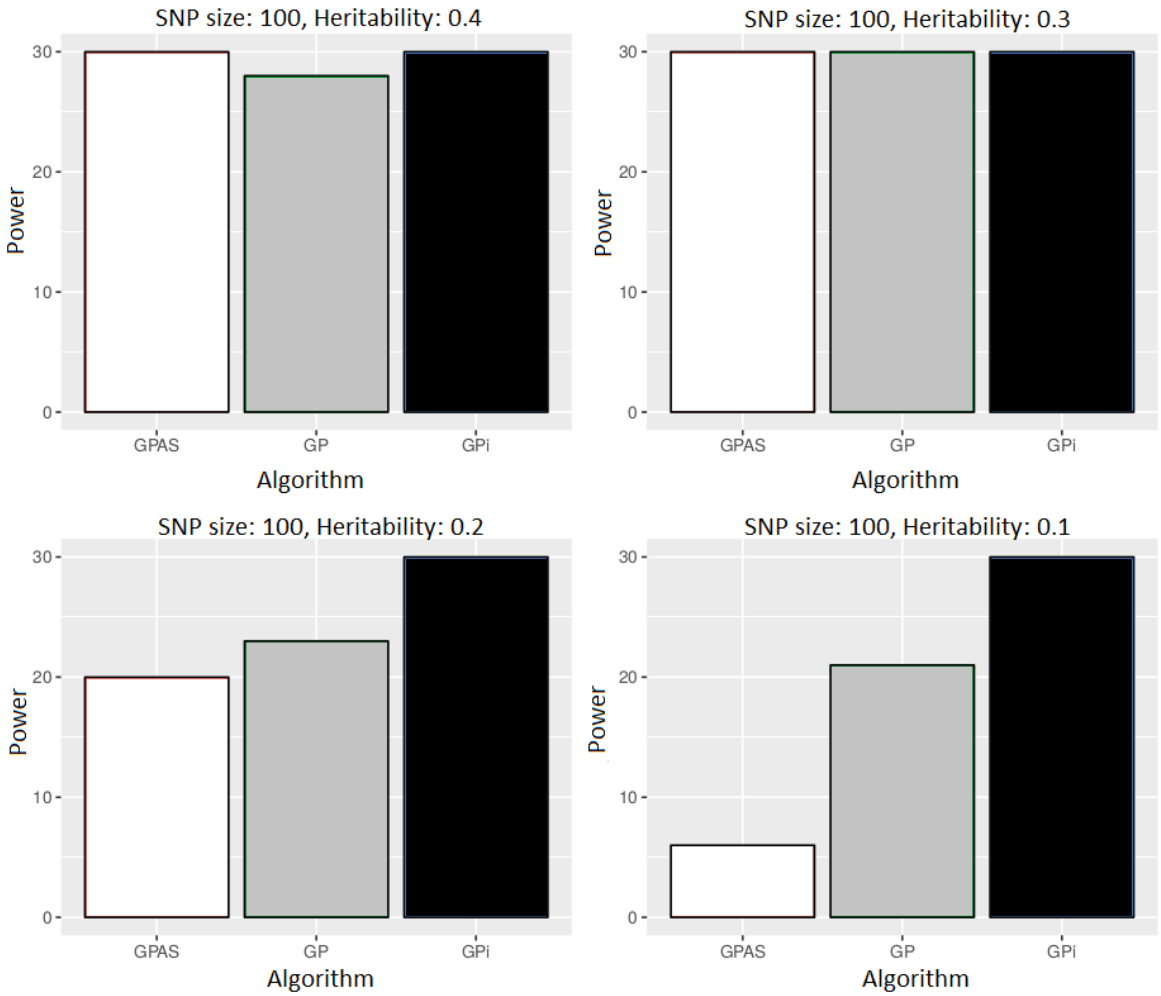


Figure 4. Graphic representing the power of each algorithm (GPAS, GP, GPi) across heritability of 0.4,0.3,0.2,0.1 with a dataset containing 100 SNPs. The power is the number of times that the algorithm identifies the correct two functional SNPs.

velopment of prevention and cure methods.

Non-random initialization methods of the initial population in evolutionary algorithms have been shown to be a strategy to aid in the search for causal SNPs in these scenarios, producing more significant results than algorithms that do not use this strategy. However, we can observe that in cases where heritability is considerable ($\geq 40\%$), initialization strategies may not be the best choice, since the other methods present significant results in this context and do not depend on this step which can be computationally expensive. To provide more conclusive basis for these analyzes, in the future, real datasets could be used, such as GWAS data from different types of complex diseases.

In addition, in order to ratify the results obtained by ana-

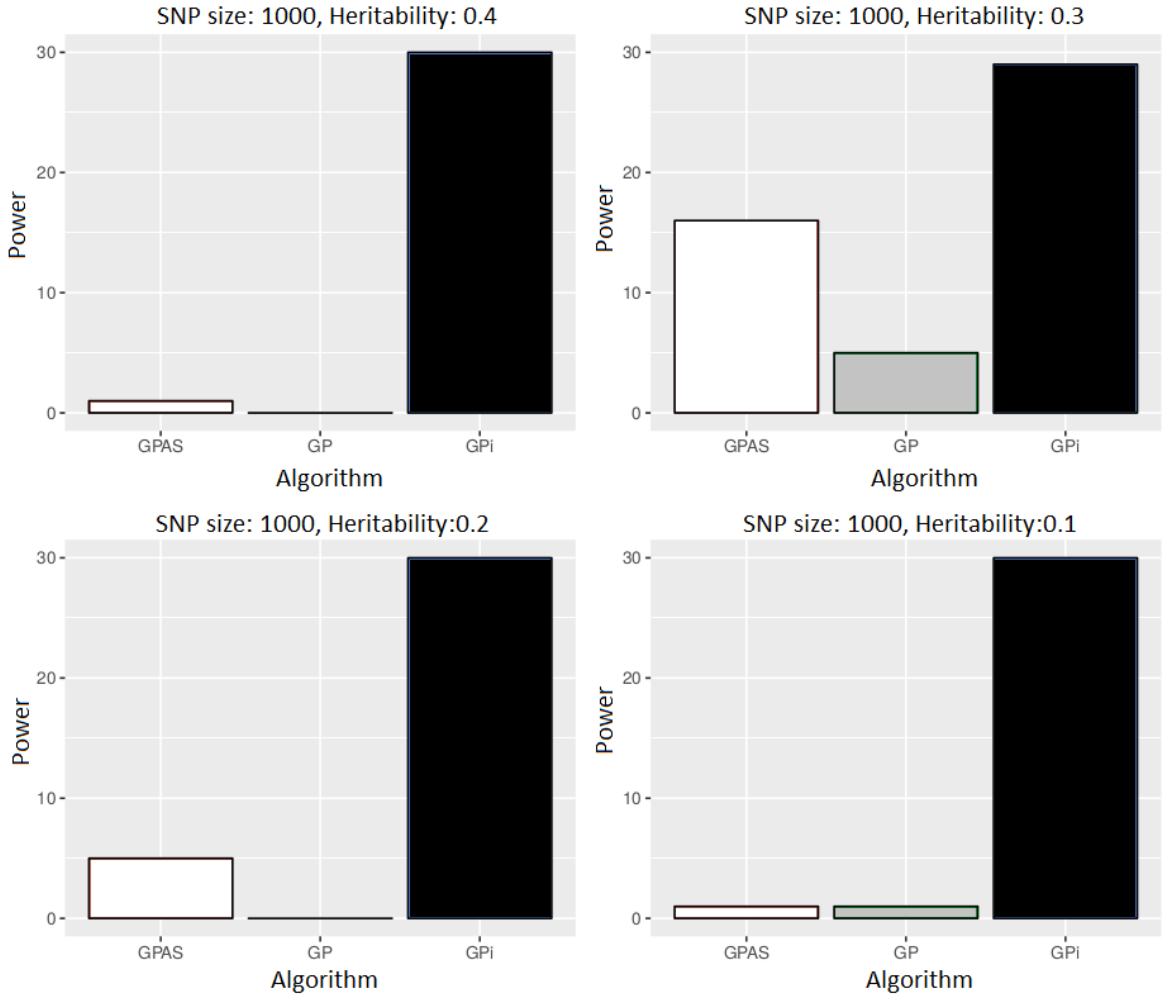


Figure 5. Graphic representing the power of each algorithm (GPAS, GP, GPi) across heritability of 0.4,0.3,0.2,0.1 with a dataset containing 1000 SNPs. The power is the number of times that the algorithm identifies the correct two functional SNPs.

lyzing the algorithm errors in the identification of the markers, other evolutionary algorithms can be used, as well as more efficient GP proficiency functions. The question of the objective function should be better investigated actually since the GP without initialization has generally presented the worst results.

Another two points to take into consideration is the initialization method and size of the databases. In the experiments present, we used databases of 100 and 1000 markers, following the experiments of similar algorithms in the literature. However, a real GWAS database has thousands or even hundreds of thousands of SNPs. This condition implies the need for efficient dimensionality reduction algorithms and / or filters. Furthermore, other classification methods can be combined to improve initialization mechanisms in cases of extremely low heritability (≤ 0.1).

5. Conclusion

A GP algorithm aims to explore all search space. However, due to the large number of possible combinations, this search may be computationally feasible. Expert knowledge approaches are recommended in these cases. in cases.

The results of the methods compared in this work showed that the use of an expert knowledge makes it possible to reduce the search space of the GP algorithm, proving to be effective, even in low heritability dataset. SNPs with higher quality of information are selected and inserted into the initial population of GP using the measure of increase in MSE of the random forest algorithm.

We show that random forest is an option among the algorithms used in other studies as expert knowledge methods and it has shown to be able to capture possible candidate markers for epistatic interactions and to be less sensitive to noises and marginal effects.

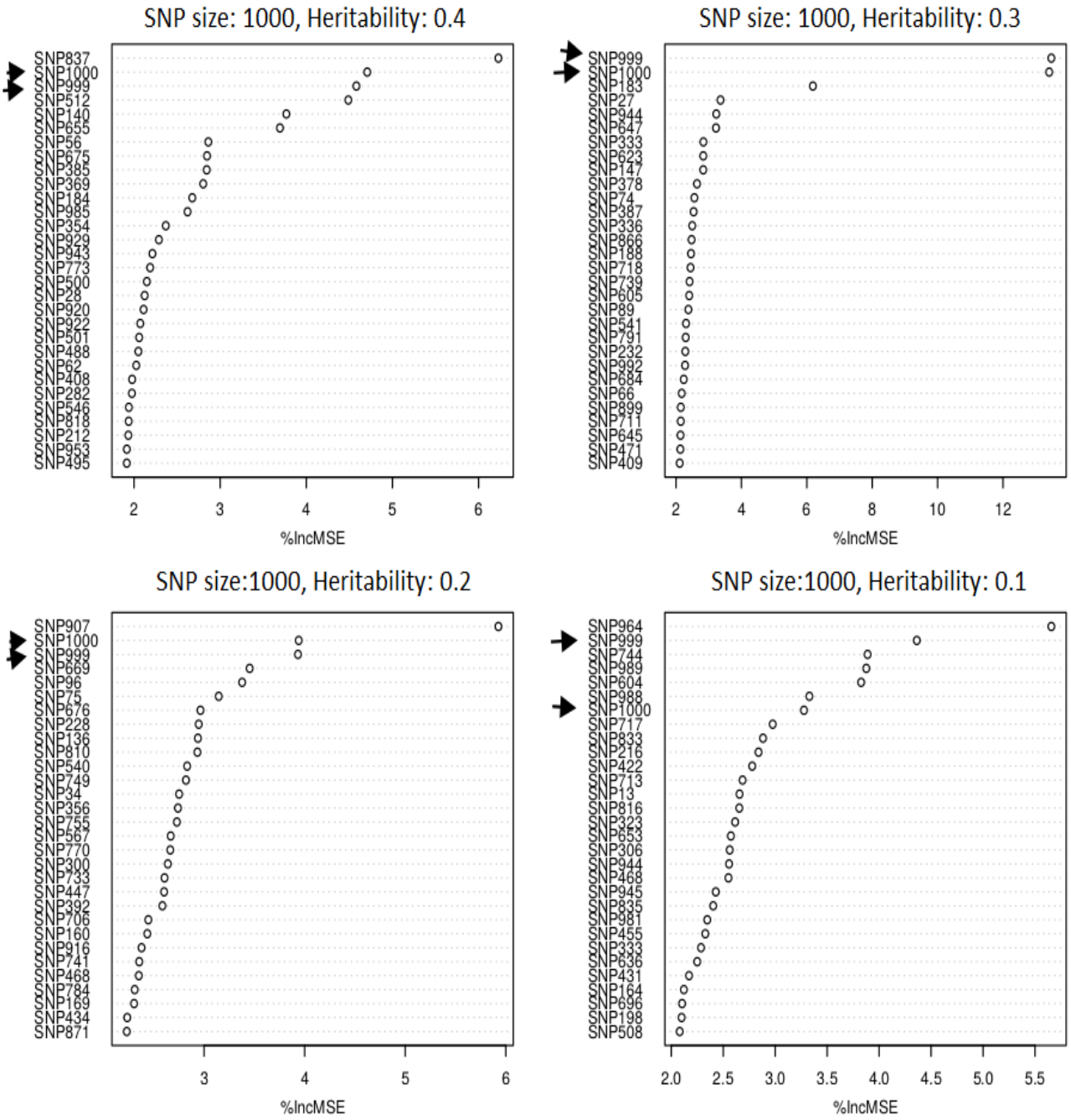


Figure 6. Ranking of the SNPs performed by the random forest algorithm. The results show the initialization step of the initial population in each scenario. The arrows indicate the positions of SNP999 and SNP1000 (the two functional SNPs).

6. Acknowledgment

The authors thanks to reviewers who gave useful comments, and would like to express thanks to the Coordination for the Improvement of Higher Level Personnel (CAPES) and National Council for Scientific and Technological Development (CNPq) and the State of Minas Gerais Research Support Agency (FAPEMIG) for the financial support for the accomplishment of this paper; and to the Postgraduate Program in Computational Modeling of Federal University of Juiz de Fora (UFJF) for the academic support.

7. Author contributions

IMG developed the proposed model, carried out the experiments, analyzed the results and contributed to the methodol-

ogy of this study. BZS structured and provided the infrastructure and computing resources needed to perform the experiments and and contributed to the proposed model. CCHB and WA are the project leaders, proposed the methodology and general approach of this study.

References

[1] MOORE, J. H.; WHITE, B. C. Tuning relieff for genome-wide genetic analysis. In: MARCHIOR, E.; MOORE, J. H.; RAJAPAKSE, J. C. (Ed.). *Proceedings of the 5th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Berlin, Heidelberg: Springer-Verlag, 2007. v. 4447, p. 166–175.

[2] BUSH, W. S.; MOORE, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput Biol.*, v. 8, n. 12, p. 1–11, 2012.

[3] MANOLIO, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med*, v. 363, n. 2, p. 166–176, 2010.

[4] GRIFFITHS, A. et al. *An Introduction to Genetic Analysis*. 7. ed. New York, USA: W. H. Freeman, 2000. v. 1.

[5] GRIFFITHS, A. J. *Introdução à genética*. 9. ed. Rio de Janeiro, Brazil: Guanabara Koogan, 2008. v. 1.

[6] TAN, H. et al. The estimation of heritability for twin data based on concordances of sex and disease. *Chronic Dis Can.*, v. 26, n. 1, p. 9–12, 2005.

[7] GU, J.; WU, X. Genetic susceptibility to bladder cancer risk and outcome. *Per Med.*, v. 8, n. 3, p. 365–374, 2011.

[8] CZENE, K.; LICHTENSTEIN, P.; HEMMINKI, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int. J. Cancer*, v. 99, n. 2, p. 260–266, 2002.

[9] CZENE, K.; LICHTENSTEIN, P.; HEMMINKI, K. Environmental and heritable causes of cancer among 9.6 million individuals in the swedish family-cancer database. *Int J Cancer.*, v. 99, n. 2, p. 260–266, 2002.

[10] POULSEN, P.; KYVIK, K. O.; VAAG A.AND BECK-NIELSEN, H. Heritability of type ii (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance – a population-based twin study. *Diabetologia*, v. 42, n. 2, p. 139–145, 1999.

[11] MOORE, J.; WHITE, B. Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge. In: RIOLO, R.; SOULE, T.; WORZEL, B. *Genetic Programming Theory and Practice IV*. 1. ed. Boston, USA: Springer, 2007. (Genetic and Evolutionary Computation, v. 1), cap. 2, p. 11–28.

[12] SZE-TO, H.-Y. et al. Gp-pi: Using genetic programming with penalization and initialization on genome-wide association study. In: RUTKOWSKI, L. et al. (Ed.). *Artificial Intelligence and Soft Computing*. 1. ed. Berlin, Germany: Springer, 2013, (Lecture Notes in Computer Science, v. 7895). cap. 30, p. 330–341.

[13] GREENE, C. S.; WHITE, B. C.; MOORE, J. H. Using expert knowledge in initialization for genome-wide analysis of epistasis using genetic programming. In: RYAN, C.; KEIJZER, M. (Ed.). *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*. New York, NY, USA: ACM, 2008. v. 1, p. 351–352.

[14] MOORE, J.; WHITE, B. *Genome-Wide Genetic Analysis Using Genetic Programming: The Critical Need for Expert Knowledge*. 2007.

[15] KIRA, K.; RENDELL, L. A. A practical approach to feature selection. In: SLEEMAN, D.; EDWARDS, P. (Ed.). *Proceedings of the Ninth International Workshop on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992. v. 1.

[16] NUNKESSER, R. et al. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics*, v. 23, n. 24, p. 3280–3288, 2007.

[17] BLEULER, S. et al. Multiobjective Genetic Programming: Reducing Bloat by Using SPEA2. In: CEC 2001. *Congress on Evolutionary Computation*. Seoul, South Korea: IEEE, 2001. v. 9.

[18] LUKE, S. et al. *ECJ 16: A Java-based Evolutionary Computation Research System*. 2007.

[19] R.C.R TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2008.

[20] LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2, n. 3, p. 18–22, 2002.

[21] URBANOWICZ, R. J. et al. METHODOLOGY GAMETES : a fast , direct algorithm for generating pure , strict , epistatic models with random architectures. *BioData Min.*, v. 5, n. 16, p. 1–14, 2012.