



Universidade Federal de Pernambuco UFPE

Centro de Informática CIn

Graduação em Ciência da Computação

Trabalho de Graduação

**Estudo de Métodos para Identificar**

**Interações entre Polimorfismos**

Eduardo Gade Gusmão

Recife, Dezembro de 2010

Universidade Federal de Pernambuco - UFPE  
Centro de Informática - CIn

Graduação em Ciência da Computação

Trabalho de Graduação

**Estudo de Métodos para Identificar**

**Interações entre Polimorfismos**

Eduardo Gade Gusmão  
egg@cin.ufpe.br

*Trabalho de Graduação apresentado no Centro de Informática (CIn) da Universidade Federal de Pernambuco (UFPE) como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.*

Orientador: Katia Silva Guimarães  
*Ph.D. in Computer Science*

Recife, Dezembro de 2010

Dedico este trabalho à minha família, que me forneceu todo o apoio necessário para o meu crescimento em todos os aspectos e à orientadora Katia, por despertar em mim o interesse na pesquisa científica e na carreira acadêmica.

Agradecimentos

Nem se eu escrevesse um trabalho de graduação inteiro apenas agradecendo o apoio de Christiani Gade Gusmão seria suficiente. Caso tal trabalho fosse escrito ele conteria a mais incrível história de amizade e companheirismo entre duas pessoas que concordam até nos aspectos discordantes. Em 1987 ela assumiu o posto de mãe, cargo que assumiu com louvor apesar dos excessos de cuidados em alguns momentos. Em 2003 assumiu o cargo de mãe de adolescente, bastante difícil e que serve apenas de preparação para o cargo de mãe de adolescente vestibulando em 2005. Sempre que precisei, sem exceção, ela estava presente para me inspirar e fornecer o tipo de conselho lúcido e acolhedor que apenas essas virtuosas da empatia (mães) podem dar.

Não menos importantes foram Herta Gade Gusmão e Celmilo José Evangelista Gusmão, respectivamente minha avó e meu avô, para o meu crescimento. A primeira me serviu de exemplo por ter concluído um mestrado nesta mesma instituição de ensino no qual pretendo receber o grau. Apesar de ter concluído os estudos do segundo grau de forma tardia através de supletivo, já sendo mãe da minha mãe e tias, tarefa tão difícil quanto sequenciar um genoma completo munido de um papel e uma caneta (sem tinta). O meu avô não poderá presenciar fisicamente a conclusão desta etapa de minha vida, porém ele sabe que isso só foi possível graças ao exemplo de caráter que recebi do homem mais importante em minha vida. Agradeço na mesma proporção às minhas tias Ingrid Gade Gusmão (segunda mãe) e Sthephanie Sophie Gade Gusmão (tia, amiga e irmã), cujas diferentes personalidades permitiram conselhos complementares pelos quais serei eternamente grato.

A professora Katia Silva Guimarães teve um papel fundamental numa fase da minha vida marcada por dúvidas sobre qual trilha percorrer. Apesar de rica e cheia de desafios interessantes, nenhuma área da ciência da computação parecia despertar completamente meu interesse. Esta docente dedicada e amiga me introduziu então à Bioinformática, ciência pelo qual me encantei e pretendo me dedicar durante o resto da minha jornada. Ela não foi responsável apenas por conselhos ou dicas e sim por todos os ensinamentos fundamentais que um indivíduo precisa para iniciar uma vida acadêmica, e por isso serei grato eternamente.

Aos meus grandes companheiros de curso Paulo Ricardo, Felipe Kühner, Nelson Gutemberg e João Rufino, pessoas com personalidades tão distintas e que admiro bastante. Me ajudaram ao longo de todo este curso de graduação e me aceitaram em seus grupos de projetos de disciplina ou extra-curriculares (ping-pong ou poker). Com eles compartilho o perfeccionismo quase paranóico, a seriedade levemente ranzinza, o senso de humor ingênuo e os sonhos de uma carreira promissora sempre com os pés no chão.

Aos docentes Sílvio Melo e Ruy Queiroz que me permitiram aprofundar meus conhecimentos em disciplinas excelentemente ministradas e projetos de monitoria e Ivan Gesteira cujos questionamentos bioinformáticos me intrigavam em todas as aulas. Também à atual aluna de mestrado e futura colega de profissão Flávia Roberta Barbosa de Araújo por me suportar e me ensinar o foco necessário para se conduzir uma pesquisa científica.

Por fim, agradeço ao meu amigo e companheiro Eduardo Henrique Farias de Carvalho, cujo maior erro foi ter esperado tanto tempo para entrar em minha vida e me fornecer um suporte sem o qual eu não consigo me imaginar sem. Durante os meus últimos anos da graduação ele foi capaz de me apoiar em momentos de dúvida ao mesmo tempo que, sem nem ele mesmo perceber, me ensinar lições de valor inestimável.

*“*[*We can only see a short distance ahead, but we can see plenty there that needs to be done.*](http://thinkexist.com/quotation/we_can_only_see_a_short_distance_ahead-but_we_can/334124.html)*”*

̶ Alan Turing

Resumo

Após o término do primeiro sequenciamento humano completo iniciou-se a chamada era pós genômica. Esta fase da biologia molecular tem como principal objetivo encontrar respostas no código genético para questionamentos a respeito da quantificação e qualificação do impacto genético nas características fenotípicas do ser que possui um dos sistemas metabólicos mais complexos e fascinantes. Na incessante busca pelo autoconhecimento, o *homo sapiens* está rapidamente aperfeiçoando as metodologias de sequenciamento, gerando uma enorme quantidade de dados biológicos que são interpretados por cientistas de diversas áreas como biologia, química, física e estatística. Naturalmente, como forma de organização e esquematização, alguns paradigmas de pesquisa estão sendo criados, como o *Genome-Wide Association Studies (GWAS)*. Como sugere o título, neste esquema são analisados diversos trechos do genoma humano a procura de variações genéticas entre indivíduos que consigam explicar doenças comuns e complexas como alguns tipos de câncer, esquizofrenia, Alzheimer, transtorno bipolar e muitas outras. Estudos que buscam interações entre essas variações, em particular, têm se mostrado bastante promissores, porém o espaço de possibilidades de combinações entre esses sítios variantes excede os limites de métodos *in vitro* ou estatísticos. Isso motivou o surgimento de vários métodos de aprendizagem de máquina e mineração de dados para restringir este espaço de possibilidades, direcionando as pesquisas apenas às combinações mais promissoras. A principal proposta deste trabalho é comparar o desempenho computacional de alguns desses métodos a partir de um conjunto de dados estatisticamente sintetizado de forma a simular um experimento biológico real, analisando os resultados à luz de ambas as ciências da computação e biológica, e criar uma ferramenta web disponível publicamente, onde os pesquisadores podem ter acesso a esses métodos e analisar seus dados de forma fácil e segura.

**Palavras-chave:** Entropy-Based SNP-SNP Interaction Method, Epístase, Genome-Wide Association Studies, Multi-Approach SNP-SNP Interaction Analysis, Multifactor Dimensionality Reduction, Polimorfismos de único nucleotídeo, Polymorphism Interaction Analysis

Abstract

After the conclusion of the first complete human sequencing, we are facing the so-called post-genomic era. The main goal now is to find answers in the genetic code for questions about the quantification and qualification of the genetic impact in certain phenotypic characteristics of the being that has one of the most complex and fascinating metabolic systems. In the endless search for self-knowledge, the *homo* sapiens is rapidly improving the sequencing methodologies, generating a massive amount of biological data that are analyzed by scientists from several areas such as biology, chemistry, physics and statistics. Naturally, to organize and schematize the analysis procedure, some research paradigms are being created, as the *Genome-Wide Association Studies (GWAS)*. As the title suggests, in this schema several parts of the human genome are analyzed in the search of genetic variations between individuals that can explain common complex diseases like some types of cancer, schizophrenia, Alzheimer, bipolar disorder and many others. Studies that search for interactions among this variations, particularly, have shown to be very promising, nevertheless the space of possibilities of combinations between these variant loci exceeds *in vitro* or statistical methods capacities. This motivated the appearance of many machine learning and data mining methods to narrow this massive space of possibilities, directing the research to the most interesting combinations only. The main proposal of this work is to compare the computational performance of some of these methods using a statistically synthesized dataset simulating a real biological experiment, analyzing the results in the sense of both computer and biological sciences, and create a web-based tool publicly available, where researchers can access the implemented methods and analyze their data easily and safely.

**Palavras-chave:** Entropy-Based SNP-SNP Interaction Method, Epistasis, Genome-Wide Association Studies, Multi-Approach SNP-SNP Interaction Analysis, Multifactor Dimensionality Reduction, Polymorphism Interaction Analysis, Single Nucleotide Polymorphism

Sumário

[1. Introdução 1](#_Toc279935759)

[1.1. Importância do Estudo 1](#_Toc279935760)

[1.2. Contribuições 3](#_Toc279935761)

[1.3. Estrutura do Documento 4](#_Toc279935762)

[2. Contexto 5](#_Toc279935763)

[2.1. Dogma Central da Biologia Molecular 5](#_Toc279935764)

[2.1.1. Estruturas envolvidas 6](#_Toc279935765)

[2.1.2. Fases do dogma central 7](#_Toc279935766)

[2.2. Polimorfismos de Único Nucleotídeo 9](#_Toc279935767)

[2.3. Epístase 10](#_Toc279935768)

[2.4. Estudos de Associação 14](#_Toc279935769)

[2.5. Análise de Interações entre Atributos 16](#_Toc279935770)

[2.6. Métodos Computacionais para Análise de Interações SNP-SNP 19](#_Toc279935771)

[2.6.1. Estatística tradicional 20](#_Toc279935772)

[2.6.2. Redução de dimensionalidade 21](#_Toc279935773)

[2.6.3. Árvores de decisão e florestas randômicas 21](#_Toc279935774)

[2.6.4. Funções de pontuação 21](#_Toc279935775)

[2.6.5. Buscas heurísticas 22](#_Toc279935776)

[2.6.6. Algoritmos de classificação 22](#_Toc279935777)

[2.6.7. Seleção de atributos tradicional 22](#_Toc279935778)

[2.6.8. Fontes de dados atuais 23](#_Toc279935779)

[2.7. Considerações Finais 23](#_Toc279935780)

[3. Métodos 25](#_Toc279935781)

[3.1. Formato da Entrada 26](#_Toc279935782)

[3.2. Polymorphism Interaction Analysis ̶ PIA 28](#_Toc279935783)

[3.2.1. Seleção da próxima combinação de atributos a ser analisada 28](#_Toc279935784)

[3.2.2. Criação da tabela de genótipo-fenótipo 29](#_Toc279935785)

[3.2.3. Validação cruzada 29](#_Toc279935786)

[3.2.4. População da tabela de contingência 29](#_Toc279935787)

[3.2.5. Cálculo da pontuação da combinação 31](#_Toc279935788)

[3.3. Multifactor Dimensionality Reduction ̶ MDR 33](#_Toc279935789)

[3.3.1. Validação Cruzada 34](#_Toc279935790)

[3.3.2. Seleção da próxima combinação de atributos a ser analisada 34](#_Toc279935791)

[3.3.3. Classificação 34](#_Toc279935792)

[3.3.4. Predição 35](#_Toc279935793)

[3.4. Entropy-Based SNP-SNP Interaction Method ̶ ESNP2 36](#_Toc279935794)

[3.4.1. Cálculo da entropia individual 37](#_Toc279935795)

[3.4.2. Seleção da próxima combinação de atributos a ser analisada 37](#_Toc279935796)

[3.4.3. Criação da tabela de genótipo-fenótipo 38](#_Toc279935797)

[3.4.4. Cálculo dos efeitos de interação da combinação 38](#_Toc279935798)

[3.5. Multi-Approach SNP-SNP Interaction Analysis ̶ MASS 39](#_Toc279935799)

[3.5.1. Motivação 39](#_Toc279935800)

[3.5.2. Cálculo das métricas individualmente 41](#_Toc279935801)

[3.5.3. Seleção da próxima combinação de atributos a ser analisada 42](#_Toc279935802)

[3.5.4. Criação da tabela de genótipo-fenótipo 42](#_Toc279935803)

[3.5.5. Cálculo dos efeitos de interação intermediários (opcional) 42](#_Toc279935804)

[3.5.6. Cálculo dos efeitos de interação da combinação 42](#_Toc279935805)

[3.6. Considerações Finais 43](#_Toc279935806)

[4. Experimentos 44](#_Toc279935807)

[4.1. Conjunto de Dados 44](#_Toc279935808)

[4.2. Experimento Comparativo 47](#_Toc279935809)

[4.2.1. Taxa de acerto 47](#_Toc279935810)

[4.2.2. Tempo de execução 47](#_Toc279935811)

[4.2.3. Parâmetros utilizados nos métodos 48](#_Toc279935812)

[5. Resultados e Discussão 50](#_Toc279935813)

[5.1. Taxa de Acerto 50](#_Toc279935814)

[5.1.1. Tamanho da entrada 50](#_Toc279935815)

[5.1.2. Variação de parâmetros biológicos 54](#_Toc279935816)

[5.1.3. Método MASS 56](#_Toc279935817)

[5.1.4. Discussão 56](#_Toc279935818)

[5.2. Tempo de Execução 58](#_Toc279935819)

[5.2.1. Tempo de execução geral 58](#_Toc279935820)

[5.2.2. Tempo de execução relativo 59](#_Toc279935821)

[5.2.3. Tempo de execução específico para cada método 60](#_Toc279935822)

[5.2.4. Discussão 61](#_Toc279935823)

[5.3. Ferramenta Web 62](#_Toc279935824)

[6. Conclusão 64](#_Toc279935825)

[6.1. Objetivos Atingidos 64](#_Toc279935826)

[6.2. Limitações de Escopo 65](#_Toc279935827)

[6.3. Dificuldades 65](#_Toc279935828)

[6.4. Trabalhos Futuros 66](#_Toc279935829)

[Referências Bibliográficas 67](#_Toc279935830)

Lista de Figuras

Figura 2.1. Dogma Central da Biologia Molecular.......................................................... 6

Figura 2.2. Definição das Epístases Biológica, Genética e Estatística............................... 14

Figura 3.1. Formato da Base de Dados........................................................................... 27

Figura 3.2. Polymorphism Interaction Analysis.............................................................. 28

Figura 3.3. Multifactor Dimensionality Reduction.......................................................... 33

Figura 3.4. Entropy-based SNP-SNP Interaction Method................................................ 36

Figura 3.5. Motivação para o Algoritmo MASS.............................................................. 40

Figura 3.6. Multi-Approach SNP-SNP Interaction Analysis............................................. 41

Figura A.1. Página Inicial.............................................................................................. 76

Figura A.2. Tela de Análise........................................................................................... 77

Figura A.3. Resultados na Tela (parte 1)........................................................................ 78

Figura A.4. Resultados na Tela (parte 2)........................................................................ 78

Figura A.5. Download dos Resultados............................................................................ 79

Figura A.6. Resultados por Email.................................................................................. 79

Figura A.7. Resultados no Formato Excel...................................................................... 80

Figura A.8. Resultados no Formato de Texto................................................................. 80

Figura C.1. Modelos Epistáticos 0 até 23........................................................................ 99

Figura C.2 Modelos Epistáticos 24 até 47....................................................................... 100

Figura C.3 Modelos Epistáticos 48 até 69....................................................................... 101

Lista de Tabelas

Tabela 2.1. Tabela de Códons........................................................................................ 8

Tabela 2.2. Exemplificação da Epístase.......................................................................... 12

Tabela 5.1. Trecho dos Resultados Detalhados do Método MASS.................................... 56

Tabela B.1. Taxas de Acerto (%) para 20 SNPs e 200 Indivíduos..................................... 82

Tabela B.2. Taxas de Acerto (%) para 20 SNPs e 400 Indivíduos..................................... 83

Tabela B.3. Taxas de Acerto (%) para 20 SNPs e 800 Indivíduos..................................... 84

Tabela B.4. Taxas de Acerto (%) para 20 SNPs e 1600 Indivíduos................................... 85

Tabela B.5. Taxas de Acerto (%) para 30 SNPs e 200 Indivíduos..................................... 86

Tabela B.6. Taxas de Acerto (%) para 30 SNPs e 400 Indivíduos..................................... 87

Tabela B.7. Taxas de Acerto (%) para 30 SNPs e 800 Indivíduos..................................... 88

Tabela B.8. Taxas de Acerto (%) para 30 SNPs e 1600 Indivíduos................................... 89

Tabela B.9. Taxas de Acerto (%) para 40 SNPs e 200 Indivíduos..................................... 90

Tabela B.10. Taxas de Acerto (%) para 40 SNPs e 400 Indivíduos................................... 91

Tabela B.11. Taxas de Acerto (%) para 40 SNPs e 800 Indivíduos................................... 92

Tabela B.12. Taxas de Acerto (%) para 40 SNPs e 1600 Indivíduos................................. 93

Tabela B.13. Taxas de Acerto (%) para 50 SNPs e 200 Indivíduos................................... 94

Tabela B.14. Taxas de Acerto (%) para 50 SNPs e 400 Indivíduos................................... 95

Tabela B.15. Taxas de Acerto (%) para 50 SNPs e 800 Indivíduos................................... 96

Tabela B.16. Taxas de Acerto (%) para 50 SNPs e 1600 Indivíduos................................. 97

Lista de Gráficos

Gráfico 5.1. Eficácia dos Métodos para 20 SNPs............................................................. 51

Gráfico 5.2. Eficácia dos Métodos para 30 SNPs............................................................. 52

Gráfico 5.3. Eficácia dos Métodos para 40 SNPs............................................................. 53

Gráfico 5.4. Eficácia dos Métodos para 50 SNPs............................................................. 54

Gráfico 5.5. Eficácia dos Métodos em Relação aos Parâmetros Herdabilidade e MAF....... 55

Gráfico 5.6. Tempo de Execução Geral.......................................................................... 58

Gráfico 5.7. Tempo de Execução Relativo...................................................................... 59

Gráfico 5.8. Tempo de Execução das Fases de cada Método............................................ 60

Lista de Quadros

Quadro 3.1. Variáveis da Base de Dados........................................................................ 27

Quadro 3.2. Métricas Baseadas na Tabela de Contingência............................................. 32

Quadro 3.3. Métricas Baseadas em Separação................................................................. 32

Quadro 3.4. Métricas do ESNP2.................................................................................... 37

Quadro 3.5. Métricas Adicionais do MASS.................................................................... 41

Quadro 4.1. Detalhes da Herdabilidade e MAF dos Modelos Epistáticos.......................... 46

Quadro 4.2. Parâmetros Gerais da Criação do Conjunto de Dados................................... 46

Quadro 4.3. Parâmetros de Execução do PIA.................................................................. 49

Quadro 4.4. Parâmetros de Execução do MDR................................................................ 49

Quadro 4.5. Parâmetros de Execução do MASS.............................................................. 49

Nomenclaturas e Padrões

**Notação Biológica:** Termos biológicos tais como: nomes de genes ou proteínas, nome de testes, procedimentos, partículas específicas, componentes celulares e outros serão representados em itálico.

**Palavras Estrangeiras:** Termos advindos de outros idiomas que não o português (Brasil) serão representados em itálico.

**Siglas e Acrônimos:** As siglas e acrônimos utilizadas sem definição de seu significado estão detalhadas na seção de Siglas e Acrônimos deste documento. Nomes de algoritmos ou métodos que não consistem em siglas (Ex. Algoritmo Apriori-Gen ou Base de Dados AlzGene) serão definidos no próprio texto e representados em itálico.

**Notas de Rodapé:** As notas de rodapé serão utilizadas para explicar fatos especiais tal como referências ambíguas, etc. Todos os conceitos especiais das ciências da computação e biológicas que não são explicados no texto serão definidos no glossário deste documento.

**Parâmetros:** Todos os parâmetros dos métodos implementados serão representados em letra capitalizada e entre «aspas francesas». Esses parâmetros foram definidos pelos autores da publicação original dos métodos e não são necessariamente os mesmos disponíveis no sistema Web implementado neste trabalho.

**Endereços Virtuais:** Endereços virtuais serão representados entre <menor que e maior que>.

Siglas e Acrônimos

**1L:R:** Single-Gene Recessive Model

**APD:** Absolute Probability Difference (Diferença de Probabilidade Absoluta)

**CART:** Classification and Regression Tree (Árvore de Classificação e Regressão)

**dbSNP:** Single Nucleotide Polymorphism Database (Base de Dados de Polimorfismos de Único Nucleotídeo)

**DNA:** Desoxyribonucleic Acid (Ácido Desoxirribonucléico)

**EC:** Evaporative Cooling Algorithm

**ESNP2, ENSP2-S, ESNP2-Mx:** Entropy-Based SNP-SNP Interaction Method, Entropy-Based SNP-SNP Interaction Method Standard Option, Entropy-Based SNP-SNP Interaction Method Model Option

**FN:** False Negative (Falso Negativo)

**FP:** False Positive (Falso Positivo)

**GWAS:** Genome-Wide Association Studies (Estudo de Associação Genômica Ampla)

**KNN:** k-Nearest Neighbors Algorithm

**LD:** Linkage Disequilibrium

**LOO:** Leave-One-Out Procedure

**MAF:** Minor Allele Frequency

**MDR:** Multifactor Dimensionality Reduction

**MOD:** Modifying-Effect Model

**mRNA:** Messenger RNA (RNA Mensageiro)

**MUDICA:** Multiblock Discriminant Correspondence Analysis

**PCA:** Principal Component Analysis (Análise de Componentes Principais)

**PCA-BCIT:** Principal Component Analysis Bootstrap Confidence Interval Test

**PIA:** Polymorphism Interaction Analysis

**RNA:** Ribonucleic Acid (Ácido Ribonucléico)

**SNP:** Single Nucleotide Polymorphism (Polimorfismo de Único Nucleotídeo)

**SURF:** Spatially Uniform ReliefF

**SVM:** Support Vector Machines (Máquinas de Vetores de Suporte)

**TN:** True Negative (Verdadeiro Negativo)

**TP:** True Positive (Verdadeiro Positivo)

**tRNA:** Transport RNA (RNA Transportador)

**TURF:** Tuned ReliefF

**WTCCC:** Wellcome Trust Case Control Consortium

Glossário

**Aminoácido:** Molécula orgânica formada por átomos de carbono, hidrogênio, oxigênio, nitrogênio e possivelmente enxofre. É o componente básico das proteínas.

**Autovalores:** Juntamente com os autovetores formam um conjunto de informações importantes a respeito de uma matriz de dados. É utilizado como medida de importância dos atributos de uma base de dados no método PCA.

**Base de Dados:** Neste trabalho o termo "Base de Dados" está sendo utilizado com dois significados que dependem do contexto: um "repositório de dados", geralmente online, que contém informações biológicas tais como o *dbSNP* e uma entrada para um algoritmo de identificação de interações entre polimorfismos consistindo de uma matriz de indivíduos por características genéticas ou ambientais.

**Biologia Sistêmica:** É o estudo das interações dos componentes de um sistema biológico e de como essas interações fazem emergir as funções metabólicas.

**Bioquímica:** É a ciência que estuda os processos químicos que ocorrem nos seres vivos.

**Busca Gulosa:** Busca pertencente ao paradigma guloso, onde são realizadas escolhas locais ótimas em cada etapa do algoritmo.

**Conjunto de Dados:** Neste trabalho o termo "Conjunto de Dados" está sendo usado para se referir a um conjunto de bases de dados.

**Dados Faltosos:** Em uma base de dados, são chamados de dados faltosos as informações que, por algum motivo tal como limitações do processo de coleta, foram perdidas.

**Diplótipo:** Diplótipo está para haplótipo assim como genótipo está para alelo. É a combinação específica dos dois haplótipos.

**Eucariotas:** Conjunto dos seres vivos que possuem a célula com núcleo bem organizado, rodeado por uma membrana, com DNA separado do citoplasma e com várias organelas.

**Fenótipo:** É a expressão do genótipo no meio ambiente para um dado indivíduo, ou seja, suas características observáveis.

**Filtro:** Algoritmo de aprendizagem de máquina utilizado durante a fase de pré-processamento de uma determinada base de dados, geralmente transformando esses dados de forma a adequá-los à análise posterior.

**Genética:** Ramo da biologia que estuda como as características biológicas são transmitidas de geração para geração.

**Genômica:** Subárea da biologia que estuda os fenômenos relativos ao conjunto completo de informações genéticas dos seres.

**Genótipo:** É o nome dado ao conjunto dos genes de um organismo. Também é utilizado neste texto como uma determinada combinação de alelos.

**Haplótipo:** Uma combinação específica de alelos em diversos *loci* genéticos.

**Heurística:** Parte de alguns métodos computacionais onde, apesar da exploração ser realizada de forma algorítmica, a avaliação do resultado é puramente empírica.

**Holístico:** Paradigma metodológico onde a abordagem de solução leva em consideração que o problema não pode ser explicado pela subdivisão do mesmo em componentes menores e devendo ser visto como um "todo".

**Interactômica:** Disciplina que estuda as interações entre componentes biomoleculares de uma célula.

***K-Nearest Neighbors*:** Método de aprendizagem de máquina onde uma dada instância é classificada a partir da classe das k instâncias mais próximas.

***Leave-One-Out*:** Forma de validação cruzada onde o conjunto de teste é composto por 1 elemento.

**Ligação Fosfodiéster:** É um tipo de ligação covalente produzida entre dois grupos hidroxila de um grupo fosfato e duas hidroxilas de outras duas moléculas formando uma ligação éster.

***Linkage Disequilibrium*:** É a associação não randômica entre alelos de dois ou mais *loci* não necessariamente no mesmo cromossomo.

***Locus*:** É o local específico no cromossomo onde está localizado determinado gene ou marcador genético. O plural é nomeado *loci*.

**Lógica de Primeira Ordem:** Sistema lógico que estende a lógica proposicional e tem poder expressivo suficiente para formalizar praticamente toda a matemática.

**Maldição da Dimensionalidade:** Dada uma tabela de genótipo-fenótipo, esse termo define o problema ocorrido pela adição de dimensões a essa tabela (considerando novas variáveis genéticas) fazendo com que cada célula contenha cada vez menos informação suficiente para modelar uma determinada condição fenotípica.

**Mendeliano:** Refere-se aos trabalhos de Mendel a respeito da chamada genética clássica publicados em 1865 e 1866. Esses estudos são relacionados à transmissão hereditária das características de um organismo a seus filhos.

**Metabolômica:** É uma abordagem global que tem como objeto de estudo o conjunto de todos os metabolitos de uma célula (metaboloma) de uma célula, tecido ou organismo.

**Microarranjo:** Pequenos chips que contêm material genético utilizado em geral para medir a expressão gênica celular sob as mais diversas condições. Um chip deste tipo utilizado para detectar polimorfismos é chamado de microarranjo de SNP (do inglês, SNP *microarray*).

***Mutual Information*:** Em ambas as teorias da probabilidade e da informação, esse termo é utilizado para nomear a quantificação da dependência mútua entre duas variáveis aleatórias.

***Naive Bayes*:** Termo em inglês para o classificador Bayes ingênuo, que consiste na aplicação simples do teorema de Bayes. É dito ingênuo porque assume fortemente que os dados sejam independentes.

**Nucleotídeo:** Compostos orgânicos formados por uma pentose um grupo fosfato e uma base nitrogenada que servem de componente básico nas moléculas de DNA e RNA. Quando compõem o DNA essas bases podem ser a adenina, guanina, citosina e timina. Quando compõem o RNA as bases são as mesmas substituindo a timina pela uracila.

**Polímero:** Compostos químicos de alta massa molecular formadas a partir de unidades estruturais menores (monômeros).

**Ponte de Hidrogênio:** Tipo de interação que ocorre entre moléculas que possuem átomos de hidrogênio ligados a átomos de flúor, oxigênio ou nitrogênio.

**Redes Neurais Artificiais:** São algoritmos computacionais que resolvem problemas como o da classificação de dados a partir de redes compostas por unidades estruturais chamadas neurônios imitando a forma como a informação é processada pelo cérebro.

**Reducionismo:** Paradigma metodológico onde a abordagem de solução leva em consideração que o problema pode ser explicado por componentes menores e, portanto, deve ser subdividido em problemas menores cujas soluções serão posteriormente unidas para formar a solução geral.

**Sequenciamento:** O sequenciamento do DNA é formado por uma série de métodos bioquímicos que têm como finalidade a determinação da ordem das bases nitrogenadas A, C, G e T num trecho específico da molécula de DNA.

**Sobreajuste (*Overfit*):** Ocorre quando uma metodologia estatística descreve erros aleatórios ou ruídos nos dados ao invés das relações reais. No contexto computacional, geralmente é causado pelo excesso de aprendizado do método, sendo que este se adéqua especificamente aos dados aprendidos, sendo incapaz de generalizar a solução para outros conjuntos de dados.

***Suport Vector Machine*:** Método de aprendizado supervisionado que pode ser utilizado para classificação ou regressão que consiste basicamente em ajustar a dimensionalidade do problema para um valor onde os dados são mais facilmente classificados.

**Trio:** Estudos de trios (do inglês, *case parent trios*) são estudos onde se associa alguma condição fenotípica a três indivíduos sendo eles o pai, a mãe e o filho, visando procurar relações hereditárias diretas.

**Validação Cruzada:** Estratégia de aprendizagem de máquina que consiste na divisão do conjunto de dados original em dois conjuntos: treino e teste. O conjunto de treino é utilizado para treinar o algoritmo de classificação em questão enquanto o de teste é utilizado para verificar a taxa de acerto. Essa divisão geralmente é realizada diversas vezes, cada vez com um conjunto de teste diferente, até que todos os trechos dos dados tenham sidos utilizados como teste. Em inglês é nomeado *n-fold cross-validation*, sendo o "n" o número de intervalos obtidos na divisão dos dados.

***Wrapper*:** Algoritmo de pré-processamento que ocorre externamente ao método de classificação utilizado. Basicamente, são gerados subconjuntos candidatos do conjunto de treinamento em questão para serem avaliados a respeito da precisão total resultante do algoritmo.

1. Introdução

"Good writing is like a windowpane."

- George Orwell

Neste capítulo introdutório serão discutidas a importância deste estudo computacional no contexto das ciências biológicas, suas principais contribuições para estudos do mesmo gênero e a estrutura deste documento de uma maneira geral.

## 1.1. Importância do Estudo

Os Polimorfismos de Único Nucleotídeo (SNPs) ou simplesmente polimorfismos são tipo de variação genética mais comum entre os seres humanos. Alguns SNPs estão sendo apontados como causadores de certas características fenotípicas, entre elas estão o maior risco de desenvolvimento de algumas patologias como o Alzheimer [Bertram et al. 2007; Williams et al. 2010], transtorno bipolar [Fallin et al. 2005], esquizofrenia [Aguiar-Pulido et al. 2010; Fallin et al. 2005], câncer de mama [Feng e Radivoyevitch 2009; Onay et al. 2006; Pharoah et al. 2007; Ritchie et al. 2001], câncer de cólon e reto [Goodman et al. 2006; Mukherjee et al. 2008], hepatite crônica [Saangyong et al. 2009], leucemia [Guyon et al. 2002], resistência da malária [Dong et al. 2007], asma [Boulesteix et al. 2007], artrite reumatóide [Briggs et al. 2010; Peng et al. 2010], diabetes [Ban et al. 2010; Nakamichi et al. 2004], autismo [Allen-Brady et al. 2008; Maestrini et al. 2010; Stone et al. 2007], degeneração macular [Jiang et al. 2009], doenças autoimunes [Sirota et al. 2009], doenças cardíacas [Kooperberg e Ruczinski 2005] e várias outras. E também fatores como a reposta negativa a tratamentos com drogas ou terapias [Chatterjee e Carroll 2005; Chu e Chen 2008; McKinney et al. 2009] ou reações alérgicas [Boulesteix et al. 2007]. Por estarem associadas com doenças complexas e tão comuns, o número de estudos deste tipo está crescendo exponencialmente.

A associação entre características genéticas e ambientais com certos padrões fenotípicos tornou-se possível devido às melhorias nas tecnologias de sequenciamento. Novas estratégias de obtenção de SNPs [Macdonald et al. 2005] estão propondo maneiras eficientes e de baixo custo para o sequenciamento em alto desempenho de sítios variantes do genoma humano através do uso de conhecimento a respeito desses polimorfismos. Como o número de estudos sobre variabilidade genética cresce bastante, é natural que a quantidade de técnicas de obtenção de dados cresça na mesma proporção [Chakravarti 1998].

Diversos artigos revisam periodicamente a importância desse tipo de estudo. São destacados [Moore et al. 2005; Moore et al. 2006; Moore 2007; Moore e Ritchie 2010; Moore et al. 2010]. Todas as publicações citadas concordam que são muitos os motivos desencadeadores de estudos que traçam um paralelo entre polimorfismos e o aumento do risco de desenvolvimento de doenças comuns, porém que fogem ao esquema mendeliano simples, entre eles podemos citar:

* Pacientes poderão ser mais facilmente diagnosticados com fontes de dados mais precisas e estatisticamente relevantes.
* Medicinas e terapias preventivas poderão ser desenvolvidas em indivíduos com maior propensão a certas patologias.
* Os médicos e terapeutas encaminharão os pacientes a terapias com maior chance de sucesso, ganhando bastante no fator tempo, muitas vezes valioso em casos clínicos mais severos.
* Os médicos, farmacêuticos e terapeutas terão novas ferramentas para que a dosagem de certos fármacos seja aplicada de forma correta. Muitas dosagens são feitas atualmente de forma gradativa, novamente isso pouparia o tempo de espera de resposta e reduziria drasticamente o óbito por dosagem irregular [Jones 2007].
* Novos subtipos de patologias aparecem naturalmente nessas pesquisas, isso contribui para pesquisas cada vez mais focadas.
* Os dados genéticos podem ser utilizados para criar uma anotação genética das patologias, isto é, os médicos poderão procurar o histórico de casos semelhantes a partir de características muito mais confiáveis do que o próprio fenótipo.
* Algumas doenças estão sendo relacionadas e outras estão sendo desassociadas, permitindo o entendimento mais aprofundado de efeitos colaterais, relação entre patologias e relação entre patologias e tratamentos.
* Testes como os de alergia poderão ser feitos sem que o usuário entre de fato em contato com a substância.
* O desenvolvimento de novos fármacos terá uma nova e poderosa ferramenta, além dos testes *in vivo* ganharem características muito menos cruéis do que testes simples de observação de efeitos.

Além disso, estudos de associação tal qual o proposto neste trabalho, são passos importantes no desenvolvimento da chamada medicina personalizada, onde os pacientes são tratados com base no seu genoma. Esta é uma realidade não muito distante, que já é utilizada por alguns fabricantes de fármacos [Jones 2007]. Cientistas importantes da atualidade como J. Craig Venter defendem o fato de que, no futuro, existirão marcadores genéticos bem estabelecidos que calcularão o risco absoluto de um indivíduo de desenvolver certa doença dada parte do seu genoma (obtida, por exemplo, através da saliva) e métodos que calculem fatores de risco relativos e populacionais [Pauline et al. 2009]. Os principais motivos pelos quais esta revolução ainda não se tornou uma realidade tão presente são às restrições éticas e sociais. Segundo [Khoury et al. 2010] a ciência já possui conhecimentos suficientes, porém a manipulação dessas técnicas deve ser realizada de forma bastante cuidadosa, para que os diagnósticos não violem a privacidade dos indivíduos e os resultados não sejam mal interpretados.

## 1.2. Contribuições

Este trabalho fornece três contribuições para pesquisas do tipo estudo de associação, que relacionam fatores genéticos tais como interação entre SNPs ou interações estatísticas entre genes (epístase) com características fenotípicas como alto risco de desenvolvimento de certa patologia ou grau de resposta a uma terapia ou fármaco: um estudo comparativo entre técnicas relativamente recentes para análise de interações entre polimorfismos, o desenvolvimento de um novo algoritmo que enfatize as vantagens e atenue os pontos fracos destes métodos estudados e a construção de uma ferramenta Web a partir destes métodos.

No estudo comparativo, três métodos computacionais serão analisados sob sua capacidade de encontrar SNPs que interagem no meio de outros que possuem apenas relações aleatórias entre si. Além disso, também serão comparados os tempos computacionais dos algoritmos. Este estudo pretende apontar os aspectos positivos e negativos de cada método mostrando também as semelhanças e diferenças entre eles.

O novo algoritmo pretende combinar os melhores aspectos segundo a análise comparativa e ao mesmo tempo resolver os problemas gerados por cada abordagem, resultando em um método que consiga detectar combinações interessantes entre SNPs e seja computacionalmente eficiente no que diz respeito ao uso de memória disponível e tempo de execução.

Os métodos estudados estarão disponíveis publicamente para que pesquisadores possam realizar suas análises em uma ferramenta Web. Os principais aspectos considerados na construção deste engenho são: a facilidade do uso dos algoritmos, leitura e obtenção das respostas, a segurança com que os dados que muitas vezes são sigilosos serão manipulados e a eficiência com que o algoritmo é implementado, utilizando todas as otimizações cabíveis.

## 1.3. Estrutura do Documento

Na seção Contexto será realizada uma revisão bibliográfica fornecendo toda a base biológica e computacional necessária para o entendimento da totalidade deste trabalho. Na seção Métodos serão detalhados os algoritmos estudados e implementados com seus respectivos parâmetros e variações, será mostrado o novo método desenvolvido e também será definido o formato da entrada dos algoritmos.

No capítulo Experimentos será mostrado o esquema utilizado para comparar os métodos, definindo o conjunto de dados e o procedimento experimental. No capítulo seguinte serão mostrados os principais resultados a respeito da eficácia do algoritmo e do tempo de execução. Será realizada também uma discussão a respeito desses resultados. As considerações finais a respeito do trabalho juntamente com as dificuldades, limitações do escopo e propostas futuras serão discutidas na seção Conclusão.

Nos apêndices serão exibidos os resultados em sua totalidade, a definição completa dos modelos epistáticos utilizados no processo experimental e um guia para utilização do engenho Web.

"All science is experiential, but all experience must be related back to and derives its validity from the conditions and context of consciousness in which it arises, i.e., the totality of our nature."

- [Wilhelm Dilthey](http://www.brainyquote.com/quotes/quotes/w/wilhelmdil311644.html)

2. Contexto

Este capítulo tem como objetivo o fornecimento da teoria computacional e biológica necessária para o entendimento deste trabalho e sua contextualização a respeito das áreas do conhecimento envolvidas, a saber, genômica, biologia sistêmica, aprendizagem de máquina e mineração de dados.

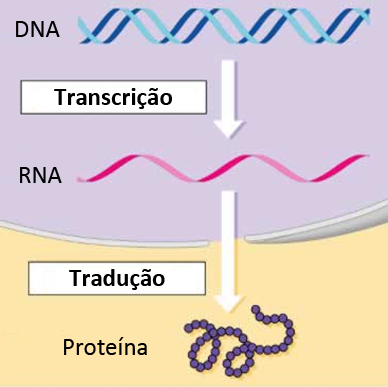
No primeiro momento será explicada a teoria biológica devido ao fato de que ela é a base de entendimento do problema. A seguir será explicitada a importância de análises deste gênero em paradigmas tais como os estudos de associação e sob o ponto de vista das teorias biológicas e estatísticas da epístase. Finalmente será introduzido o contexto computacional, principal foco do trabalho, juntamente com uma rica revisão literária a respeito dos métodos disponíveis para a análise de interações entre polimorfismos e suas relações com doenças complexas.

## 2.1. Dogma Central da Biologia Molecular

A Biologia Molecular consiste no estudo das relações biológicas no nível molecular. O principal foco desta área de conhecimento, que intersecta com áreas como bioquímica e genética, é o estudo do material genético contido dentro da célula (ou do núcleo celular nos seres eucariotas) e de seu produto de expressão, as proteínas. Este tópico, na sua totalidade, será baseado em [Alberts et al. 2002].

O dogma central da biologia molecular rege o procedimento que ocorre dentro da célula responsável pela criação de proteínas a partir do material genético fundamental, o DNA. Este procedimento contém três etapas fundamentais: a replicação do DNA, transcrição do DNA em mRNA e tradução do mRNA em proteínas. O procedimento completo contém ainda mais etapas porém esse nível de detalhamento escapará do foco principal deste trabalho, portanto serão detalhados neste item apenas essas três etapas principais e seus produtos elementares.

A Figura 2.1 exibe o esquema do dogma central da biologia molecular. Nos próximos subtópicos serão descritas as estruturas envolvidas neste procedimento e também as fases necessárias para o entendimento do trabalho.



**Figura 2.1.** Dogma Central da Biologia Molecular

Fonte: Addison Wesley Longman (com tradução livre)

### 2.1.1. Estruturas envolvidas

*2.1.1.1. Cromossomo:* Nos seres eucariotas os cromossomos estão dispostos dentro do núcleo celular. No caso dos seres humanos, objeto de estudo deste trabalho, existem 23 pares de cromossomos, fazendo com que sejam intitulados seres diplóides. Cada cromossomo é caracterizado como uma longa molécula de DNA enlaçada em proteínas estruturais chamadas histonas. O cromossomo contém o material genético de um ser, que será utilizado para criar as proteínas necessárias para o funcionamento metabólico correto do organismo. O conjunto de todos os cromossomos de um ser é chamado de genoma.

*2.1.1.2. DNA:* A molécula de DNA consiste em dois longos polímeros compostos por unidades estruturais chamadas nucleotídeos. O seu *backbone* é formado pelos trechos de açúcar e fosfato e ligados entre seus componentes vizinhos por ligações fosfodiéster. Ligada à molécula de açúcar está uma das quatro bases nitrogenadas possíveis: adenina (A), citosina (C), guanina (G) ou timina (T). Os dois polímeros ligam-se entre si através dessas bases tendo duas possíveis ligações: adenina-timina consistindo de duas pontes de hidrogênio e citosina-guanina consistindo de três pontes de hidrogênio. É exatamente este longo código, formado por essas quatro letras que representam as bases nitrogenadas (formando os nucleotídeos), que corresponde à sequência genética de cada indivíduo.

*2.1.1.3. RNA:* A molécula de RNA consiste numa cadeia simples (e não dupla como o DNA) de nucleotídeos bastante semelhante ao DNA com exceção de que a base nitrogenada timina é substituída pela base nitrogenada uracila (U). Existem quatro classes de RNA: RNA mensageiro, RNA transportador, RNA ribossômico e RNA pequenos nucleares. No âmbito deste trabalho só serão definidos os RNA mensageiros (mRNA) que é o resultado do processo de transcrição e entrada para o processo de tradução, e o RNA transportador (tRNA) que auxilia o processo da tradução.

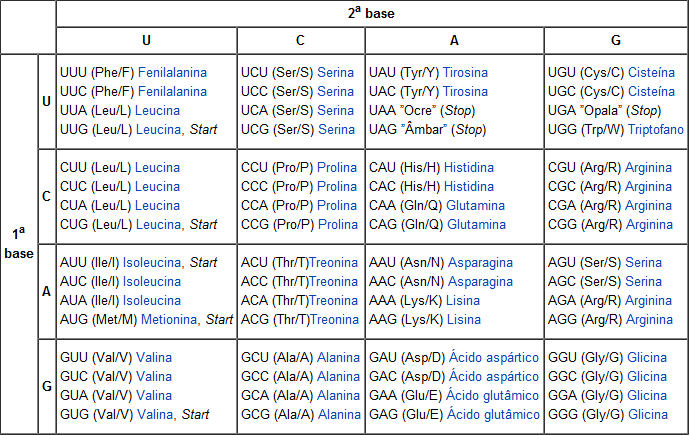
*2.1.1.4. Ribossomo:* O ribossomo é uma proteína que auxilia o processo da tradução. É nela que o mRNA se acopla para iniciar o procedimento de criação das proteínas. Elas estão presentes fora do núcleo celular, no citoplasma da célula.

*2.1.1.5. Proteína:* São compostos químicos de alto peso molecular formados por uma longa cadeia de aminoácidos correspondendo a aproximadamente 80% do peso seco de uma célula. As proteínas são responsáveis por diversas atividades fundamentais para a manutenção da vida tais como: função enzimática (otimizam reações químicas dentro da célula), funções estruturais, sinalização da célula, respostas imunológicas, funções metabólicas na célula, entre diversas outras. A função das proteínas está diretamente relacionada à sua estrutura, fazendo com que pequenas alterações estruturais inutilizem a molécula ou causem funcionamento irregular da mesma, ocasionando desequilíbrios metabólicos letais ao organismo.

### 2.1.2. Fases do dogma central

*2.1.2.1. Replicação do DNA:* A replicação do DNA consiste na cópia desta molécula por uma proteína chamada *DNA polimerase*. Esta cópia é essencial para processos metabólicos comuns e básicos como a divisão celular. Este processo consiste na leitura de uma fita específica do DNA (chamada de senso ou *template*) e a molécula resultante da replicação é idêntica à outra fita (o anti-senso). Alguns procedimentos de detecção e correção de erro dentro desta fase asseguram que a fita gerada possua o máximo possível de fidelidade com o DNA original, porém ainda assim, erros podem ocorrer, e como será visto posteriormente neste trecho, estes "erros" são o principal objeto de estudo deste trabalho.

*2.1.2.2. Transcrição:* A transcrição é a etapa responsável pela geração de uma molécula de RNA a partir de um trecho específico da molécula de DNA. Nesta fase, uma proteína chamada *RNA polimerase* se acopla a um trecho específico da molécula de DNA (procedimento regulado por diversos mecanismos presentes nas regiões promotoras) iniciando o processo de criação do RNA mensageiro. A *RNA polimerase* então percorre toda a extensão necessária até que, em um certo trecho, ela se desestabiliza e interrompe esta produção. O mRNA resultante deste processo irá migrar para fora do núcleo celular para iniciar a etapa de tradução. Em termos gerais essa extensão do DNA que foi "lido" é chamada de gene que geralmente está associado, individualmente ou em conjunto com outros genes, à produção de uma proteína. Existem etapas intermediárias entre a transcrição e a tradução tal como o *splicing*, onde são cortadas certas regiões do mRNA chamadas *íntrons* e conectados os trechos restantes chamados *exons*, porém tal nível de detalhamento foge ao escopo deste projeto e não será abordado.



**Tabela 2.1.** Tabela de Códons

Fonte: <<http://bit.ly/dQFUGo>>

Representa a associação entre todos os possíveis códons com os aminoácidos existentes. Por exemplo, a combinação UUU (topo da primeira célula) é responsável pela adição do aminoácido *Fenilalanina* na cadeia. É interessante observar dois casos: no primeiro não importando qual a última base da trinca, caso os dois primeiros sejam UA o aminoácido gerado será a *Tirosina*, já caso os dois primeiros sejam UU, é necessária a determinação do último para saber se o aminoácido resultante será a *Fenilalanina* ou a *Leucina*.

*2.1.2.3. Tradução:* A tradução inicia quando o mRNA é acoplado ao ribossomo no citoplasma da célula. Cada trinca de bases do mRNA (códon) é "lida" pelo ribossomo que acopla um aminoácido correspondente à trinca na sequência de aminoácidos sendo gerada, resultando no produto final: a proteína. Existe uma trinca específica que corresponde ao início desta produção e existe um códon específico que determina o fim deste procedimento. Os RNA transportadores são responsáveis por armazenar cada aminoácido que será posteriormente acoplado. Eles são estruturas formadas por um códon específico de um lado e um aminoácido de outro e estão presentes em número muito grande no citoplasma. Quando determinado códon do mRNA é lido, um tRNA eventual que esteja próximo do ribossomo é alinhado com este códon, acarretando na junção do aminoácido que está em uma de suas extremidades à sequência de aminoácidos corrente. A Tabela 2.1 mostra todos os possíveis tRNAs associando códons a aminoácidos. Através da observação desta tabela percebe-se que a alteração de certas bases não causam mudanças no aminoácido resultante, porém algumas alterações em apenas uma base podem alterar o aminoácido da sequência para outro com propriedades bioquímicas bastante diferentes, este fato também é muito importante para o entendimento do resto desta pesquisa.

## 2.2. Polimorfismos de Único Nucleotídeo

O objeto de estudo deste trabalho é o Polimorfismo de Único Nucleotídeo (SNP) ou simplesmente polimorfismo. Durante o procedimento de replicação do DNA, como foi mencionado anteriormente, alguns "erros" podem ocorrer. Tais "erros" podem ser: a substituição de uma base por outra, a inserção de uma base ou a remoção de uma base. Esta pequena variação genética pontual pode ser passada para os descendentes do indivíduo no qual ela ocorreu, fazendo com que este genótipo se torne presente em uma pequena parcela da população. De fato, para ser considerado polimorfismo, a variante menos comum deve ser observada em pelo menos 1% da população, porém este conceito está sofrendo modificações, em vista que já existem SNPs catalogados em bases de dados que não cumprem exatamente essa regra [Chakravarti 2001].

O Genoma humano é composto por 3 bilhões de pares de bases, sendo estimados cerca de 10 milhões SNPs, fazendo com que esse tipo de variação genética seja a mais comum nos seres humanos (cerca de 80% de toda a variação). A maioria dos SNPs não causam efeitos observáveis entre os indivíduos, enquanto outros causam apenas mudanças não nocivas tais como diferenças na cor dos olhos ou altura. Os SNPs que não produzem mudanças no fenótipo, em sua maioria, são encontrados em regiões não codificantes do DNA, isto é, regiões que não contêm genes, enquanto outros estão em regiões que são codificadas nos ribossomos, porém cujo aminoácido correspondente ao códon provavelmente é o mesmo ou com propriedades físico-químicas semelhantes ao do códon original. Porém caso ocorra uma deleção ou inserção de alguma base, mudando assim a janela de leitura dos códons chamada *frame*, ou então caso o aminoácido resultante a partir de um códon com uma base modificada possua propriedades totalmente diferentes, o resultado pode ser o mal funcionamento metabólico, levando a diversas complicações tais como patologias complexas ou a resposta ineficiente de um indivíduo a certo tratamento. Tais SNPs nocivos são ditos possuidores de uma mudança não-sinônima entre as bases nitrogenadas.

Essas pequenas modificações podem alterar profundamente a proteína fabricada a partir do gene em questão, mudando drasticamente sua forma tridimensional e consequentemente sua função ou forma de interação com outras proteínas ou complexos protéicos, causando diversas complicações no metabolismo. Há resultados na literatura [Ferrer-Costa et al. 2002] que estimam que, de todos os SNPs não-sinônimos causadores de patologias, 80% causam a doença pela desestabilização da proteína, 10% são nocivos porque afetam a estrutura quaternária da proteína e 10% afetam outros mecanismos como a interação entre as proteínas, etc.

## 2.3. Epístase

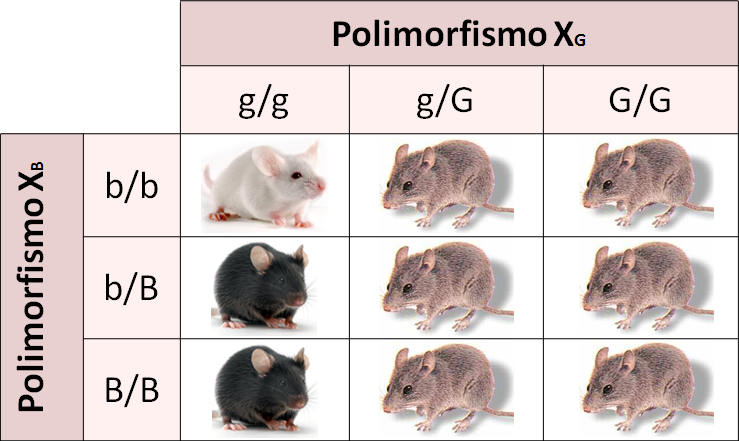
Como mencionado anteriormente, os seres humanos são diplóides, isto é, contêm dois cromossomos com DNA quase idênticos. Cada polimorfismo consiste na mudança de uma base para qualquer outra. Digamos que a variante comum seja X (que pode assumir quaisquer um dos valores A, C, G, T ou nenhum) e a variante rara seja Y podendo assumir qualquer um desses valores mencionados atendo-se à restrição de que X ≠ Y. A cada uma dessas variantes é dado o nome de alelo. Considerando o fato de que existem duas cópias desta base, uma em cada cromossomo, temos quatro possibilidades: ambos os cromossomos contêm X, ambos contêm Y, um contém X e o outro Y e vice-versa. Uma dessas variantes será a mais expressa na população (digamos X) e a outra será mais rara (digamos Y). A ciência genética nomeia essas possíveis combinações alélicas:

* XX: Homozigoto dominante
* YY: Homozigoto recessivo
* XY ou YX: Heterozigoto

Percebe-se então que cada polimorfismo no nosso DNA pode assumir três possibilidades (já que a ordem dos alelos nos heterozigotos não faz diferença na expressão fenotípica resultante). A partir deste ponto podemos definir um conceito bastante importante neste trabalho: a epístase.

O termo epístase foi usado pela primeira vez em [Bateson 1909] para descrever o efeito de mascaramento no qual um alelo de um determinado *locus* impede que outro alelo em outro *locus* manifeste seu efeito. Um pouco mais tarde, no estudo [Fisher 1918] foi cunhada uma definição puramente estatística para a epístase em complemento à abordagem biológica de Bateson. Estas duas definições, em diferentes níveis populacionais, são consideradas por muitos cientistas atuais como o elo que liga estas duas ciências no que diz respeito a estudos onde se pretende extrair informações a respeito de relações entre polimorfismos. Moore mostra que são importantes estudos que pretendem fornecer respostas para questionamentos tais como: "É válido pensar na ocorrência da epístase biológica mesmo quando não se existe epístase estatística, porém evidências estatísticas de epístase implicam necessariamente na ocorrência de epístase biológica?" [Moore 2005].

A definição da epístase no sentido biológico, conforme mostra [Cordell 2002], pode ser ilustrado pela Tabela 2.2. Esta figura representa uma tabela de genótipo-fenótipo, isto é, uma tabela que contém o fenótipo observado para cada combinação de genótipo possível. Neste caso temos dois polimorfismos em dois *loci* diferentes nomeados XG e XB. Esses dois *loci* influenciam na cor dos pelos de um rato, podendo assumir respectivamente, conforme mostrado acima, os estados: homozigoto dominante (GG e BB), homozigoto recessivo (gg e bb) e heterozigoto (Gg e Bb). Caso observemos indivíduos com um dos alelos G, então a cor deste indivíduo será cinza, isto é, o alelo G é dito dominante sobre o alelo g dentro deste *locus*. De forma semelhante, caso o genótipo no locus XG seja gg, o alelo B será dominante sobre b pois sua presença definirá se a cor do indivíduo será preta (contém o alelo dominante) ou branca (não contém o alelo dominante). Porém perceba que caso o genótipo no *locus* XG contenha o alelo G dominante, não importa qual o genótipo do *locus* XB porque o indivíduo sempre será cinza. É dito então que o alelo G no *locus* XG é epistático ao alelo B no *locus* XB. Afrouxando um pouco este conceito dizemos que o polimorfismo XG é epistático em relação ao polimorfismo XB. No caminho inverso dizemos que o polimorfismo XB é hipostático à XG.



**Tabela 2.2.** Exemplificação da Epístase

Fonte: Criação própria com base na Tabela 1 de [Cordell 2002]

A definição estatística de epístase criada por Fisher a explica como desvios da aditividade em modelos estatísticos, que pode ser traduzida como o efeito observado na população através das diferenças nas epístases biológicas entre indivíduos. De maneira geral, são atribuídos valores de penetrância para cada possível combinação de genótipos. A penetrância consiste na probabilidade P(D|G) de um fenótipo D ocorrer dada aquela combinação de genótipos G. É comum o uso de métodos estatísticos paramétricos tais como: regressão linear ou logística; Ou não paramétricos como: partição combinatória ou restrita, análise de associação de conjuntos, programação genética em redes neurais artificiais e técnicas de redução de dimensionalidade para avaliar a não-aditividade indicativa da epístase estatística. Métodos computacionais não paramétricos baseados em disciplinas como a mineração de dados estão sendo cada vez mais usados como alternativa aos métodos paramétricos estatísticos tradicionais, pelo fato de que a epístase no seu contexto estatístico é difícil de ser detectada utilizando ferramentas onde não se conhece o modelo *a priori* [Moore and Williams 2005].

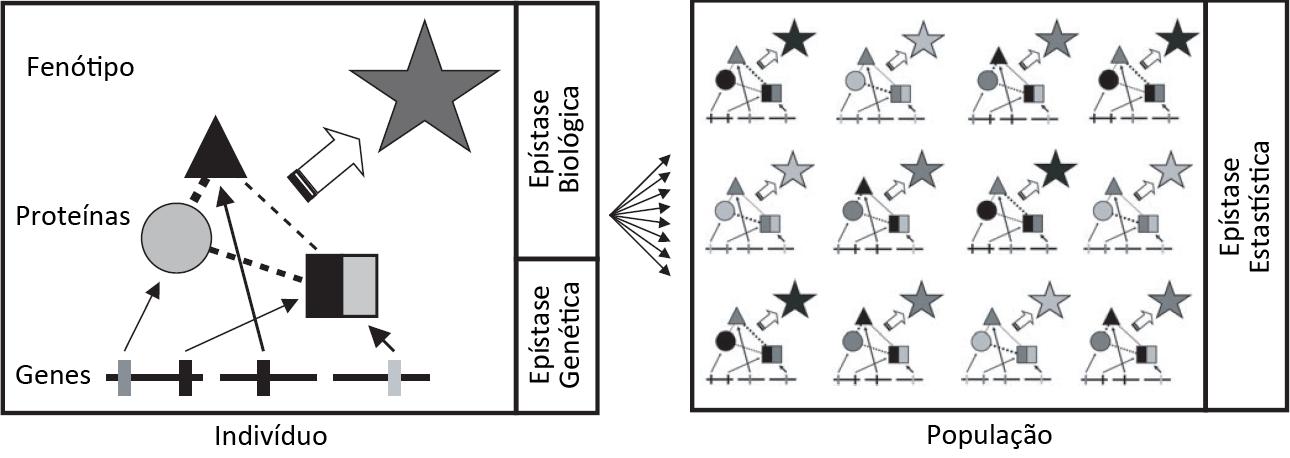
No estudo realizado por [Culverhouse 2002] é retratado o fato de que existe uma grande classe de modelos genéticos onde a abordagem tradicional (procurar associações baseadas em significância sobre modelos que são ajustados aos dados) vai falhar, devido à não presença de aditividade (em conformidade com a descrição estatística) nem variação de dominância em nenhum dos *loci* suscetíveis (conforme definido no exemplo dos ratos, quando havia dominância de um alelo sobre outro dentro do mesmo *locus*).

Em [Moore 2003], este pesquisador, especialista em epístase, mostra a natureza ubíqua dessa relação entre fatores genéticos e fornece evidências interessantes que permeiam sua hipótese:

* Como mostrado anteriormente, a epístase não é um conceito recente, já que há cerca de 100 anos atrás já se formulavam hipóteses que reconheciam esses desvios das frações Mendelianas como sendo ocasionado por um processo de interação entre genes.
* A ubiquidade das interações biomoleculares na regulação dos genes e dos sistemas bioquímicos e metabólicos sugere que relações entre variações na sequência de DNA com a ocorrência de complicações no organismo podem, de fato, envolver relações entre genes.
* Resultados positivos para testes com SNPs tipicamente não se replicam entre amostras independentes. Isso é verdade para ambos os testes a partir de *linkage disequilibrium* e estudos de associação.
* As relações gene-gene e gene-ambiente estão sendo comumente encontradas quando investigadas da forma correta.

Agora que foram definidos de forma clara os dois conceitos de epístase podemos retratar as diferenças entre eles. No estudo [Moore 2005] esta diferença é explicada de forma simples (Figura 2.2) repartindo o conceito da epístase biológica ainda em um subconjunto: o genético.

No próximo tópico será mostrado que o conceito de epístase foi uma das principais ferramentas que os cientistas utilizaram para mudar o paradigma de estudo dos polimorfismos de forma individual para considerar interações entre eles em estudos que associavam essas características genéticas a doenças.



**Figura 2.2 -** Definição das Epístases Biológica, Genética e Estatística

Fonte: [Moore 2005] (com tradução livre)

A epístase genética pode ser pensada como a interação entre trechos no DNA que dão origem a certos fenótipos específicos em um indivíduo. A informação genética afeta o indivíduo através de uma complexa hierarquia de proteínas (círculo, quadrado e triângulo). A interação física entre as proteínas (e outras biomoléculas) e seu impacto no fenótipo (estrela) constitui a epístase biológica. As diferenças nas epístases genéticas e biológicas entre indivíduos dá origem à epístase estatística. As epístases a nível individual (biológica e genética) podem ocorrer sem que necessariamente exista a epístase a nível populacional (estatística), bastando para isso que as variações na sequência de DNA e as biomoléculas entre os indivíduos pertencentes à população sejam as mesmas [Moore 2005]. Porém o contrário é a motivação de diversos estudos que pretendem validar a utilização de técnicas estatísticas e computacionais para explicar o fenômeno epistático.

## 

## 2.4. Estudos de Associação

Dados de polimorfismos são muito utilizados nos mais diversos tipos de pesquisas atualmente. No paradigma classificatório, o mais atual entre todos, podemos citar [Aguiar-Pulido et al. 2010] que realizou uma classificação de pacientes com esquizofrenia, [Sirota et al. 2009] conduziu um estudo a respeito de pacientes com doenças auto-imunes, comparando os efeitos dessas doenças através com sua carga de variabilidade genética. No paradigma estatístico, o mais tradicional, existem uma grande quantidade de estudos. Entre eles [Boulesteix et al. 2007] realizaram um teste múltiplo a respeito de interações SNP-SNP, [Kooperberg e Ruczinski 2005] utilizaram regressão lógica de Monte Carlo, [Li et al. 2009] utilizaram regressão lógica em dados de polimorfismos baseados em casos de trios (dados de genitores e descendentes) e [Peng et al. 2010] criaram uma metodologia baseada em componentes principais para estimar intervalos de confiança com testes de *bootstrap* em dados envolvendo muitos polimorfismos.

Este trabalho, entretanto, está situado no paradigma dos estudos de associação. Neste tipo de estudo são analisadas as configurações dos polimorfismos de diversos indivíduos procurando a associação entre esses fatores genéticos com uma dada característica fenotípica em questão, que geralmente é uma patologia complexa e comum (como as descritas no capítulo introdutório), que foge aos padrões mendelianos simples e raros (como a fibrose cística e anemia falciforme). O termo estudo de caso-controle é utilizado como sinônimo para estudo de associação e se refere ao fato de que neste tipo de estudo geralmente existem dados de pacientes com a patologia em questão (casos) e pacientes onde não se observa tal fenótipo (controles).

No contexto da genômica atual, estudos de associação deste tipo que utilizam dados de variabilidade genética coletados a partir de um grande trecho do genoma de vários indivíduos são intitulados de *Genome-Wide Association Studies* (GWAS). Com o término do sequenciamento completo do ser humano entramos na era pós genômica, marcada por uma grande expectativa no que diz respeito a estudos do tipo GWAS. Infelizmente, apesar da grande expectativa, métodos bioestatísticos revelaram pouquíssimos trechos de variabilidade genética individualmente associados com as doenças complexas e não foram capazes de identificar outros trechos cuja participação no mecanismo de ativação de algumas doenças já se conhecia [Moore et al. 2010].

Tais fracassos, segundo [Moore and White 2007] são fruto da dificuldade de se encontrar tais características no meio de tantas outras, sendo este problema rotulado como "procurar uma agulha num palheiro". Segundo [Moore et al. 2010] os cientistas na época atribuíram esses insucessos a dois motivos especiais:

* Os métodos bioestatísticos utilizados na época são, a partir do seu próprio design, agnósticos ou não enviesados no sentido de que eles ignoram os resultados atuais a respeito da divisão em vias metabólicas dos genes, explicados pela biologia sistêmica.
* A modelagem linear utilizada até então para analisar os dados do GWAS considerou os polimorfismos de forma independente, ignorando o mesmo em seu contexto genômico ou ambiental.

A partir deste momento diversos estudos foram realizados considerando interações entre os polimorfismos ao invés de seus efeitos individuais, tendo a teoria dos efeitos epistáticos como guia que unia as teorias biológicas e estatísticas. Neste momento a bioinformática teve um papel importante nessas análises estando sempre ao lado de, ou até substituindo, métodos bioestatísticos bastante utilizados. Os resultados começaram a se mostrar mais interessantes, fazendo com que os cientistas convergissem para este paradigma de estudo de associação.

Em um capítulo de livro bastante interessante, [Moore e White 2007] revisaram três desafios principais para se conseguir realizar analises promissoras com dados de GWAS:

* Métodos potentes de aprendizagem de máquina e mineração de dados têm que ser desenvolvidos para modelar estatisticamente essas relações entre variações na sequência de DNA e a suscetibilidade a doenças. Métodos tradicionais como regressão logística tem poder limitado para modelar interações não lineares de ordens mais altas devido a aspectos como a maldição da dimensionalidade.
* Na presença de vários SNPs a serem analisados, é necessária uma seleção prévia de alguns deles que sejam mais promissores. Métodos que utilizam filtros e *wrappers* ou até heurísticas de pré-processamento terão um importante papel nestes tipos de análise já que o total de combinações de polimorfismos possíveis está muito além da capacidade computacional de métodos exaustivos.
* Finalmente, são necessários instrumentos para interpretar esses modelos de interação. Mesmo que esses métodos consigam fornecer um modelo interessante de risco de desenvolvimento de uma determinada doença, ele não é tão facilmente convertido em um tipo específico de prevenção à patologia ou estratégia de tratamento da mesma. Portanto, a interpretação no contexto da biologia humana se faz necessária.

Este trabalho propõe o estudo exatamente do primeiro desafio citado, porém cita os outros dois dentro de um escopo resumido.

## 2.5. Análise de Interações entre Atributos

A análise de importância de atributos dentro de um conjunto de dados específico é uma subárea com grande número de pesquisas de métodos na aprendizagem de máquina e na mineração de dados. De fato essa não é uma subárea tão recente como comprova o estudo [Mingers 1988]. A grande motivação de estudos como este e também como [Buntine e Niblett 1992; White e Liu 1994] são as árvores de decisão. Tal estrutura é criada através da seleção de um atributo que melhor explica as classes do conjunto de instâncias do problema. A eficácia de tal estrutura de classificação de dados depende fortemente no algoritmo escolhido para se selecionar o atributo mais interessante dentre todas as possibilidades [Breiman 2001][[1]](#footnote-1).

Em [Mingers 1988] são comparados diversos algoritmos para seleção de atributos: métrica de informação de Quinlan, Chi-quadrado sobre tabela de contingência, estatística G, índice Gini, a medida *Gain-ratio* e a correção de Marshall. Em todos eles cada atributo é atribuído a um valor que representa o quão interessantemente ele divide as classes associadas ao problema. Mais tarde [Buntine e Niblett 1992] realizaram um estudo mais aprofundado sobre a mesma ótica de Mingers. No mais recente dos três, [White e Liu 1994] mostram através de algumas técnicas estatísticas o fato de que algumas métricas, tais como o ganho de informação, são tendenciosas em favor dos atributos com grandes números de valores, sendo preferíveis métricas que utilizem a distribuição de Chi-quadrado como a técnica do Chi-quadrado sobre tabela de contingência.

Essas métricas se conservaram bastante, sendo utilizadas recentemente em diversos algoritmos de construção de árvores de decisão, árvores de classificação e regressão e em algoritmos que criam essas estruturas como componentes básicos de uma floresta randômica. Em [Strobl et al. 2006] o índice Gini é estudado profundamente sugerindo uma abordagem baseada na consideração dos *p-values* obtidos em distribuições derivadas a partir do ganho máximo deste índice de diversidade num contexto de classificação binária e mostra que essa é uma forma eficaz e sem viés para critérios de seleção de atributos em algoritmos de partição recursiva.

No contexto da interação entre atributos, muito importante para a análise de dados de polimorfismos conforme discutido nos itens anteriores, vários autores deram contribuições. Neste item será realizada uma rápida revisão nas métricas baseadas em entropia (como algumas das anteriores para atributos individuais), métricas baseadas em componentes principais e métricas baseadas em técnicas estatísticas como a análise de correspondência.

Aleks Jakulin forneceu um trabalho bastante completo no que diz respeito à análise de interações entre atributos em sua tese de doutorado [Jakulin 2005]. Ele inicia seu trabalho fornecendo a motivação na incerteza sobre os modelos estatísticos que devem ser assumidos baseados na distribuição específica de cada tipo de dado. Posteriormente, ele realiza uma revisão intensa no conceito de entropia utilizado em disciplinas como teoria da informação, e mostra diversos modelos para se construir e visualizar interações das mais diversas naturezas. Essa tese compila o estado da arte na interação entre atributos, discutindo conceitos como os apresentados em trabalhos prévios pelo próprio autor na sequência deste texto.

Em [Jakulin e Bratko 2004a] é discutido que a abordagem mais correta deve ser a de construção de um modelo através de testes que demonstrem a presença de interações ao invés de se assumir previamente modelos específicos de interação e testá-los em sua adequação aos dados. Neste trabalho eles generalizam as análises para interações de qualquer ordem através de dois modelos: um modelo reducionista de arquitetura "parte para o todo" onde o modelo do "todo" é reconstruído por aproximações em modelos de suas "partes" e um método mais holístico onde o "todo" é modelado diretamente.

Em [Jakulin e Bratko 2004b] é discutida a interação entre atributos no contexto da classificação dos dados, principalmente no sentido de se identificar os atributos que são redundantes para esta tarefa. Neste trabalho são discutidos dois conceitos importantes: o de miopia, causado por algoritmos de aprendizagem de máquina que assumem independência dos dados ao invés de interação e o de fragmentação, surgindo pelo motivo oposto, quando se assume interação ao invés de independência dos dados.

Em [Jakulin et al. 2003] são discutidas as interações em um contexto bastante próximo ao deste trabalho, utilizando dados médicos reais a respeito de atroplastia de quadril. É criada então uma nova abordagem baseada em "votação" a partir de classificadores como *Naive Bayes* para se testar a interação entre atributos e uma forma para se levar em consideração tal teste durante a classificação propriamente dita.

Em um contexto mais computacional, [Freitas 2001] discute a importância do entendimento das interações entre atributos para a mineração de dados, através de problemas como: construção de atributos, indução de regras de lógica de primeira ordem, detecção do paradoxo de Simpson e a detecção de regras simples de interação. Ele também enfatiza o fato de que a maioria dos algoritmos de indução de regras é baseada em uma busca gulosa, que não mostra de forma tão elucidativa as interações.

A análise de componentes principais (PCA) foi bastante utilizada principalmente para encontrar polimorfismos individualmente interessantes no contexto da patologia. Segundo [Gauderman et al. 2007] é possível capturar efeitos de LD através do uso de PCA em uma estratégia com duas etapas: na primeira o PCA é utilizado para encontrar os SNPs que capturam a estrutura de correlação na matriz de dados de entrada, enquanto na segunda etapa os componentes principais são utilizados diretamente para se testar a associação com a doença. Já no estudo [Oh e Park 2007] o PCA é utilizado no contexto de testes de associação baseados em haplótipos, dado o conhecimento que haplótipos incorporam informação de LD em múltiplos *loci*. Num estudo ainda mais recente e bastante referenciado, [Peng et al. 2010] utilizam uma técnica denominada PCA-BCIT que utiliza os autovalores obtidos diretamente pelo PCA para se acessar a associação entre combinações de atributos na base de dados e a informação classificatória.

Outra abordagem que está sendo explorada atualmente é a análise de correspondência discriminante [Perrière et al. 1996]. No estudo [Williams et al. 2010] essa idéia é elaborada, sendo criado o MUDICA, utilizado com sucesso para associar os atributos de um conjunto de dados a respeito do mal de Alzheimer.

Além das técnicas mencionadas nesta revisão, existem diversas outras abordagens não citadas neste texto, utilizadas ou não no contexto de detecção de interações entre polimorfismos e efeitos epistáticos. Porém esta linha de pesquisa está se estendendo rapidamente principalmente no que diz respeito à agregação de conhecimentos da biologia sistêmica pura. Dados de vias metabólicas e interações entre componentes biológicos de qualquer ordem estão moldando a ciência da interactômica, que deverá liderar as pesquisas de associação entre relações epistáticas e doenças humanas complexas [Aittokallio e Schwikowski 2006].

## 2.6. Métodos Computacionais para Análise de Interações SNP-SNP

A partir das definições acima, podemos definir o contexto exato deste trabalho como:

* A partir de dados de polimorfismos de único nucleotídeo provindos de estudos de associação, métodos computacionais precisam ser criados para encontrar efeitos epistáticos nos genes onde tais SNPs se encontram de forma a vencer as limitações impostas pela grande quantidade de dados disponível. De forma eficaz e rápida, esses métodos precisam encontrar, através de métricas que quantifiquem e qualifiquem interações entre atributos, as combinações mais promissoras de polimorfismos de uma certa base de dados, para que elas possam ser estudadas com mais profundidade por cientistas da área biológica.

Diversos trabalhos já propuseram métodos para realizar tal tarefa, obtendo resultados bastante interessantes. Nos próximos subtópicos serão revisados alguns desses métodos, que estarão divididos por abordagens. Esta classificação dos métodos em diferentes abordagens é ilustrativa e não quer dizer que métodos em uma abordagem específica não utilizem elementos particulares de outra. No final deste item serão feitas considerações finais a respeito desta breve contextualização teórica retratando as fontes de dados atuais e as perspectivas sobre estudos nesta linha.

### 2.6.1. Estatística tradicional

A própria caracterização do problema se encaixa diretamente em abordagens estatísticas bastante conhecidas como regressão logística ou lógica. Esses métodos tradicionais geralmente são utilizados para analisar dados médicos a nível não-genômico, por exemplo, em estudos que associam indivíduos fumantes ao maior risco de doenças pulmonares ou câncer.

No contexto genético a regressão logística é utilizada de forma pura em estudos como [Nakamichi et al. 2004] ou em associação ou comparação com outros métodos como em [García-Margariños et al. 2009], onde esse método é utilizado ao lado de métodos baseados em árvores de classificação e regressão ou em [He et al. 2009] onde esta técnica é comparada à uma estratégia de redução de dimensionalidade.

A regressão lógica é amplamente proposta pelo fato de sua teoria estar diretamente relacionada com o problema em questão [Schwender e Ickstadt 2007]. Variações são comumente encontradas como a utilização de regressão lógica com Monte-Carlo [Kooperberg e Ruczkinsi 2005] ou a aplicação deste método em dados de trios [Li et al. 2009].

Outras metodologias estatísticas como análises empíricas Bayesianas no seu formato generalizado [Schwender e Ickstadt 2008] ou testes múltiplos [Boulesteix et al. 2007] também são comuns.

### 2.6.2. Redução de dimensionalidade

Esta técnica consiste em reduzir a tabela de genótipo-fenótipo que possui k dimensões para apenas uma dimensão, no formato de um atributo multivalorado que geralmente envolve os possíveis estados de risco no qual a doença pode ocorrer [He et al. 2009].

A adaptação desta técnica, nomeada MDR, como descrita em [Hahn et al. 2003; Ritchie et al. 2001] foi uma das escolhidas neste trabalho pela sua grande referência em outros trabalhos e pela sólida base computacional que possui. Este método, bastante estudado por renomados pesquisadores como Dr. Jason Moore, é alvo de diversos estudos que apontam formas de se manipular seus parâmetros e é um dos favoritos quando se trata de bases de dados muito desbalanceadas (geralmente quando o número de controles é muito maior do que o número de casos) [Velez et al. 2007].

### 2.6.3. Árvores de decisão e florestas randômicas

As árvores de decisão ou de classificação e regressão consistem na estrutura básica para as florestas randômicas, que consistem de uma simples coleção dessas árvores com intuito de classificar uma determinada instância do problema com base em algoritmos de votação. Parte deste algoritmo é baseado numa estratégia de *bootstrapping*, que fornece as bases teóricas [Breiman 2001] necessárias para que esta técnica seja utilizado no contexto de detecção de interações entre atributos genéticos [Bureau et al. 2005; Jiang et al. 2009].

Em revisão recente, [Kim et al. 2009] enfatizaram que este método é tão atrativo porque é capaz de lidar com dados advindos de sequenciamento em larga escala, que majoritariamente não possuem um modelo associado.

### 2.6.4. Funções de pontuação

Métricas baseadas em entropia [Dong et al. 2007] ou baseadas em valores de uma tabela de contingência obtida após uma validação cruzada [Goodman et al. 2006; Mechanic et al. 2008] se mostraram bastante eficientes para detectar padrões epistáticos em bases de dados com os mais variados modelos (não conhecidos a priori) de interação. De fato, este trabalho foca bastante nestes métodos, sendo as duas abordagens citadas no início deste parágrafo escolhidas para estudo e implementação. A nova metodologia que este trabalho propõe é fortemente baseada nestas idéias, discutidas em detalhes no próximo capítulo.

### 2.6.5. Buscas heurísticas

Inevitavelmente, heurísticas são necessárias para que grandes massas de dados possam ser analisadas, fugindo ao padrão lento da estimação de parâmetros da abordagem estatística e ao caráter combinatório da redução da dimensionalidade e das funções de pontuação. Em [Brinza e Zelikovsky 2007] é proposto um algoritmo guloso nos moldes de algoritmos conhecidos de alinhamento de sequências para propor a modelagem de um Fator de Risco como um alinhamento a um diplótipo. Já em [Mao et al. 2009] é utilizado o algoritmo do *Apriori-Gen* em dados de pacientes com a Doença de Crohn.

### 2.6.6. Algoritmos de classificação

Algoritmos de classificação tradicionais como Redes Neurais Artificiais [Gunther et al. 2009; Ritchie et al. 2003] ou Máquinas de Vetores de Suporte [Guyon et al. 2002] também são amplamente utilizados. Geralmente nessas abordagens é realizado primeiramente o aprendizado dos dados a partir da própria base de dados obtida no estudo ou a partir de bases de dados conhecidas na literatura. Depois são criadas tabelas de contingência a partir de testes utilizando o procedimento *Leave-One-Out* e calculando as estatísticas relativas aos valores resultantes [Ritchie et al. 2003].

### 2.6.7. Seleção de atributos tradicional

Nesta abordagem foram incluídos basicamente dois grandes grupos de algoritmos. O primeiro é composto pela família *Relief*, algoritmo de seleção de atributos capaz de detectar dependências condicionais entre atributos, tradicionalmente utilizado na fase de pré-processamento de dados em algoritmos de aprendizagem de máquina [Robnik-Sikonja e Kononenko 2003]. Este algoritmo que passou por um processo de evolução gerando novas versões como o *ReliefF* ou o *Tuned ReliefF* (TURF), foi aplicado no contexto genético por [Greene et al. 2009] resultando no *Spatially Uniform ReliefF* (SURF), que combinado com o algoritmo TURF consegue resultados muito melhores do que este último sozinho sob um contexto epistático.

O segundo grupo é formado por uma técnica baseada no modelo físico para evaporação de partículas. O algoritmo de seleção de atributos *Evaporative Cooling* (EC) [McKinney et al. 2007] busca o melhor subconjunto a partir de um certo conjunto de dados de forma a minimizar a quantidade de ruído total, combinando ReliefF e a teoria da *Mutual Information* [Jakulin 2005]. De fato, este algoritmo é capaz de remover variantes genéticas que não participam do contexto interativo de forma eficaz [McKinney et al. 2009].

### 2.6.8. Fontes de dados atuais

Com o aperfeiçoamento de técnicas de obtenção de dados biológicos tais como seqüenciamento de alto-desempenho ou microarranjos de alta densidade, surgiram várias fontes de dados que disponibilizam muito material nos quais esses estudos podem se basear, como *SNP500Cancer* [Packer et al. 2006], *HapMap Project* [The International HapMap Consortium 2003], *AlzGene Database* [Bertram et al. 2007] e *Wellcome Trust Case-Control Consortium (WTCCC)* [Wellcome Trust Case-Control Consortium 2010] e isso enfatizou a necessidade de métodos computacionais que ajudem os cientistas a processar essa numerosa quantidade de informação. Dados reais de estudos de associação (que é o caso da última mencionada - WTCCC) geralmente são mais sigilosos porque envolvem pacientes e isso está retratado como uma das principais dificuldades deste trabalho, porém estas bases de dados estão submetidas a procedimentos cada vez menos burocráticos de requisição de dados e espera-se que no futuro fontes de dados em estudos de associação sejam mais facilmente encontradas.

Como será detalhado posteriormente, o conjunto de dados utilizado para experimentar os métodos foi obtido pronto em [Velez et al. 2007] que geraram os dados através de funções estatísticas. Apesar de ter obtido os dados prontos, estão disponíveis diversos pacotes para simulação de dados de polimorfismos. Os mais conhecidos e utilizados são o HAP-SAMPLE [Wright et al. 2007] que utiliza dados do *HapMap* para sintetizar novos dados e o genomeSIMLA [Edwards et al. 2008] que é capaz de simular padrões realísticos em *Linkage Disequilibrium*.

## 2.7. Considerações Finais

Estudos deste gênero são um passo adiante para o estado da arte no que diz respeito a estudos de associação, chamado de *Whole-Genome Studies*. Como mencionado anteriormente, as tecnologias atuais estão gerando cada vez mais dados biológicos. Neste paradigma mais recente existem dados de até 300.000 polimorfismos sequenciados a partir de um número muito grande de indivíduos para serem analisados. Todos os métodos citados neste texto ainda não são ideais para a análise de uma quantidade tão grande de dados biológicos, nem se todos os computadores do mundo trabalhassem apenas neste problema [Moore e Ritchie 2010]. A crença é de que estes métodos sejam aperfeiçoados por técnicas computacionais e estatísticas de forma a permitir a realização de estudos do tipo *Whole-Genome*.

"But man has still another powerful resource: natural science with its strictly objective methods."

- [Ivan Pavlov](http://www.brainyquote.com/quotes/quotes/i/ivanpavlov284473.html)

3. Métodos

Neste capítulo serão descritos os métodos estudados e o método proposto. A descrição dos métodos não originais será feita com base nas suas respectivas publicações e fornecerá uma ampla visão de suas motivações, objetivos, variantes, parâmetros e utilização dos mesmos em estudos biológicos reais, revisando os resultados interessantes que eles obtiveram. O formato de entrada de todos os algoritmos, que consiste em um conjunto de dados a nível individual típico em estudos de caso-controle, será explanado no primeiro item.

Algumas considerações gerais devem ser feitas para evitar repetições desnecessárias, dado que todos os métodos seguem uma mesma lógica combinatória. Todos eles são métodos exaustivos, ou seja, exploram todas as possíveis combinações entre os atributos (características genéticas ou ambientais) e atribuem uma pontuação à cada combinação. Algumas figuras ou fórmulas a seguir mostrarão os cálculos de tais pontuações apenas para os casos onde há dois atributos que interagem, por motivo de simplificação ou fácil visualização, porém a generalização é diretamente proporcional ao aumento da dimensionalidade da tabela de genótipo-fenótipo. Todos os métodos a seguir podem ser aplicados em bases de dados de qualquer tamanho. O tamanho da população costuma ter pouco impacto no tempo computacional (como será mostrado nos próximos capítulos) porém a quantidade de atributos pode tornar a execução demasiadamente lenta, portanto deve-se evitar números muito elevados. Outro parâmetro decisivo é a escolha da ordem de interação, isto é, a quantidade de SNPs presentes em cada combinação a ser pontuada. Caso a ordem seja 4 estaremos procurando por interações quádruplas entre polimorfismos. Também é recomendado evitar análises com interações mais altas em bases com muitos atributos.

Todos os métodos possuem em comum os parâmetros «ORDER» e «TOP» que definem respectivamente a ordem de interação e a quantidade máxima de combinações que será ranqueada de acordo com as funções de pontuação.

Apenas o método PIA possui uma política de tratamento de dados faltosos, que é a mais simples possível, desconsiderando o dado faltoso para a população da tabela de genótipo-fenótipo. Na implementação realizada neste estudo adotou-se esta política a todos os métodos, sendo este tratamento realizado na própria fase de leitura da entrada de dados.

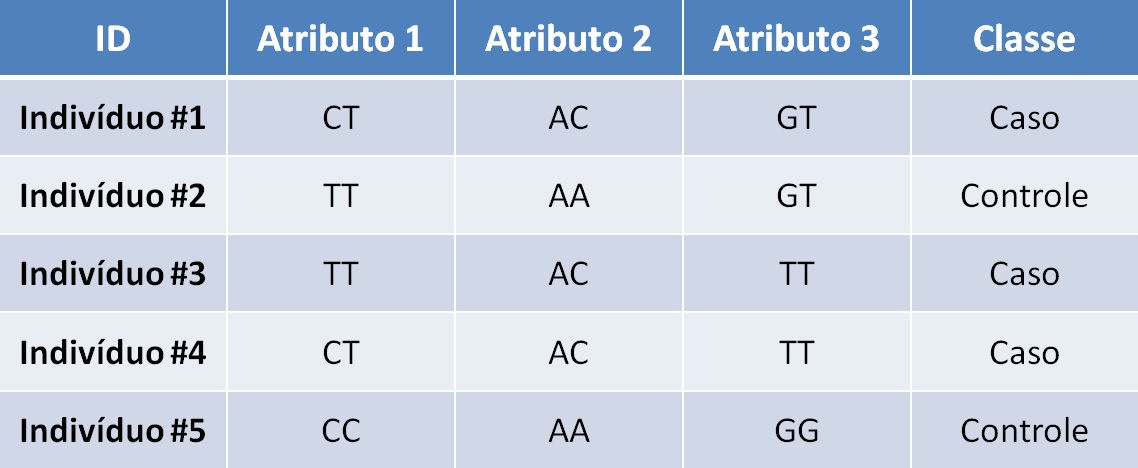
## 3.1. Formato da Entrada

O formato da base de dados utilizado como entrada para os algoritmos segue o padrão de estudos de associação no nível individual. Essa entrada consiste numa matriz onde as linhas representam os indivíduos e as colunas representam as características genéticas ou ambientais a serem analisadas. Uma coluna adicional representa a informação sobre o estado do indivíduo dada uma condição fenotípica, como por exemplo, se o indivíduo é caso (possui a patologia em questão) ou controle (não possui a doença).

Cada indivíduo possui uma informação associada a cada variável genética ou ambiental. Essas variáveis devem ser categóricas e indicam o estado desse indivíduo em relação àquele atributo. No caso de um polimorfismo esse atributo, como discutido no capítulo anterior, assume três possíveis valores: *homozigoto dominante*, *homozigoto recessivo* e *heterozigoto*. Como os seres humanos são diplóides, isto é, possuem duas cópias de cada cromossomo, podem acontecer os seguintes casos: os dois cromossomos contêm o mesmo alelo que é o mais presente na população (*homozigoto dominante*), ambos contêm o mesmo alelo que é o mais raro na população (*homozigoto recessivo*) e um cromossomo possui o alelo mais comum enquanto outro possui o mais raro (*heterozigoto*).

A Figura 3.1 mostra um exemplo fictício desta entrada para os algoritmos. Nela estão retratados os pacientes com seus respectivos identificadores e os atributos que podem ser informações genéticas (SNPs) ou ambientais (por exemplo, raça e tempo de exposição ao cigarro). A última coluna contém a classe do indivíduo, isto é, a informação de caso-controle.

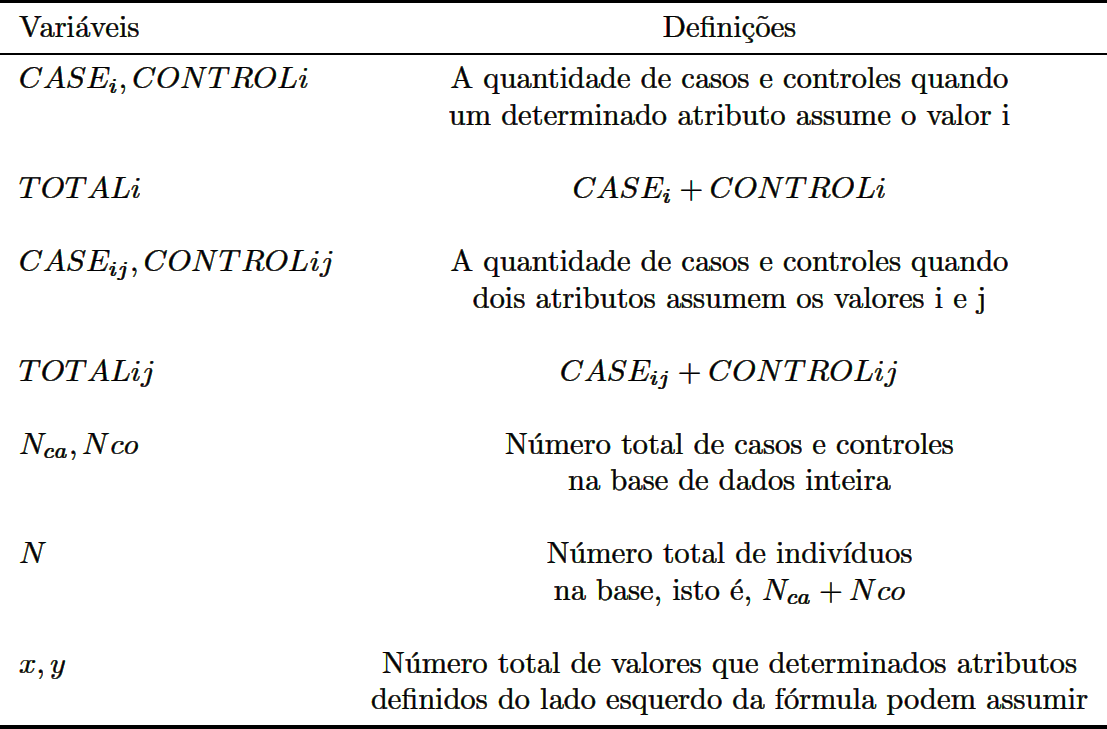
É importante ressaltar que quando se trabalha com interações entre atributos, os dados devem, obrigatoriamente, serem considerados neste formato a nível individual, isto é, contendo a informação de cada indivíduo para cada característica. Outros estudos como os de procura de marcadores genéticos únicos podem utilizar dados a nível coletivo (ou resumido), que consistem nos mesmos tipos de dados, porém contendo apenas a quantidade total de indivíduos portadores da patologia e não portadores para cada característica.



**Figura 3.1.** Formato da Base de Dados

Fonte: Criação própria

O Quadro 3.1 mostra todas as variáveis associadas ao conjunto de dados que serão posteriormente utilizadas nas fórmulas que avaliam a importância de cada combinação pelos métodos descritos nos próximos tópicos.

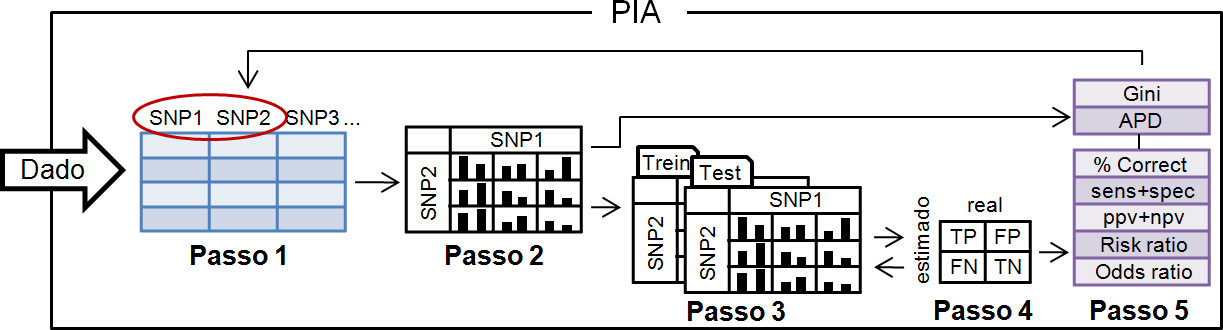


**Quadro 3.1.** Variáveis da Base de Dados

Fonte: Criação própria

## 3.2. Polymorphism Interaction Analysis ̶ PIA

PIA v.2.0 é um método de seleção de atributos com base na idéia de que não existe um único método que identifique de forma ótima as relações complexas entre atributos em todos os casos existentes [Mechanic et al. 2008]. Ele possui sete métricas diferentes que avaliam todas as possíveis combinações de SNPs e as classificam de acordo com uma pontuação geral calculada com base em todas as métricas [Goodman et al. 2006]. A Figura 3.2 resume todos os passos descritos nos próximos itens.



**Figura 3.2.** Polymorphism Interaction Analysis

Fonte: Criação própria baseada no método descrito em [Mechanic et al. 2008]

Este método foi utilizado com sucesso em estudos sobre câncer de cólon. Ele foi capaz de identificar polimorfismos nulos no gene relativo à *glutationa-S transferase T1 (GSTT1)* como um dos melhores indicadores do risco de desenvolvimento do câncer de cólon, sendo estes já relatados na literatura. Outros resultados também foram obtidos e confirmados por regressão logística como a relação entre a modificação da associação entre o câncer e polimorfismos no *GSTT1* com polimorfismos no gene *TP53* (*Arg72Pro* na *p53*) e *CASP8* (*Asp302His* na *caspase 8*) [Goodman et al. 2006].

### 3.2.1. Seleção da próxima combinação de atributos a ser analisada

O algoritmo é iniciado com a seleção da próxima combinação a ser analisada (Passo 1). Como explicitado no início deste capítulo todos os métodos possuem essa etapa pois todos eles avaliam todas as possíveis combinações de SNPs. No caso do PIA, esta seleção não possui nenhum critério ou heurística em especial, apenas uma restrição de integridade onde uma combinação é descartada da análise quando a fração entre o maior e o menor número entre casos e controles excede um certo limite, estabelecido pelo usuário na forma do parâmetro «RATSL».

### 3.2.2. Criação da tabela de genótipo-fenótipo

Escolhida a próxima combinação a ser analisada, o algoritmo procede com a criação da tabela de genótipo-fenótipo. A tabela de genótipo-fenótipo, conforme exemplificada de forma bidimensional no Passo 2 da Figura 3.2, é uma tabela k-dimensional, onde k é a ordem de interação em questão. Cada dimensão está associada a um atributo da combinação sob análise e pode assumir todos os valores deste atributo. Cada célula desta tabela irá conter o número total de casos e de controles (as barras pretas na imagem) que contêm tal combinação específica de valores nos atributos em cada dimensão. Esta etapa é fundamental em todos os métodos.

Para o cálculo das métricas baseadas na tabela de contingência, o algoritmo procede para o item 3.2.3, porém para o cálculo das métricas baseadas em separação, o algoritmo pula para o passo 3.2.5 (as duas setas saindo do Passo 2).

### 3.2.3. Validação cruzada

Nesta etapa é realizado o procedimento de validação cruzada ou *n-fold cross-validation*, com o usuário definindo n, para que as métricas baseadas na tabela de contingência possam ser calculadas. Esta estratégia divide o conjunto de dados em subconjuntos disjuntos de treino e teste (Passo 3). O valor n, representado pelo parâmetro «FRACT», dita o tamanho do conjunto de teste, geralmente 10% do tamanho total.

Para cada divisão intervalar, a tabela de contingência é populada conforme descrito no passo 3.2.4. Cada etapa representa um incremento à tabela de contingência existente. Este procedimento de validação cruzada pode ser repetido diversas vezes, definido no parâmetro «NTIME», acrescentando-se novamente os dados à mesma tabela de contingência, sendo maior a precisão das métricas baseadas nesta tabela à medida que a quantidade de repetições aumenta. No início de cada procedimento os dados são aleatoriamente embaralhados numa quantidade de vezes regulada pelo parâmetro «SHUFFLE».

### 3.2.4. População da tabela de contingência

Este é o passo crucial para as métricas baseadas nesta tabela e que permite o maior número de variantes do algoritmo. Será explicada a forma mais simples pela qual a tabela de contingência pode ser populada e depois serão detalhadas as variantes. Uma tabela de contingência padrão é exibida no Passo 4 da Figura 3.2, nesta estrutura as linhas representam os valores que foram estimados como pertencentes às classes envolvidas no problema e as colunas representam os valores reais.

Cada célula da tabela de genótipo-fenótipo é classificada como célula de caso ou de controle baseando-se na quantidade destes indivíduos no conjunto de treinamento na célula. Caso a quantidade de casos seja maior do que a de controles a célula é rotulada como sendo 'célula de caso', caso contrário ela é rotulada como 'célula de controle'. Esta quantidade de casos ou de controles pode ser absoluta ou relativa, conforme definido no parâmetro «IFRACT». No caso da utilização de números relativos de casos ou controles, isto é, o número absoluto em cada célula dividido pelo número total em todo o conjunto de dados, diz-se estar utilizando uma fração ocupacional. Caso alguma célula não possua nenhum indivíduo ela é não é rotulada e diz-se que está 'vazia'. Quando o número de casos e controles em uma célula são iguais ela é considerada metade 'célula de caso' e metade 'célula de controle'.

Após a classificação das células da tabela de genótipo-fenótipo, todos os indivíduos do conjunto de teste são verificados. Caso eles sejam casos e estejam numa célula caso é incrementada a quantidade de verdadeiros positivos (TP), caso sejam controles e estejam numa célula de controle é incrementada a quantidade de verdadeiros negativos (TN), caso sejam casos e estejam numa célula de controle ou a célula seja vazia incrementa-se a quantidade de falsos negativos (FN) e caso sejam controles e estejam numa célula de caso ou vazia, incrementa-se a quantidade de falsos positivos (FP). Quando a célula é metade caso e metade controle são incrementados os respectivos contadores falsos e verdadeiros em 0.5 para cada indivíduo.

Neste ponto existem algumas variações para este algoritmo básico. A primeira delas é a capacidade de se utilizar o conjunto de treino para popular também a tabela de contingência, regida pelo parâmetro «ITRAIN», sendo o modo de utilização idêntico ao uso do conjunto de teste. A segunda variação se dá com o parâmetro «LOOTR» que define a forma como os valores da tabela de contingência são incrementados com o conjunto de teste. Três maneiras podem ser executadas: No procedimento de *Leave One Out* o indivíduo que está sendo testado não é levado em consideração durante a fase de rotulação das células da tabela de genótipo-fenótipo; No procedimento *Maximum Rules* o indivíduo em teste é considerado durante a fase de rotulação das células da tabela de genótipo-fenótipo; E na opção *PIA v.1.0* é executado o algoritmo conforme sua primeira versão descrita em [Goodman et al. 2006], onde duas modificações acontecem: a primeira ocorre quando o número de casos e controles é igual, sendo todos os indivíduos testados nesta célula classificados como corretos, e a segunda é o fato de que quando existe apenas um controle na célula, ela é incorretamente classificada sempre.

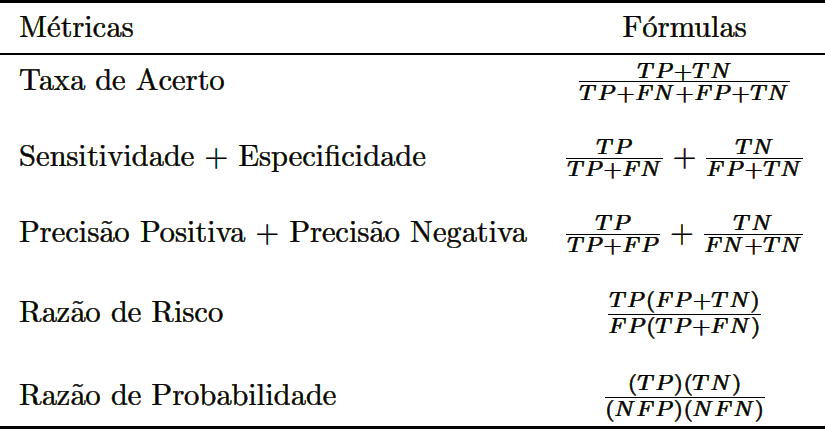
No final deste procedimento, a tabela de contingência conterá o somatório de todos os valores de TP, FP, TN, FN para todos os intervalos da validação cruzada e todas as rodadas deste procedimento. Estes valores são utilizados para o cálculo das métricas baseadas na tabela de contingência conforme será mostrado no item 3.2.5. É importante ressaltar o fato de que as métricas de separação não precisam deste procedimento retratado nos itens 3.2.3 e 3.2.4 pois utilizam a tabela de genótipo-fenótipo inteira.

### 3.2.5. Cálculo da pontuação da combinação

O PIA v.2.0 possui sete métricas, sendo cinco delas métricas baseadas na tabela de contingência e duas delas métricas de separação. As primeiras geralmente são utilizadas em estudos que contém viés estatístico, para correção, visualização, estreitamento de intervalos, cálculo de risco ou margens de erros, etc. Neste contexto elas tendem a retratar uma generalização de cada viés individual de cada célula da tabela de genótipo-fenótipo entre as possíveis classes caso ou controle. As últimas duas são comumente utilizadas como critério de escolha de nós em algoritmos de construção de árvores de decisão, árvores de classificação e regressão (CART, do inglês *Classification and Regression Trees*) e florestas randômicas compostas por estes tipos de árvores. Em todas as métricas, quanto maior o valor, mais interessante é o resultado.

As métricas baseadas na tabela de contingência utilizada neste método (tabela roxa de baixo no Passo 5 da Figura 3.2), mostradas no Quadro 3.2, com suas abreviações dos termos em inglês entre parênteses são: taxa de acerto (*%Correct*), sensitividade + especificidade (*Sensitivity + Specificity*), precisão positiva + precisão negativa (*Positive Predictive Value + Negative Predictive Value*), Razão de Risco (*Risk Ratio*), Razão de Probabilidade (*Odds Ratio*).

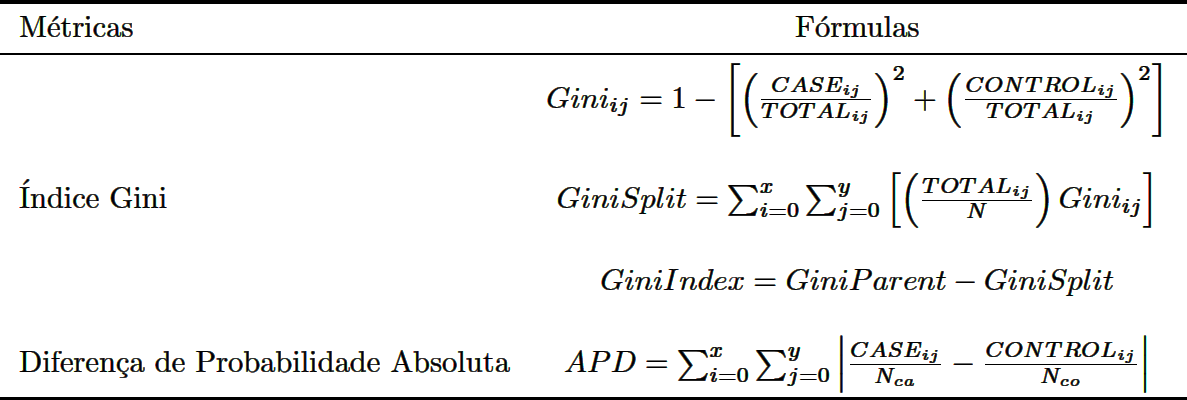
As métricas baseadas em separação (tabela roxa de cima no Passo 5 da Figura 3.2), que utilizam o conjunto de dados inteiro, são definidas no Quadro 3.3. São elas o índice Gini (*Gini Index*) e a diferença da probabilidade absoluta (*Absolute Probability Difference*).



**Quadro 3.2.** Métricas Baseadas na Tabela de Contingência

Fonte: Criação própria baseada no método descrito em [Mechanic et al. 2008]

O cálculo de todas as métricas descritas nesta tabela é de forma direta através do uso dos valores da tabela de contingência. Observe que os valores da taxa de acerto e da sensitividade + especificidade serão os mesmos caso as bases de dados sejam balanceadas ou caso estejam sendo utilizadas frações ocupacionais.



**Quadro 3.3.** Métricas Baseadas em Separação

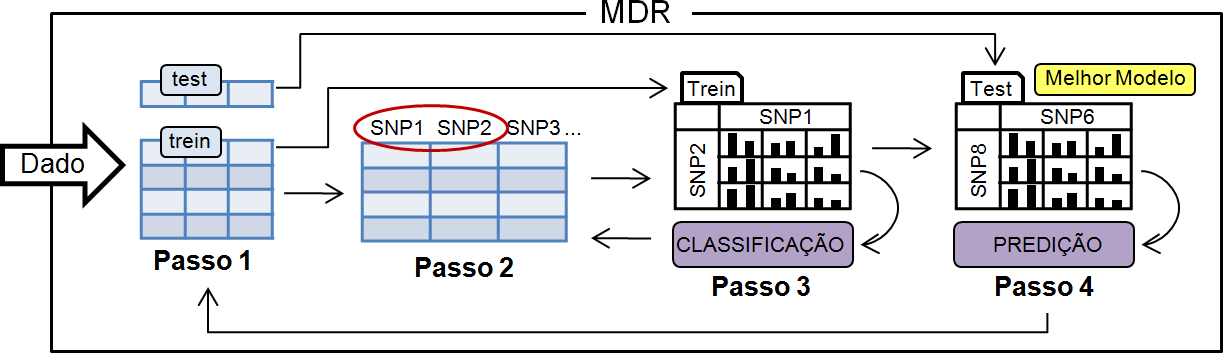
Fonte: Criação própria baseada no método descrito em [Mechanic et al. 2008]

O índice Gini consiste de três partes. A primeira é o cálculo do índice parcial para cada célula da tabela de contingência (assumindo uma tabela com duas dimensões neste quadro). O cálculo completo do índice é realizado então com base na média ponderada entre todos os índices parciais. A métrica GiniParent se refere ao cálculo do GiniSplit para a base de dados inteira tomando-se o total de casos e controles (dessa forma levando em consideração o quão desbalanceada a base está). Por fim o índice final é calculado através da subtração do GiniParent do GiniSplit. O cálculo do índice Gini para variáveis de dimensões diferentes de dois é realizado através do simples cálculo do índice Gini parcial para todas as células desta da nova tabela de genótipo-fenótipo. Essa mesma observação procede para a diferença da probabilidade absoluta, que consiste na aplicação direta da fórmula descrita nesta tabela.

Após o cálculo de todas essas sete métricas para a combinação em questão, é calculada uma pontuação final que corresponde à importância da combinação e que servirá para ranqueá-la. Esta pontuação final (*Overall*) é obtida através do somatório de todas as outras métricas normalizadas de forma que o maior valor seja 50. Como são sete métricas o maior valor possível é 350. Caso o número de casos e controles seja igual, isto é, a base é balanceada, ou então caso estejamos utilizando frações ocupacionais, o valor da sensitividade + especificidade é desconsiderado pois sempre será igual à taxa de acerto. Neste caso o maior valor que a pontuação final pode assumir é 300.

## 3.3. Multifactor Dimensionality Reduction ̶ MDR

A principal idéia por trás do MDR é selecionar as características genéticas em grupos de alto ou baixo risco, reduzindo a dimensionalidade dessas variáveis que predizem o risco para apenas uma dimensão. A partir disso, essa nova variável unidimensional é submetida a um processo de validação cruzada e a um teste de permutação para avaliar sua capacidade de classificar e predizer a doença [Hahn et al. 2003; Ritchie et al. 2001]. A Figura 3.3 resume todos os passos descritos nos próximos itens.



**Figura 3.3.** Multifactor Dimensionality Reduction

Fonte: Criação própria baseada no método descrito em [Hahn et al. 2003]

O método MDR já foi aplicado em diversos estudos. Em um deles, a respeito de câncer de mama esporádico, MDR foi aplicado em 10 polimorfismos presentes nos genes *COMT*, *CYP1A1*, *CYP1B1*, *GSTM1* e *GSTT1*, encontrando uma associação entre o risco da doença e uma interação quádrupla entre alguns polimorfismos [Ritchie et al. 2001].

### 3.3.1. Validação Cruzada

O MDR possui esquema um pouco diferente dos demais pois divide o conjunto de dados em treino e teste antes das demais etapas (Passo 1). Esta divisão é feita da mesma forma que o definido no item 3.2.3 da seção anterior. Essa construção se dá porque uma das características do MDR é a escolha do melhor modelo a cada procedimento completo de validação cruzada. Os parâmetros que definem quantos procedimentos serão realizados e quantas vezes os dados serão aleatoriamente misturados são, respectivamente, o «NTIME» e o «SHUFFLE». O tamanho do conjunto de teste é regido pelo parâmetro «FRACT».

### 3.3.2. Seleção da próxima combinação de atributos a ser analisada

Após a divisão do conjunto em teste e treino, a próxima combinação de SNPs possível é então definida (Passo 2). Este método inverte a ordem natural dos outros algoritmos fazendo a iteração entre todas as combinações possíveis dentro do *loop* da validação cruzada.

São criadas as tabelas de genótipo-fenótipo do conjunto de teste e do conjunto de treino para a combinação de atributos em questão. A tabela referente ao conjunto de treino é utilizada na fase de classificação enquanto a de teste é mantida para o uso na última etapa de predição.

### 3.3.3. Classificação

Com os dados da tabela de treino, é realizado um processo semelhante ao do método PIA em que cada célula desta tabela de genótipo-fenótipo é rotulada (Passo 3). Neste método cada célula pode ser rotulada em 'alto risco', 'baixo risco' e 'vazia', sendo feita da seguinte forma: se a fração entre casos e controles for maior do que um certo limiar a célula é considerada 'alto risco', caso esta fração seja menor ou igual a esse limiar ela é considerada 'baixo-risco' e caso a célula não contenha nenhum dado então é considerada 'vazia'. Caso o número de controles seja igual a 0 porém o número de casos seja maior que 0 a razão é considerada infinita e a célula é rotulada 'alto risco'. Este limiar é definido pelo parâmetro «THRESHOLD».

De posse desses rótulos é realizado um processo de classificação com o próprio conjunto de treino no formato *Leave-One-Out* para se ter acesso às taxas de acerto da combinação em questão. Neste procedimento, cada indivíduo do conjunto é testado sobre esses rótulos e classificado de acordo com algum classificador. Qualquer algoritmo de classificação em aprendizagem de máquina é valido, como por exemplo Naive Bayes, Suport Vector Machines, classificadores baseados em instância (KNN), Árvores de Decisão, Florestas Randômicas, Redes Neurais Artificiais ou qualquer outro. Nesta implementação em específico foi utilizado o chamado "Core MDR", que usa a simples função de acurácia balanceada (do inglês, *Balanced Accuracy*), que consiste na simples média entre a sensitividade e a especificidade, tornando a pontuação deste algoritmo levemente parecida com a sensitividade + especificidade do PIA.

Esta iteração entre os itens 3.3.2 e 3.3.3 é realizada até que todas as possíveis combinações entre os atributos, dada a ordem de interação escolhida, sejam classificadas. A combinação que possuir a maior taxa de acerto é escolhida como sendo o melhor modelo temporário de rotulação da tabela de genótipo-fenótipo. O modelo temporário que obtiver a melhor pontuação entre todos esses modelos calculados para cada intervalo da validação cruzada (iteração dos itens 3.3.1 até 3.3.3) procede para o Passo 4.

De fato, cada iteração completa do procedimento de validação cruzada gera apenas um melhor modelo. O ranqueamento feito nesta implementação leva em consideração todas as taxas de acerto em todas as etapas de todos os intervalos da validação cruzada. Caso uma taxa de acerto para uma combinação específica seja maior do que outra taxa de acerto prévia para esta mesma combinação ela é simplesmente atualizada e realocada no ranking. Caso a taxa de acerto seja menor do que outra previamente calculada, ela é descartada. Desta forma cada combinação guarda sua métrica correspondente à sua melhor rodada em todas as etapas da validação cruzada corrente.

### 3.3.4. Predição

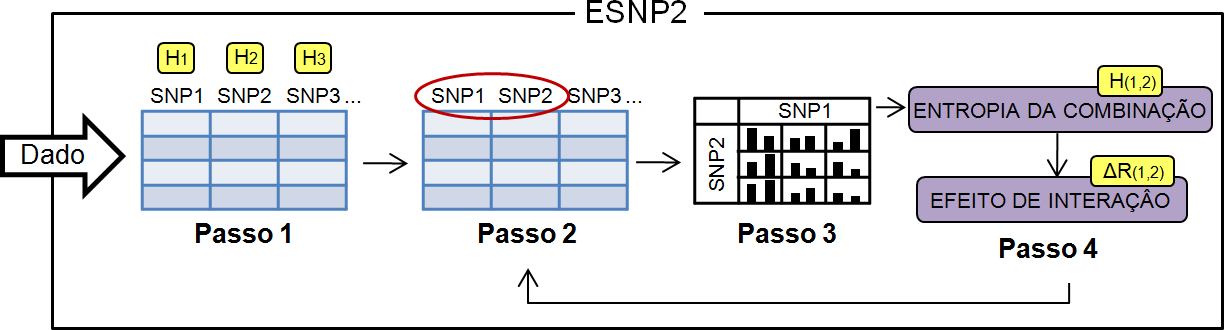
De posse do melhor modelo, o método implementado neste trabalho também retorna uma lista contendo a taxa de acerto da fase de predição. Esta lista possui o tamanho igual à quantidade de vezes que se realizou o procedimento completo da validação cruzada (parâmetro «NTIME»).

A predição é feita com o mesmo algoritmo da fase de classificação, porém desta vez utilizando o conjunto de teste sobre o de treino ao invés do procedimento de *Leave One Out* (Passo 4). Além da lista contendo a as taxas de acerto da predição, esta implementação calcula a média entre as taxas de acerto das fases de classificação e de predição.

A lista contendo os melhores modelos pode ser utilizada como uma outra métrica para avaliar a significância das combinações mais promissoras. Por exemplo, caso realizemos dez vezes o procedimento de validação cruzada e em todos eles ou em grande maioria uma certa combinação aparece como sendo o melhor modelo, então é bastante provável que a sua seleção pelo algoritmo de classificação não foi por simples chance. Além disso, as médias entre as taxas de acerto dessas últimas fases podem ser utilizadas como indicadores de uma ordem de interações ótima. Para isso realizamos todo o procedimento aumentando a ordem de interação. Quando maior for a média entre as taxas de acerto de predição em relação à média entre as taxas de acerto de classificação (naturalmente mais altas), poderemos afirmar com mais certeza que o algoritmo de classificação não está sendo prejudicado com o aprendizado demasiado devido a altas ou baixas ordens de interação, isto é, não está sobreajustando (*overfitting*) os dados [Hahn et al. 2003].

## 3.4. Entropy-Based SNP-SNP Interaction Method ̶ ESNP2

Existem duas versões deste método descritas em [Dong et al. 2007]. O *ESNP2-standard* (ESNP2-S) tem como objetivo detectar as interações da mesma forma que os métodos descritos nos outros tópicos o fazem, e o *ESNP2-model-option* (ESNP2-Mx) testa a interação contra vários modelos genéticos pré-concebidos. ESNP2-Mx não foi desenvolvido neste trabalho porque os resultados são bastante próximos do ESNP2-S e o objetivo é construir uma ferramenta que analise todas as ordens de interação, sendo muito difícil aplicar todos os modelos genéticos prévios, dadas ordens mais altas [Li e Reich 1999]. A definição original do ESNP2-S avalia apenas o caso das interações de ordem dois, porém ela pode ser facilmente generalizada para avaliar todas as ordens de interação possíveis (como descrito neste estudo). A Figura 3.4 resume todos os passos descritos nos próximos subtópicos.



**Figura 3.4.** Entropy-based SNP-SNP Interaction Method

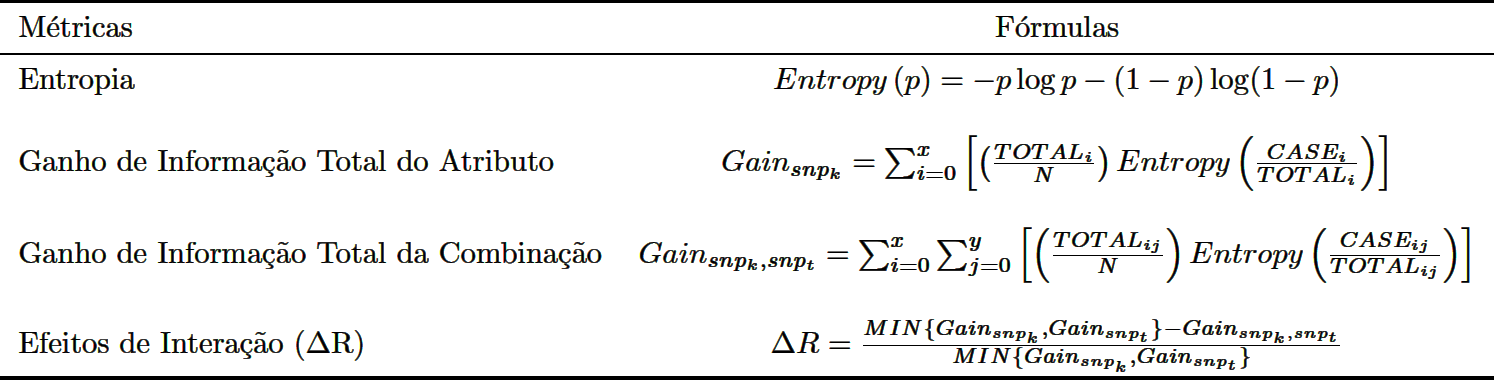
Fonte: Criação própria baseada no método descrito em [Dong et al. 2007]

O método ESNP2 foi aplicado em dados de malária, resultando na comprovação bem sucedida de um efeito epistático negativo entre anemia falciforme e talassemia alfa com a malária. Enquanto uma regressão logística com teste de Wald só pôde dizer que existia interação entre a anemia falciforme com a talassemia alfa e que isso afetava a resistência da malária, o ESNP2, além de fornecer esses resultados, conseguiu mostrar o modelo epistático negativo apropriado para a relação [Dong et al. 2007].

### 3.4.1. Cálculo da entropia individual

O algoritmo é iniciado com o cálculo da entropia para cada atributo individualmente (Passo 1). A entropia é uma medida, assim como o índice Gini e o APD do método PIA, muito utilizada como forma de seleção de nós em algoritmos de construção de árvores de decisão. A primeira fórmula do Quadro 3.4 mostra o cálculo da entropia pura para cada possível valor de um atributo enquanto a segunda fórmula mostra o cálculo da entropia total para cada atributo.

Essas entropias, que também são chamadas de Ganho de Informação Associado, são mantidas para uso posterior no cálculo dos efeitos de interação finais da combinação.



**Quadro 3.4.** Métricas do ESNP2

Fonte: Criação própria baseada no método descrito em [Dong et al. 2007]

### 3.4.2. Seleção da próxima combinação de atributos a ser analisada

O algoritmo procede selecionando a próxima combinação da lista de todas as possíveis combinações entre os atributos do conjunto de dados de entrada (Passo 2). Da mesma forma que o MDR, este método não realiza nenhuma heurística ou seleção de combinações específica, a seleção é feita em ordem até que todas as combinações tenham sido avaliadas.

### 3.4.3. Criação da tabela de genótipo-fenótipo

A tabela de genótipo-fenótipo então é criada a partir da combinação escolhida (Passo 3). Como este método não necessita de estratégias do gênero da validação cruzada realizada no PIA e no MDR essa tabela será única e servirá exclusivamente para o cálculo do grau de importância da combinação para posterior ranqueamento (tratando-se do ESNP2-S).

No ESNP2-Mx existiria uma fase posterior a essa onde essa tabela seria sobreposta a diversos modelos genéticos preexistentes tais como o modelo recessivo-recessivo (RR), modelo recessivo de único gene (1L:R), modelo de modificação de efeito (MOD), etc. Essa análise se torna muito complexa conforme a ordem de interação cresce e a publicação original onde este método foi concebido [Dong et al. 2007] não realiza interações de ordens maiores que dois atributos. Por este motivo e os explicitados no início deste tópico apenas o ESNP2-S foi implementado.

### 3.4.4. Cálculo dos efeitos de interação da combinação

De posse da tabela de genótipo-fenótipo, são calculados os efeitos de interação com base novamente na entropia associada a cada combinação entre atributos conforme a terceira fórmula do Quadro 3.4. A métrica final de interação (ΔR) é calculada através da razão mostrada na última fórmula do Quadro 3.4.

O ESNP2 é um método bastante simples e possui apenas um parâmetro intitulado «BOOTSTRAP» que permite a análise dos *p-values* para cada combinação em relação a um teste de *bootstrapping*. Este teste consiste na aplicação do método ESNP2 uma grande quantidade de vezes, definida no parâmetro mencionado, porém com um conjunto de dados obtido a partir de uma amostra com reposição dos indivíduos contidos nos dados originais. Este novo conjunto de dados deverá ter a mesma quantidade de indivíduos do conjunto de dados original porém como a amostragem foi feita com reposição, cerca de um terço [Breiman 2001] dos indivíduos do conjunto original ficarão de fora desta nova base. O *p-value* de cada combinação é dado pela proporção de combinações cujo ΔR foi maior do que na aplicação do método no conjunto original. Este tipo de teste pode ser utilizado em qualquer outro método descrito nesta seção, porém o escopo foi restringido à sua não implementação, constando a mesma como melhoria futura.

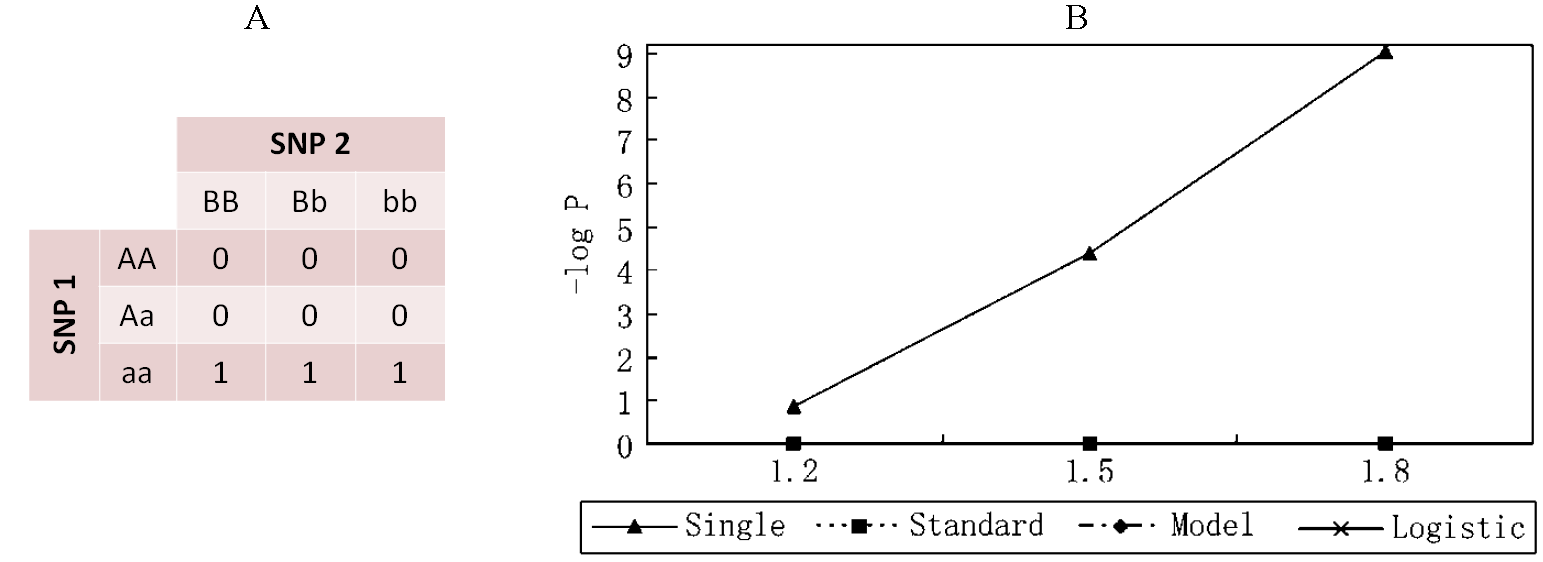
## 3.5. Multi-Approach SNP-SNP Interaction Analysis ̶ MASS

Este método foi desenvolvido a partir da observação dos pontos fortes e fracos dos métodos discutidos nas seções anteriores. Basicamente ele consiste na junção das seguintes idéias que funcionaram respectivamente nos métodos PIA e ESNP2:

* Nenhum método isolado é capaz de prever de forma ótima as interações entre os atributos. Por isso a melhor abordagem é unir várias métricas simples em uma métrica geral. Com o uso de diversas funções, a probabilidade de viés é menor e combinações interessantes geralmente são enfatizadas.
* Métricas que utilizam o conjunto inteiro de dados ao invés de realizar um procedimento no estilo de validação cruzada se sobrepõem tanto na procura pelas melhores combinações quanto no tempo computacional. Estratégias de validação cruzada ou amostragem com reposição tomam muito tempo computacional e suas vantagens são percebidas apenas quando as bases de dados estão drasticamente desbalanceadas, e ainda assim, o desempenho de métricas baseadas em separação é competitiva.
* Considerar os efeitos de interação de cada atributo pertencente à combinação em questão se mostrou uma abordagem bastante interessante, sendo capaz de capturar de forma eficaz as interações na presença tanto de efeitos marginais quanto de efeitos principais.

### 3.5.1. Motivação

No estudo realizado por [Dong et al. 2007] foi verificado que o ESNP2 obteve baixo desempenho em modelos genéticos tais como o modelo recessivo de único gene 1L:R (Figura 3.5-A). Neste modelo o efeito de um único gene consegue explicar boa parte da condição da doença. Porém na última fórmula do Quadro 3.4 que define o efeito interativo do ESNP2 é observado que apenas o menor entre os valores das entropias individuais é levado em consideração. Quanto menor o valor obtido através da aplicação da segunda fórmula do Quadro 3.4 em um atributo singular, maior a capacidade do atributo, individualmente, dividir de forma correta a base de dados. Isto faz com que, caso um atributo individualmente consiga dividir de forma satisfatória o conjunto de dados, em contraposição ao outro que divide pessimamente, a entropia deste atributo ótimo será pequena o suficiente para não acrescentar informação relevante na última fórmula do Quadro 3.4. Portanto os efeitos de interação aleatórios podem e vão, porventura, se sobrepor aos efeitos de interação reais, mesmo que os efeitos da interação se sobreponham de forma razoável ao efeito da entropia de um *locus* apenas, o que geralmente acontece.



**Figura 3.5.** Motivação para o Algoritmo MASS

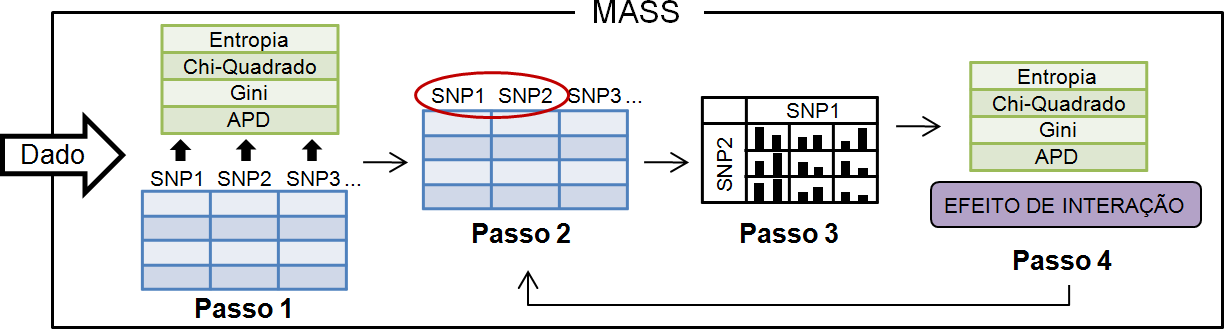
Fonte: A) Criação própria com base nos modelos descritos em [Li e reich 1999]

B) Figura 3 [Dong et al. 2007]

A imagem mostra o modelo 1L:R (parte A) e os resultados obtidos através da aplicação dos algoritmos ESNP2-S e ESNP2-Mx ao lado de uma análise de regressão logística neste modelo em [Dong et al. 2007] (parte B). O gráfico mostra os *p-values* obtidos em um teste de *bootstrap* conforme descrito no item 3.4.4 para cada modelo de dados com um OR (*Odds Ratio*) específico (utilizado neste estudo como parâmetro básico para definir a relação entre o genótipo com a doença) e uma população de 1000 casos e 1000 controles. O triângulo representa o valor mínimo do p-valor da associação entre dois genes, isto é, o máximo de efeitos devidos a um único gene.

Dado este fato observado, uma simples modificação na última fórmula do Quadro 3.4 resultando na última fórmula do Quadro 3.5 consegue fazer com que algumas interações reais onde um dos genes participantes contribuía de forma maior aos efeitos totais, que se perdiam em meio a interações aleatórias, fossem encontradas. A única preocupação desta mudança seria na perda da sensibilidade devido ao fato de que as reações interativas mais interessantes aos biólogos são justamente as interações onde os dois genes participam ativamente, porém percebeu-se de forma empírica que isto não aconteceu (ver seção Resultados).

Este novo método consegue taxas de acerto de detecção de interações, na maioria das vezes igual ou maiores do que o algoritmo que conseguiu as melhores taxas entre os estudados (ESNP2). O tempo de execução é levemente maior do que o ESNP2, porém ainda assim consegue ficar bem abaixo dos outros dois métodos (mais detalhes na seção Resultados).

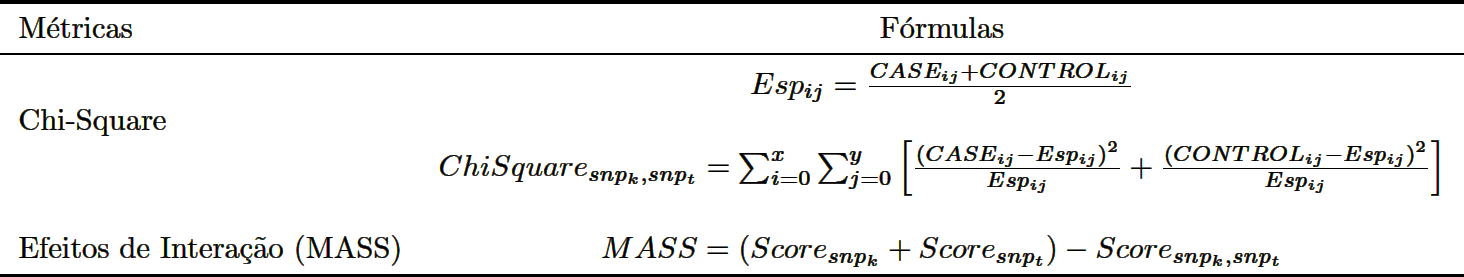
****

**Figura 3.6.** Multi-Approach SNP-SNP Interaction Analysis

Fonte: Criação própria

### 3.5.2. Cálculo das métricas individualmente

Assim como no método ESNP2, este algoritmo inicia com o cálculo das métricas individuais para cada atributo (Passo 1 da Figura 3.6). As métricas utilizadas por esse algoritmo são as seguintes: ganho de informação (utilizando a entropia de forma semelhante ao método ESNP2), índice Gini e APD (também de forma idêntica ao algoritmo PIA) e a métrica de Chi-Quadrado (do inglês, *Chi-Square*).



**Quadro 3.5.** Métricas Adicionais do MASS

Fonte: Criação própria

Este quadro exibe apenas as métricas novas para o método MASS (primeira e segunda fórmulas) e os pontos onde essa nova abordagem difere das anteriores (última fórmula). O MASS também utiliza as outras métricas: Ganho de Informação (entropia), índice Gini e APD da forma como são descritas nas tabelas 3.3 e 3.4. Na última fórmula o termo *Score* define qualquer uma das quatro métricas do algoritmo MASS.

A métrica de ganho de informação difere da terceira fórmula do Quadro 3.4 apenas pois ela foi invertida para que o resultado retornado seja melhor quanto maior for o valor. Para isso, um procedimento idêntico ao realizado no índice Gini é feito conforme exibido no Quadro 3.3.

A única métrica adicional neste método é a de Chi-Quadrado, que segue uma lógica semelhante às outras. Nela são subtraídos os efeitos de interação esperados (definidos pela primeira fórmula do Quadro 3.5) que consistem na média simples entre os casos e controles daquela célula dos efeitos de interação realmente observados tanto para casos quanto para controles. O resultado é a segunda fórmula do Quadro 3.5.

### 3.5.3. Seleção da próxima combinação de atributos a ser analisada

O algoritmo procede selecionando a próxima combinação a ser analisada dentro do conjunto de possibilidades formado por todas as possíveis combinações (Passo 2). Ele realiza uma verificação idêntica ao método PIA onde todas as combinações que são desbalanceadas demais são descartadas (ver subtópico 3.1.1). O parâmetro que indica o limite de desbalanceamento é o «RATSL».

### 3.5.4. Criação da tabela de genótipo-fenótipo

A tabela de genótipo-fenótipo é criada de forma idêntica aos outros métodos (Passo 3). Como no ESNP2 ela é única, porque este método também utiliza apenas métricas que utilizam todo o conjunto de dados.

### 3.5.5. Cálculo dos efeitos de interação intermediários (opcional)

Esta fase opcional é caracterizada pelo cálculo das funções de pontuação sobre todos os elementos do conjunto das partes da combinação em questão, excluindo o conjunto vazio, o próprio conjunto contendo todos os elementos da combinação e os conjuntos unitários (cujo cálculo das métricas foi feito na etapa 3.4.2).

Este procedimento pode ser realizado no caso em que se deseja levar em consideração todos os efeitos marginais possíveis. O procedimento é bastante custoso e não fornece vantagens interessantes em bases de dados simples, porém pode ser realizado como um estudo de caso adicional. O parâmetro que sinaliza a realização desta fase é o «SIMPLEASSIGN».

### 3.5.6. Cálculo dos efeitos de interação da combinação

Finalmente, todas as métrica são aplicadas à combinação completa (Passo 4). A última fórmula do Quadro 3.5 mostra o cálculo dos efeitos de interação para cada métrica possível. Uma pontuação final é dada a cada combinação com base na mesma idéia praticada no algoritmo PIA. Neste caso, todas as métricas são normalizadas de forma que o maior valor seja 100 e o menor seja 0, e então são somados os valores de todas as funções para a combinação em questão. Como o total de funções de pontuação é quatro, o maior valor que esta métrica final (*Overall*) pode assumir é 400.

O usuário também pode definir através dos parâmetros «USEGAIN», «USECHI», «USEGINI» e «USEAPD» a não utilização de algumas dessas funções durante o cálculo da métrica final.

## 3.6. Considerações Finais

Estes métodos estudados e implementados pertencem ao grupo de métodos exaustivos, isto é, avaliam todas as possibilidades de interações possíveis. Os resultados gerados neste trabalho podem ser utilizados em conjunto com resultados provenientes de estudos que utilizam outras abordagens para facilitar a seleção de métodos interessantes em contextos biomédicos reais.

# “Experience is the name every one gives to their mistakes.”

- Oscar Wilde

4. Experimentos

Nesta seção será detalhado o procedimento realizado para analisar comparativamente os métodos estudados. Este paradigma experimental foi utilizado previamente por [Velez et al. 2007; Mechanic et al. 2008], adjetivado pelos mesmos como uma forma robusta e estatisticamente relevante de se medir o desempenho de métodos capazes de identificar interações entre atributos.

Primeiramente será detalhada a forma como o conjunto de dados utilizado para testar os métodos foi criado. Os tópicos subsequentes consistem na explicitação do procedimento experimental e suas pretensões neste estudo.

## 4.1. Conjunto de Dados

O conjunto de dados usado para testar os métodos foi utilizado previamente por [Velez et al. 2007]: um estudo sobre a variação de parâmetros no método MDR relacionada à detecção de interações epistáticas.

Cada conjunto de dados corresponde a uma matriz definida no tópico 3.1 do capítulo anterior. Esses dados foram gerados a partir de diversos modelos epistáticos criados através de funções estatísticas que modelam parâmetros biológicos reais. Os parâmetros modelados são os seguintes:

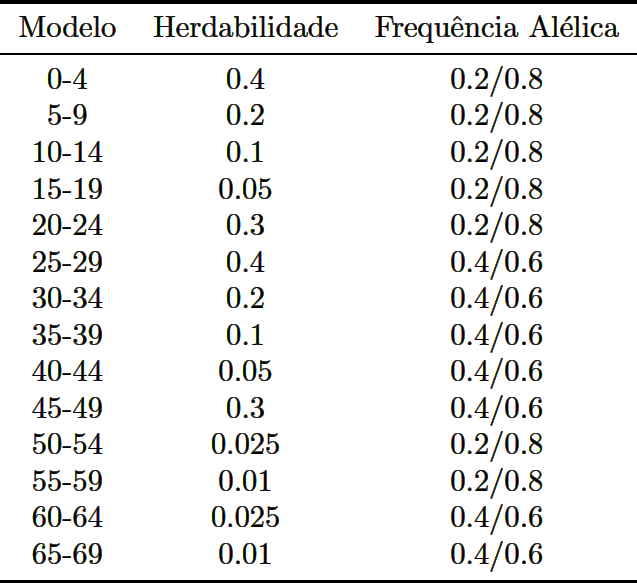
* *Herdabilidade:* As diferenças no fenótipo entre indivíduos podem ser devidos a fatores genéticos ou ambientais. O conceito estatístico da herdabilidade é a proporção de diferenças no fenótipo que são causadas por diferenças no genótipo, isto é, uma forma de medir a quantidade de impacto da carga genética dos indivíduos nas suas características fenotípicas. Essa medida é um valor real entre 0 e 1. Naturalmente, o valor (1 - herdabilidade) nos fornece a quantidade de impacto que o ambiente tem em certa característica[[2]](#footnote-2). É importante ressaltar que este conceito deve ser aplicado a populações e não a um indivíduo apenas e também que, num contexto real, ele não é estático, uma vez que as populações podem sofrer mudanças.
* *Frequência do Menor Alelo (MAF):* É a frequência com que a variante menos comum do polimorfismo ocorre. Pela nomenclatura utilizada no capítulo 2 onde as possíveis combinações alélicas eram GG, Gg e gg, essa é a frequência com a qual o alelo g ocorre na população. A pouco tempo restringia-se o conceito de polimorfismo apenas a variações pontuais que continham a frequência do menor alelo maior ou igual a 1%, porém esse conceito está sendo reconstruído atualmente, inclusive já existindo alguns polimorfismos na base *dbSNP* que contêm MAF menor do que 1%.

A partir dos parâmetros mencionados, é possível criar uma tabela de penetrância de dimensionalidade igual à ordem de interação que se deseja modelar. Penetrância é definido como sendo a probabilidade P de uma condição póstuma de doença D dada uma particular combinação entre genes G. isto é, P(D|G). A tabela é gerada ao se calcular tal probabilidade para todas as combinações possíveis de genótipos [Moore e White 2007].

O escopo deste trabalho não incluiu a forma como esses dados são gerados e não houve acesso ao detalhamento dessa simulação de dados (como a explicitação de quais funções estatísticas foram utilizadas) no estudo de [Velez et al.]. O estudo realizado por [Li e Reich 1999] discute esse assunto de uma forma interessante. Eles enumeraram todas as possíveis tabelas de penetrância não-redundantes (efeitos não óbvios) com modelos genéticos de interação binários (tabelas bidimensionais com nove células), nomeando alguns casos particulares baseando-se em suas propriedades.

No caso utilizado neste trabalho [Velez et al. 2007], foram criados modelos de interação epistáticas através da variação de parâmetros definida no Quadro 4.1. Perceba que o total de modelos gerados é 70, consistindo em todas as possibilidades possíveis entre 2 valores de frequência de menor alelo, 7 valores de herdabilidade e 5 tipos de funções estatísticas diferentes. Todas as tabelas de penetrância ordenadas por modelo estão dispostas no apêndice C.

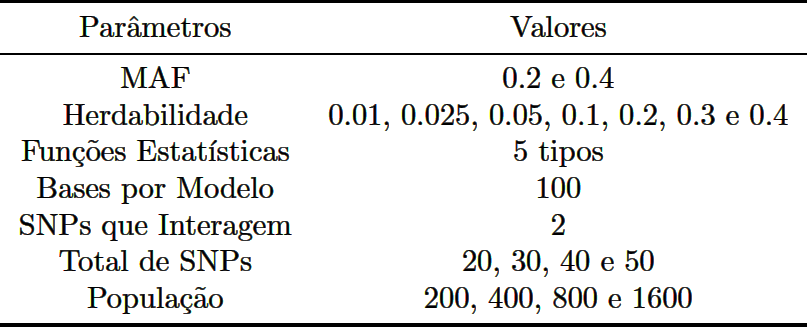
Para cada modelo epistático gerado, foram criadas 100 bases de dados. As bases de dados obtidas possuíam 1000 SNPs cada uma. As análises com números tão elevados de polimorfismos se faz impossível dada a quantidade de bases a serem testadas e o caráter combinatório dos métodos, portanto estas bases foram redimensionadas, criando bases com 20, 30, 40 e 50 SNPs. O número de indivíduos (população) das bases foi mantido idêntico, a saber, 200, 400, 800 e 1600.



**Quadro 4.1.** Detalhes da Herdabilidade e MAF dos Modelos Epistáticos

Fonte: Criada por Flávia Araújo baseada na definição dos modelos realizada em [Velez et al. 2007]

Para um dado tamanho de base (número de SNPs, número de indivíduos) existem os 70 modelos diferentes, cada um deles contendo 100 bases. O total de bases por tamanho de base é 7000, fazendo um total de 56000 bases analisadas nestes experimentos. O Quadro 4.2 mostra um resumo de todos estes parâmetros associados aos conjuntos de dados do experimento.



**Quadro 4.2**. Parâmetros Gerais da Criação do Conjunto de Dados

Fonte: Criação própria baseada na definição dos modelos realizada em [Velez et al. 2007]

## 4.2. Experimento Comparativo

O experimento visa comparar os métodos de acordo com sua capacidade de detectar interações entre polimorfismos (taxa de acerto de predições de relações epistáticas) e seu desempenho em relação ao tempo de execução.

### 4.2.1. Taxa de acerto

Cada base gerada através do procedimento descrito no item anterior possui somente dois SNPs interagindo conforme os modelos epistáticos descritos. Os outros polimorfismos são atributos aleatórios cuja interação ocorre meramente por chance.

A análise consiste na aplicação dos quatro métodos descritos no capítulo 3 nessas bases de dados com o parâmetro «ORDER» = 2, isto é, procurando por interações entre dois SNPs. Os algoritmos irão retornar uma lista contendo todas as combinações binárias de polimorfismos. Caso a combinação SNP-SNP mais bem colocada seja exatamente a combinação cuja interação é conhecida previamente, é computado um acerto para aquele método naquele modelo. Como cada modelo contém 100 bases, o total de acertos também é a percentagem de acerto, descrita neste documento como taxa de acerto. A taxa de acerto se refere, portanto, a um método aplicado a um modelo epistático.

O objetivo deste estudo comparativo é analisar se os métodos conseguem detectar esses modelos epistáticos em meio a perturbações aleatórias. Percebe-se desta forma que quanto mais baixos forem os valores da herdabilidade e a frequência do menor alelo, mais difícil essa interação é percebida. Com valores de herdabilidade baixos, o fenótipo não consegue ser explicado pelo genótipo de uma forma muito clara, diminuindo assim a sensitividade do método. Por outro lado, conforme são baixados os valores da frequência do menor alelo, é reduzida também a especificidade do método, que não terá informações suficientes em algumas células da tabela de genótipo-fenótipo dado que o alelo menos frequente ocorre mais raramente.

### 4.2.2. Tempo de execução

Durante a avaliação anterior também foram armazenados o tempo de execução de cada método aplicado em cada base de dados. Como os únicos parâmetros que afetam o tempo de execução são o número de SNPs e o número de indivíduos (definindo o tamanho da entrada), foi calculada uma média simples entre todas as execuções de todos os modelos. O resultado desta análise é um tempo de execução médio por base para cada combinação entre os parâmetros que definem o tamanho da entrada de dados.

Esta análise deve ser tratada apenas como um indicador da complexidade computacional de cada método. O fator tempo de execução é bastante difícil de ser mensurado pois está sujeito a outros fatores cujo controle é bastante difícil em um computador normal tais como os métodos como o processador ou o sistema operacional realizam as operações internas. O procedimento de extração de uma média entre uma grande quantidade de análises (7000) faz com que o resultado seja bastante confiável, porém ainda assim ele deve ser visto como um indicador apenas.

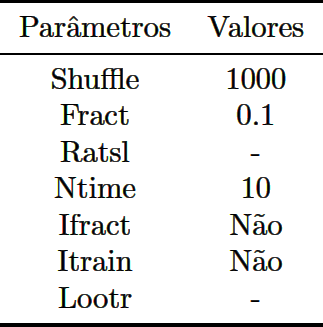
Todas as análises foram realizadas em um PC Sony VAIO modelo VPCZ12C7E (Z-Series) com processador Intel Core i7 CPU M620 2.67GHz, 6GB RAM (sendo 5.68GB usáveis), no sistema operacional Windows 7 Professional 64-bit. Todas as análises foram executadas a partir da versão final do projeto do sistema Web, porém não através da versão disposta na rede e sim através de um arquivo Java que chamava as funções necessárias nas bases que estavam em diretórios locais. As análises não foram executadas a partir do sistema Web disponível pois isso tornaria a análise bastante instável devido à instabilidade no tempo de resposta da Internet ou da rede local a partir do servidor Apache Tomcat 5.5. Em todas as análises o computador foi mantido em local regularmente resfriado e nenhum outro processo, além dos inevitáveis, ocupava a CPU.

### 4.2.3. Parâmetros utilizados nos métodos

Nos quadros 4.3, 4.4 e 4.5 são definidos os parâmetros utilizados durante a execução de todos os experimentos para os métodos, respectivamente, PIA, MDR e MASS.

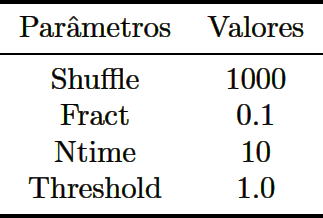
O único parâmetro do ESNP2 existe para regular o seu teste de *boostrap*, porém como essa análise foi excluída do escopo deste processo de experimentação, não existe nenhum parâmetro a ser definido para este algoritmo. O método novo (MASS) foi rodado duas vezes (R1 e R2 no quadro 4.5) com a única diferença o fato de que, na segunda rodada, a métrica APD não foi utilizada para o cálculo da métrica geral.

Os parâmetro «RATSL» do PIA e do MASS não possuem valor definido pois a base de dados não continha nenhum dado faltoso, única forma de torná-la desbalanceada. O parâmetro LOOTR do PIA não foi definido pois como mostra o parâmetro ITRAIN o conjunto de treino não foi utilizado. Como a base de dados é simples a variação destes parâmetros não iria alterar os resultados de forma significativa (isso foi observado empiricamente porém os resultados não são exibidos neste trabalho).



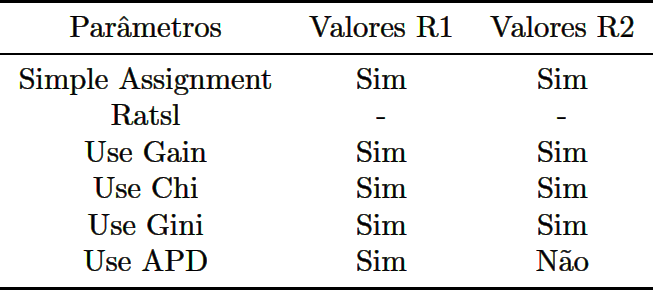
**Tabela 4.3.** Parâmetros de Execução do PIA

Fonte: Criação própria



**Tabela 4.4.** Parâmetros de Execução do MDR

Fonte: Criação própria



**Tabela 4.5.** Parâmetros de Execução do MASS

Fonte: Criação própria

"People love chopping wood. In this activity one immediately sees results."

- [Albert Einstein](http://www.brainyquote.com/quotes/quotes/a/alberteins108301.html)

5. Resultados e Discussão

Os principais resultados a respeito da taxa de acerto e tempo de execução serão mostrados nesta seção. Pretende-se mostrar os resultados de forma objetiva, isto é, este capítulo se deterá apenas aos resultados que, de alguma forma, confirmem ou não as hipóteses que motivaram o estudo. Os demais resultados serão detalhados no apêndice B.

Nos dois primeiros tópicos serão discutidos os resultados a respeito da taxa de acerto dos métodos nos diferentes modelos epistáticos e o tempo de execução dos mesmos. Em cada um destes tópicos os resultados serão exibidos de maneiras diferentes para que no final seja realizada uma discussão, destacando os pontos mais interessantes que podem ser visualizados em cada gráfico ou tabela.

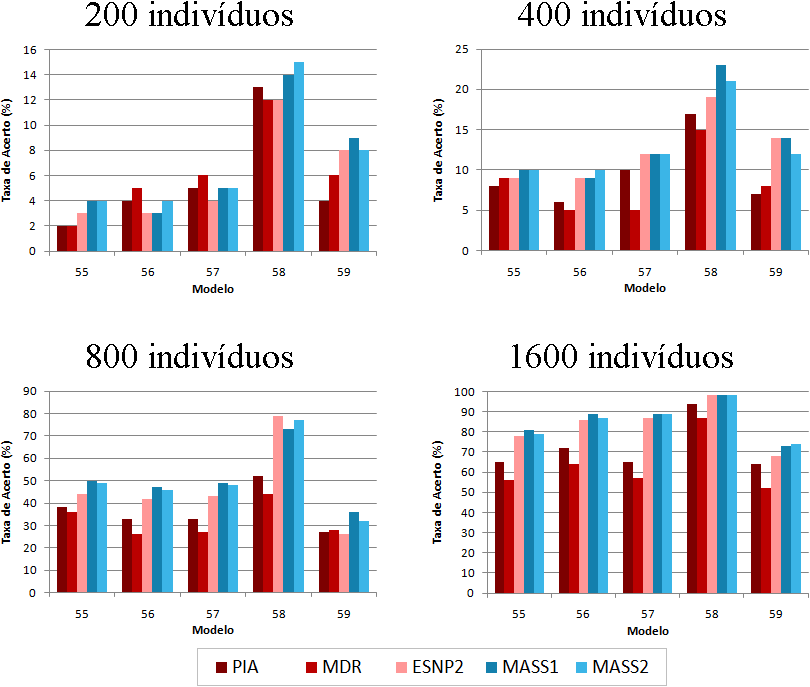
No último tópico será discutida a ferramenta Web resultante deste estudo. Será definida a forma como ela foi elaborada e com quais tecnologias. Este tópico não irá conter uma explicação detalhada sobre seu funcionamento (guia de instruções), sendo esta encontrado no apêndice A.

## 5.1. Taxa de Acerto

A taxa de acerto, conforme discutido no capítulo anterior, demonstra a eficácia do método para encontrar uma interação entre polimorfismos em meio a interações aleatórias a partir de uma base de dados gerada com a variação de parâmetros que simulam um organismo real. Os próximos itens pretendem mostrar trechos desses resultados que fornecem material interessante para discussão. Todas as taxas de acerto definidas a seguir estão em porcentagem, que também correspondem aos seus valores absolutos, dado que 100 bases foram analisadas para cada modelo.

### 5.1.1. Tamanho da entrada

Neste item são detalhadas as taxas de acerto dos métodos em relação à variação do tamanho da entrada (número de indivíduos e de polimorfismos). Os resultados dizem respeito apenas aos modelos 55 até 59 que foram gerados com herdabilidade 0.01 e MAF 0.2, isto é, são os modelos nos quais é mais difícil de se detectar as interações (conforme discutido no item 4.2.1 do capítulo anterior).

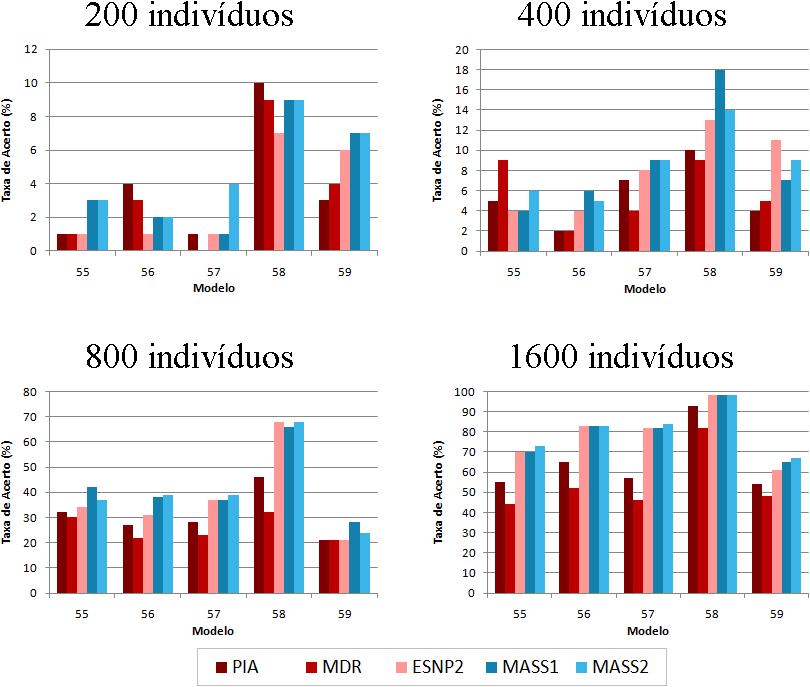


**Gráfico 5.1.** Eficácia dos Métodos para 20 SNPs

Fonte: Criação própria

Os quatro gráficos desta imagem mostram as Taxas de Acerto em porcentagem que cada método obteve para os modelos epistáticos 55, 56, 57, 58 e 59 (todos eles gerados com herdabilidade = 0.01 e MAF = 0.2). As barras em tons de vermelho mostram o desempenho dos métodos não originais estudados enquanto as barras em tons de azul mostram o desempenho do novo método. MASS1 corresponde à Taxa de Acerto do novo método considerando a métrica total formada pelas métricas: ganho de informação (entropia), Chi-Quadrado, índice Gini e APD. MASS2 corresponde à Taxa de Acerto do novo método desconsiderando a métrica APD que obteve os piores resultados entre as quatro. O eixo correspondente à Taxa de Acerto está redimensionado para facilitar a visualização das barras verticais.

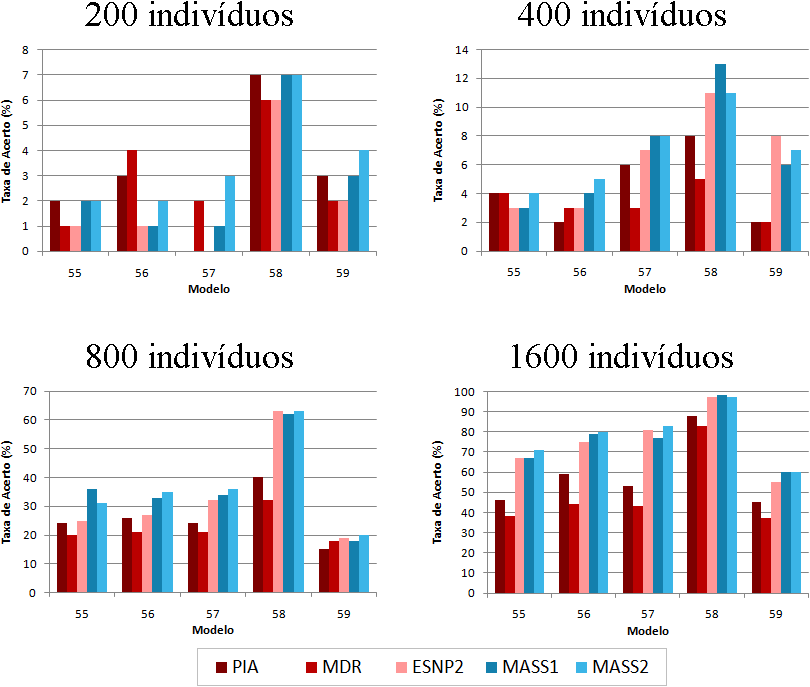
Os Gráficos 5.1, 5.2, 5.3 e 5.4 exibem tais resultados respectivamente considerando 20, 30, 40 e 50 polimorfismos em cada um. Cada gráfico contém as taxas de acerto para a quantidade de atributos determinada e para todas as quantidades de indivíduos testadas.



**Gráfico 5.2.** Eficácia dos Métodos para 30 SNPs

Fonte: Criação própria

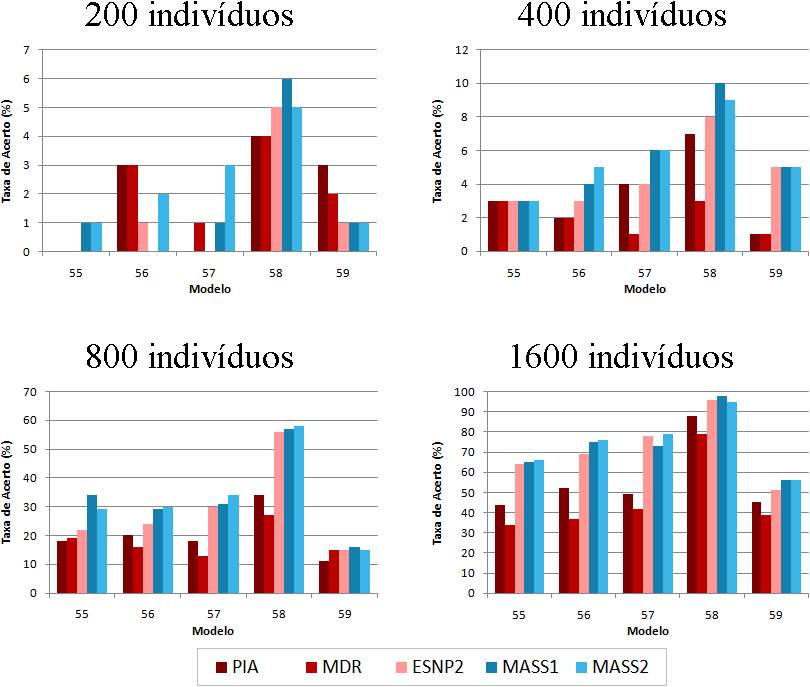
Os quatro gráficos desta imagem mostram as Taxas de Acerto em porcentagem que cada método obteve para os modelos epistáticos 55, 56, 57, 58 e 59 (todos eles gerados com herdabilidade = 0.01 e MAF = 0.2). As barras em tons de vermelho mostram o desempenho dos métodos não originais estudados enquanto as barras em tons de azul mostram o desempenho do novo método. MASS1 corresponde à Taxa de Acerto do novo método considerando a métrica total formada pelas métricas: ganho de informação (entropia), Chi-Quadrado, índice Gini e APD. MASS2 corresponde à Taxa de Acerto do novo método desconsiderando a métrica APD que obteve os piores resultados entre as quatro. O eixo correspondente à Taxa de Acerto está redimensionado para facilitar a visualização das barras verticais.



**Gráfico 5.3.** Eficácia dos Métodos para 40 SNPs

Fonte: Criação própria

Os quatro gráficos desta imagem mostram as Taxas de Acerto em porcentagem que cada método obteve para os modelos epistáticos 55, 56, 57, 58 e 59 (todos eles gerados com herdabilidade = 0.01 e MAF = 0.2). As barras em tons de vermelho mostram o desempenho dos métodos não originais estudados enquanto as barras em tons de azul mostram o desempenho do novo método. MASS1 corresponde à Taxa de Acerto do novo método considerando a métrica total formada pelas métricas: ganho de informação (entropia), Chi-Quadrado, índice Gini e APD. MASS2 corresponde à Taxa de Acerto do novo método desconsiderando a métrica APD que obteve os piores resultados entre as quatro. O eixo correspondente à Taxa de Acerto está redimensionado para facilitar a visualização das barras verticais.



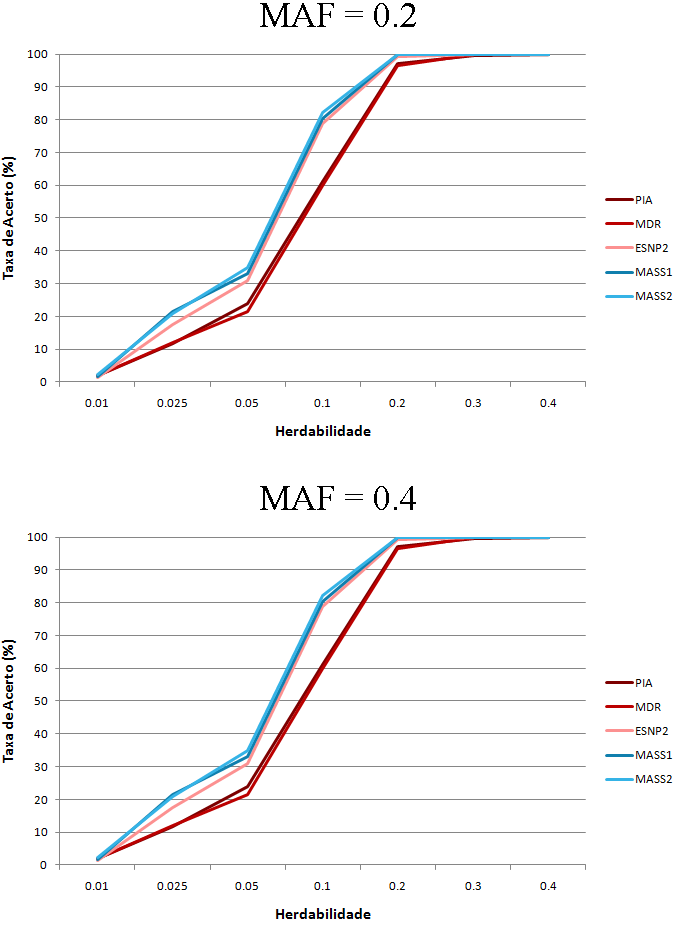
**Gráfico 5.4.** Eficácia dos Métodos para 50 SNPs

Fonte: Criação própria

Os quatro gráficos desta imagem mostram as Taxas de Acerto em porcentagem que cada método obteve para os modelos epistáticos 55, 56, 57, 58 e 59 (todos eles gerados com herdabilidade = 0.01 e MAF = 0.2). As barras em tons de vermelho mostram o desempenho dos métodos não originais estudados enquanto as barras em tons de azul mostram o desempenho do novo método. MASS1 corresponde à Taxa de Acerto do novo método considerando a métrica total formada pelas métricas: ganho de informação (entropia), Chi-Quadrado, índice Gini e APD. MASS2 corresponde à Taxa de Acerto do novo método desconsiderando a métrica APD que obteve os piores resultados entre as quatro. O eixo correspondente à Taxa de Acerto está redimensionado para facilitar a visualização das barras verticais.

### 5.1.2. Variação de parâmetros biológicos

Neste item são exibidas as taxas de acerto para a variação de parâmetros de herdabilidade e MAF (Gráfico 5.5) a partir de um tamanho de entrada específico (50 SNPs e 200 indivíduos). Tal tamanho de entrada foi escolhido nesta visualização por ser o mais "difícil" (pois possui mais características a serem analisadas com uma quantidade menor de informações), assemelhando-se aos experimentos biológicos reais.



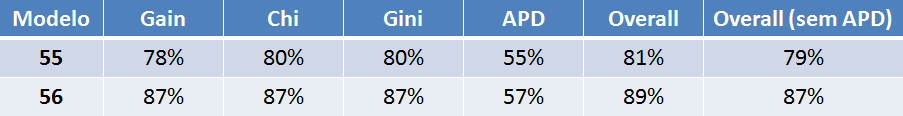
**Gráfico 5.5.** Eficácia dos Métodos em Relação aos Parâmetros Herdabilidade e MAF

Fonte: Criação própria

São exibidos os resultados de Herdabilidade vs. Taxa de Acerto para MAF = 0.2 (gráfico superior) e para MAF = 0.4 (gráfico inferior). É importante ressaltar que a escala do eixo da herdabilidade não está no formato numérico. Cada Taxa de Acerto corresponde à média entre as taxas dos 5 modelos de cada combinação de herdabilidade e MAF.

### 5.1.3. Método MASS

Os resultados para as métricas individuais que definem o novo método MASS são detalhados neste item. O principal objetivo desta exibição é defender alguns pontos importantes a respeito deste novo algoritmo desenvolvido. A Tabela 5.1 mostra trechos dos resultados mais interessantes para ilustrar algumas das hipóteses formuladas durante a fase de motivação do método (item 3.5.1).



**Tabela 5.1.** Trecho dos Resultados Detalhados do Método MASS

Fonte: Criação própria

Esta tabela foi criada com as taxas de acerto para cada métrica individual do método MASS. As informações foram extraídas dos modelos 55 e 56 para a execução com 20 SNPs e 1600 indivíduos. Gain, Chi, Gini, APD, Overall e Overall (sem APD) se referem respectivamente ao ganho de informação (entropia), Chi-Square, índice Gini, APD, métrica Overall resultante da combinação de todas as outras métricas e métrica Overall resultante da combinação de todas com exceção do APD (que obteve os menores índices de acerto).

### 5.1.4. Discussão

Em relação à taxa de acerto, todos os métodos mostraram uma notável capacidade em encontrar os polimorfismos que interagiam, destacando-se o método ESNP2 e MASS. Como os dados foram sintetizados, a diferença entre essas taxas não permite a inferência de um método ideal, dado que bases biológicas reais contêm muitas peculiaridades. Foi observado que os métodos PIA e MDR têm maior independência de modelo, isto é, não são afetados por mudanças radicais em padrões epistáticos numa mesma base, enquanto o ESNP2 toma vantagem do fato de ser independente dos efeitos marginais. Isso explica o alto desempenho do método ESNP2, visto que as bases geradas respeitavam apenas um modelo específico e efeitos marginais são inevitáveis ao se gerar dados através de funções que respeitam o equilíbrio de *Hardy-Weinberg* [Schwender e Ickstadt 2007].

Os gráficos do item 5.1.1 mostram que a nova abordagem é tão interessante ou até melhor do que o método que obteve as maiores taxas de acerto entre todos (ESNP2). Isso se dá principalmente pelo fato de que foram combinadas as estratégias mais interessantes entre os métodos: o uso de várias funções de classificação e a utilização exclusiva de métricas de separação. Além disso esse método consegue encontrar interações onde não necessariamente os dois genes participantes contribuem da mesma forma para os efeitos marginais de interação.

No Gráfico 5.5 é interessante observar que em ambos os resultados para MAF iguais a 0.2 e 0.4, existe uma notável vantagem do método MASS para as herdabilidades baixas 0.025 e 0.05 (formando uma leve curva para cima). Isso é bastante interessante visto que são justamente com esses valores mais baixos que os métodos geralmente obtêm seus piores resultados. O ESNP2 não atingiu a eficiência do MASS pois ele leva em consideração o fato de ambos os genes contribuírem de forma semelhante para o efeito epistático estatístico, porém quando a herdabilidade é baixa, a contribuição igualitária entre os efeitos de interação aleatórios se sobressai aos efeitos de interação reais porém com contribuições não tão igualitárias entre os polimorfismos. Isso pode ser pensado como um primeiro passo para se considerar os efeitos epistáticos em seus níveis de atuação (Figura 2.2), teoria condizente com as explicações derivadas do estudo do metaboloma.

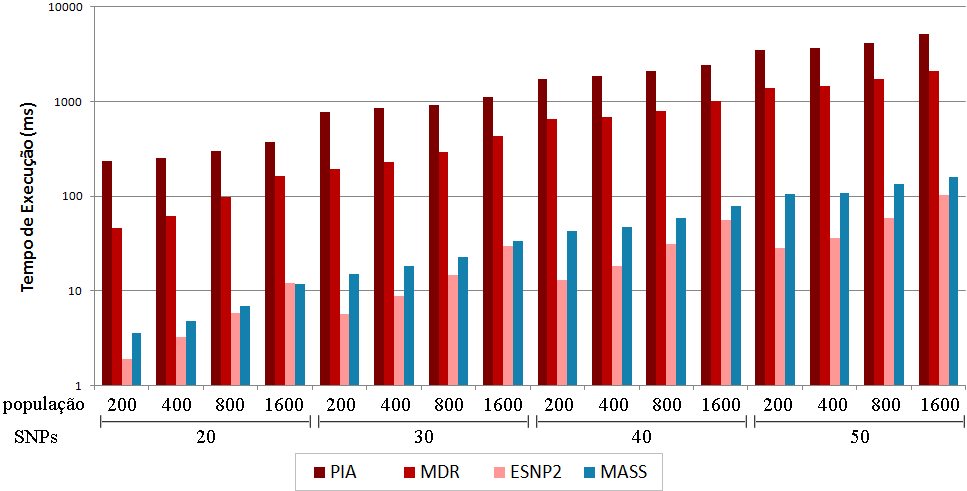
É verificada a veracidade de algumas características dos métodos, principalmente o novo método desenvolvido, a partir da observação de um pequeno trecho dos resultados exibido na Tabela 5.1. Primeiro percebe-se que a estratégia de combinar várias funções de classificação (como o PIA fez originalmente e como faz o MASS) é válida pelo fato de que as taxas de acerto da nova métrica MASS foram maiores nos do que todas as outras individualmente para alguns casos (como os exibidos na tabela). Isso também é verificado ao se observar as taxas de acerto para o MASS com e sem o APD. Mesmo esta métrica tendo obtido taxas de acertos muito inferiores às outras, removendo-a da análise final a taxa de acerto Overall cai em relação ao Overall com a métrica APD. A partir da análise deste pequeno trecho (e visto de forma periódica a partir da análise de todos os resultados no Apêndice B) as métricas baseadas em separação são muito interessantes para o problema da identificação de interação entre polimorfismos, ainda mais quando são associadas à teorias como a da *mutual information*.

## 5.2. Tempo de Execução

Os valores relativos ao tempo de execução de cada método para um dado tamanho de entrada fornece uma indicação do quão eficiente é o desempenho de cada um. Este tópico pretende exibir tais valores e realizar uma discussão a respeito de quais são os trechos dos métodos mais interessantes ao se considerar o tempo de execução. Em todos os resultados a seguir, o tempo é dado em milissegundos.

### 5.2.1. Tempo de execução geral

Este item mostra o tempo de execução de forma geral, isto é, o tempo de execução total de cada método para cada combinação de parâmetros que definem o tamanho da entrada (número de SNPs e de indivíduos).



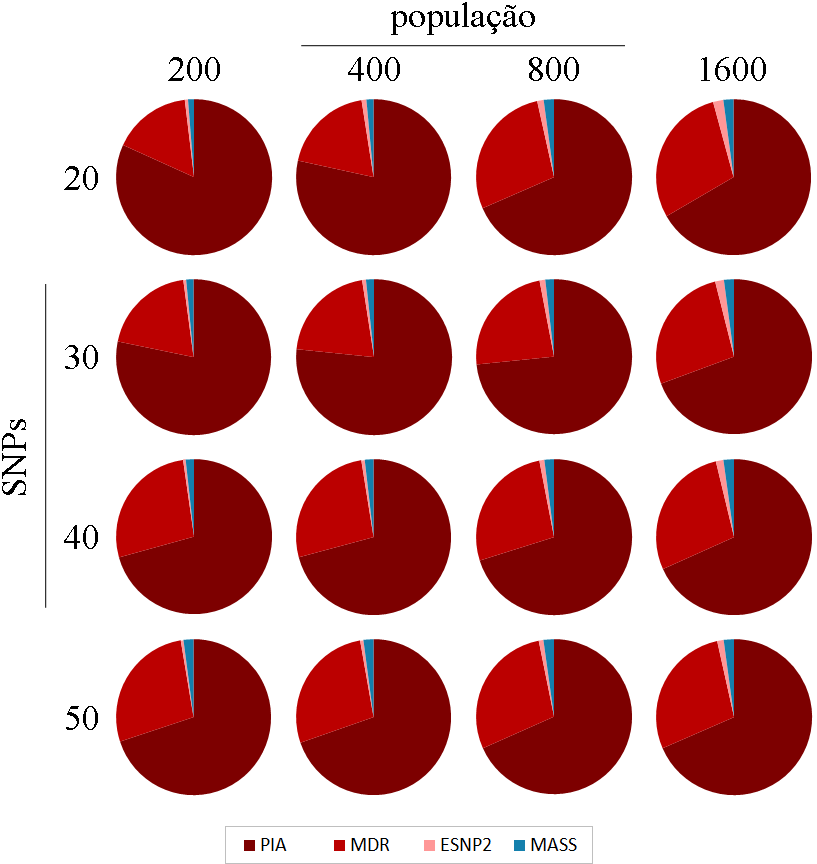
**Gráfico 5.6.** Tempo de Execução Geral

Fonte: Criação própria

O gráfico exibe o tempo de execução em escala logarítmica e em milissegundos para cada combinação dos parâmetros que definem o tamanho da entrada: número de indivíduos (população) e número de SNPs.

O Gráfico 5.6 exibe tais resultados com o tempo de execução na escala logarítmica. Tal escala foi adotada devido à discrepância entre o tempo de execução dos métodos PIA e MDR e os métodos ESNP2 e MASS. Percebe-se que, como a escala é logarítmica, saltos lineares no gráfico na verdade representam saltos exponenciais. Todos os métodos descritos crescem exponencialmente pelo fato de considerarem todas as possíveis combinações de atributos de uma determinada ordem.

### 5.2.2. Tempo de execução relativo



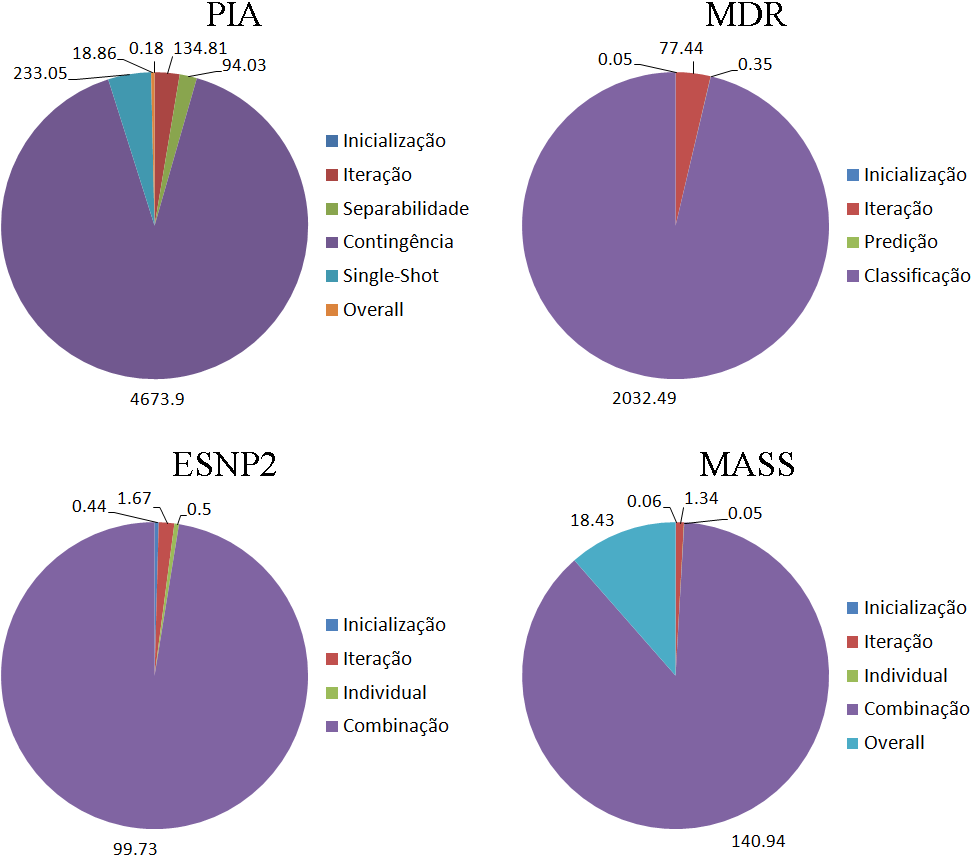
**Gráfico 5.7.** Tempo de Execução Relativo

Fonte: Criação própria

O gráfico exibe os tempos de execução relativos entre os métodos. Cada gráfico de torta representa o tempo de cada método em relação ao tempo total dos métodos para uma determinada combinação de número de SNPs e indivíduos.

Neste item são definidos os tempos de execução relativos entre os métodos para todos os tamanhos de entradas possíveis. Os valores do tempo não são exibidos pois não são importantes nesta visualização, cujo objetivo principal é mostrar o quanto cada método se comporta em relação ao outro ao se aumentar a quantidade de polimorfismos e indivíduos. Esse resultado pode ser verificado no Gráfico 5.7.

### 5.2.3. Tempo de execução específico para cada método

Neste item são focados os tempos de execução de cada trecho do algoritmo de cada um dos métodos para o modelo mais custoso apenas (50 SNPs e 1600 indivíduos).

**Gráfico 5.8.** Tempo de Execução das Fases de cada Método

Fonte: Criação própria

O gráfico exibe os tempos de execução para cada fase dos algoritmos. Esses tempos correspondem às análises com número de SNPs igual a 50 e população igual a 1600.

A etapa de inicialização e iteração, comuns a todos os métodos, dizem respeito às fases onde são realizados os procedimento iniciais de alocação de recursos ou pré-processamento (apenas no MDR) e de transações comuns durante a iteração dos métodos. As etapas de separabilidade e contingência calculam, respectivamente, as métricas de separação e baseadas na tabela de contingência. A fase de *Single-Shot* do PIA se refere a uma análise realizada a partir de um procedimento de *Leave-One-Out* a partir de toda base de dados com as métricas baseadas na tabela de contingência. Este tipo de teste pode ser utilizado como "controle" em análises mais aprofundadas porém, neste trabalho, tal metodologia não foi considerada. A fase *Overall* (assim como no MASS) se refere ao cálculo da métrica final relativa ao somatório de todas as outras métricas normalizadas. No MDR a predição e classificação dizem respeito a essas respectivas fases conforme detalhado no tópico 3.3. No ESNP2 as fases individual e combinação se referem, respectivamente, às fases onde são calculadas as entropias para cada atributo individualmente e para a combinação entre atributos. No MASS essas duas fases também tem os mesmos significados porém nelas são calculadas todas as métricas utilizadas no algoritmo.

### 5.2.4. Discussão

Em relação à análise temporal o ESNP2 obteve o melhor desempenho entre todos os métodos conforme esperado, já que ele não realiza nenhuma estratégia de validação cruzada (realizada pelo PIA e pelo MDR). O método PIA apesar de ter obtido resultados um pouco melhores quanto à taxa de acerto, teve tempo de execução maior do que o MDR, resultado também esperado dadas as dicotomias: complexo e lento (PIA) contra simples e rápido (MDR). Outra observação interessante é o fato de que o tempo de execução aumenta consideravelmente quando o número de SNPs aumenta, porém quando o tamanho da população cresce, esse tempo não tende a crescer tão rapidamente (Gráfico 5.6) sendo o ESNP2 a única exceção a essa regra já que seu processamento é suficientemente leve. Apesar do novo algoritmo MASS ter se saído bem melhor do que todos os outros em relação à eficácia da detecção de interações, ele perdeu em praticamente todos os casos para o ESNP2 em relação ao tempo computacional, porém pelo Gráfico 5.6 podemos verificar que ele não é tão custoso quanto o PIA e o MDR, perdendo para o ESNP2 por uma pequena margem.

O Gráfico 5.7 mostra algumas coisas interessantes como o fato de que o conforme aumentamos a população o MDR tende a se igualar ao PIA, porém conforme aumentamos o número de polimorfismos envolvidos, essa tendência é atenuada. Isso mostra que o MDR, apesar de obter taxas piores de acerto está mais preparado para ser executado em bases de dados grandes do que o PIA. Esse gráfico também exibe o fato de que conforme é aumentado o número de indivíduos da base de dados, o tempo computacional do ESNP2 tende a se aproximar do MASS, mostrando outra vantagem desta nova abordagem em relação a conjuntos de dados extremamente grandes tais como os fornecidos pelos GWAS.

Por fim, o gráfico 5.8 mostra claramente que a grande responsável pelo aumento no tempo computacional dos métodos PIA e MDR é a estratégia de validação cruzada. Tal estratégia é necessária em todos os métodos que utilizam algum algoritmo de classificação para que sejam acessadas as taxas de acerto mais precisas utilizando uma base de dados altamente especializada, isto é, não podemos utilizar dados de uma patologia para classificar um algoritmo de aprendizagem de máquina para outra doença. As estratégias baseadas em separação, típicas de estruturas como árvores de decisão são preferíveis inclusive pela sua maior eficácia (como pôde ser visto no tópico 5.1). Algumas estratégias estão sendo utilizadas em troca da validação cruzada como abordagens baseadas em *bootstrapping* (amostragem com reposição) [Bureau et al. 2005].

## 5.3. Ferramenta Web

Os métodos estudados neste trabalho foram implementados utilizando a linguagem de programação Java. Um módulo Web, que permite o fácil acesso a tais análises, foi construído utilizando a tecnologia Java Server Faces (JSF) v.1.1 rodando sobre o servidor Apache Tomcat v.5.5. A ferramenta está disponível publicamente no seguinte endereço: <<https://jaqueira.cin.ufpe.br/pit/faces/index.jsp>>.

A ferramenta está dividida em três seções principais: Análise, Métodos e Referências. Na primeira seção podemos realizar a análise de uma base de dados com um formato pré-estabelecido com cada um dos quatro métodos individualmente ou simultaneamente. Nesta tela de Análise os parâmetros específicos de cada algoritmo devem ser definidos de acordo com a preferência do usuário. O resultado da análise consiste em uma lista que relaciona cada interação com sua pontuação e também uma análise do tempo de execução do algoritmo para cada etapa.

Na seção Métodos, as técnicas, parâmetros e forma de exibição dos resultados estão descritos de forma breve e concisa. Também é definido o formato do arquivo que deverá conter a entrada do algoritmo, isto é, a matriz de indivíduos vs. polimorfismos descrevendo os dados genotípicos e ambientais a nível individual. Na última seção temos as referências utilizadas para cada método com um link direcionado para a publicação original.

O Apêndice A deste trabalho fornece um guia para a realização de uma análise utilizando a ferramenta proposta.

"Finally, in conclusion, let me say just this."

- [Peter Sellers](http://www.brainyquote.com/quotes/quotes/p/peterselle125575.html)

6. Conclusão

Nesta seção serão feitas as considerações finais a respeito deste trabalho de graduação. Inicialmente serão discutidos os objetivos atingidos juntamente com um sumário geral sobre este estudo, frisando as hipóteses realizadas e os resultados obtidos. Nos dois tópicos seguintes serão retratadas as limitações do escopo deste projeto e as principais dificuldades durante a realização do mesmo. Finalmente serão fornecidas as diretrizes para a continuação deste estudo, consistindo basicamente da aplicação do método descrito numa base de dados biológica real.

## 6.1. Objetivos Atingidos

Este trabalho teve como objetivo principal o estudo de métodos para analisar a relação entre variáveis genéticas e doenças comuns e complexas. Para isso foi realizado um experimento a partir de um conjunto de dados gerados estatisticamente de forma a modelar um experimento biológico real e aplicados três métodos estudados. A partir da análise dos resultados preliminares foram identificados os pontos fortes e fracos de cada método e foi proposto um novo algoritmo com objetivo de suprir as deficiências das abordagens anteriores.

Além do estudo comparativo, foi criada uma ferramenta Web disponível publicamente para realizar tais análises com todos os métodos desenvolvidos. Além das análises propriamente ditas, o site contém informações a respeito dos métodos e de suas publicações originais.

Pode ser afirmado que o estudo atingiu seus objetivos, criando um estudo de caso completo no que diz respeito à interação entre polimorfismos de único nucleotídeo. A principal contribuição deste estudo, além do novo método desenvolvido e da ferramenta Web, é a proposta de um paradigma experimental, baseado em conjuntos de dados sinteticamente criados, para que métodos biológicos afins sejam testados. Os resultados dos métodos estudados, considerados estado da arte da abordagem combinatória na resolução do problema de seleção de atributos em bioinformática, podem ser utilizados como base para futuros estudos.

## 6.2. Limitações de Escopo

O entendimento da forma como os dados foram gerados e a modelagem estatística por trás das relações epistáticas poderiam ser bastante reveladoras e interessantes porém não foram exaustivamente exploradas devido ao alto teor estatístico. Este trabalho também não gerou suas próprias bases de dados o que teria sido interessante para permitir a definição completa das funções estatísticas utilizadas e a utilização de um conjunto maior de valores dos parâmetros (herdabilidade e MAF).

O tamanho dos dados analisados foi restrito devido à falta de tempo hábil para analisar métodos combinatórios. Num estudo biológico real, principalmente no paradigma do GWAS, as bases de dados geralmente consistem de um número muito grande de polimorfismos para um número razoável de indivíduos. A falta de tempo hábil mencionada se refere ao fato de que uma quantidade muito grande de bases de dados foi executada justamente para se avaliar estatisticamente a significância dos métodos. Todos os métodos aqui descritos são aplicáveis em bases de dados maiores, já que nesses estudos reais a quantidade de bases de dados a serem analisadas é bem menor (geralmente uma, com a quantidade total de pacientes do estudo ou poucas dezenas caso se deseje realizar um estudo com quantidade variável de controles).

Para validar de forma completa os algoritmos estudados, testes estatísticos são necessários para comprovar a autenticidade dos resultados. O teste de *bootstrapping* realizado por [Dong et al. 2007] é uma forma interessante, com fácil visualização e poder de interpretação. Porém ele é bastante custoso em vista de que é necessário rodar cada método pelo menos 1000 vezes para se obter *p-values* com uma resolução de 0.001, por este motivo ele não pôde ser executado. Outros testes estatísticos também podem ser aplicados.

## 6.3. Dificuldades

A principal limitação deste trabalho consistiu na dificuldade da obtenção de dados biológicos reais para uma análise experimental mais profunda. A análise com dados reais é muito importante para testar de fato a capacidade de um método em alcançar os objetivos propostos e as hipóteses formuladas. Alguns métodos podem ser aplicados com mais sucesso em certos tipos de doença, enquanto outros são mais bem sucedidos em estudos que também envolvam características ambientais, etc. E esse tipo de conhecimento só é possível de ser obtido através da experimentação *in silico* com dados de estudos de associação reais. A dificuldade de se obter tais dados se deve a dois fatores: primeiro, os dados utilizados neste tipo de estudo que envolve interações entre atributos deve estar obrigatoriamente no nível individual. Dados a nível coletivo a respeito de câncer de mama foram obtidos [Pharoah et al. 2007], porém a partir deles só é possível estudar o efeito de um atributo individual para o dado fenótipo. A outra dificuldade é o fato de que dados a nível individual em estudos de associação com doenças envolvendo pacientes reais possuem diversas restrições éticas e de privacidade. A requisição desses dados é um processo bastante burocrático e lento, sendo as requisições feitas durante o desenvolvimento do projeto ainda pendentes na data de entrega do mesmo. Também foram obtidos dados de estudos de trios a respeito de esquizofrenia [Fallin et al. 2005], onde cada célula de dado contém informações a respeito de um indivíduo e de seus genitores, porém como esse paradigma exige algumas técnicas especiais e o foco não era exatamente esse, os dados não foram utilizados.

## 6.4. Trabalhos Futuros

Os próximo passo deste estudo é o entendimento mais aprofundado das relações epistáticas num contexto genético, biológico e estatístico através da análise da forma como os dados de polimorfismos se comportam. Com isso pretende-se fornecer uma explicação biológica completa a respeito da eficácia do novo método apresentado.

Posteriormente, o experimento poderá ser repetido para uma quantidade maior de valores dos parâmetros que definem os modelos epistáticos da doença (herdabilidade e MAF) e dos parâmetros que definem o tamanho da base de dados (número de polimorfismos e indivíduos), sendo o aumento do número de polimorfismos uma meta importante e que pode ser alcançada através do uso de computadores mais potentes ou clusters de processamento de informação.

Finalmente, para finalizar o estudo de forma completa, um caso patológico real deverá ser analisado. Nesse ponto algumas variáveis devem ser discutidas como: a fonte desses dados, a doença a ser analisada (aplicar uma metodologia exploratória numa doença não tão estudada ou analisar uma doença bem formalizada para validar os métodos), a forma como tais experimentos serão tomados e a forma como os resultados serão visualizados.

Referências Bibliográficas

Aguiar-Pulido, V., Seoane, J.A., Rabuñal, J.R., Dorado, J., Pazos, A. and Munteanu, C.R. (2010) “Machine learning techniques for single nucleotide polymorphism – disease classification models in schizophrenia”. In *Molecules*, 15(7):4875-4889.

Aittokallio, T. and Schwikowski, B. (2006) "Graph-based methods for analysing networks in cell biology". In *Briefings in Bioinformatics*, 7(3):243-255.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002) "Molecular Biology of the cell". Garland Science, 4ª ed., 1616p.

Allen-brady, K., Miller, J., Matsunami, N., Stevens, J., Block, H., Farley, M., Krasny, L., Pingree, C., Lainhart, J., Leppert, M., McMahon, W.M. and Coon H. (2008) "A high-density SNP genome-wide linkage scan in a large autism extended pedigree". In *Molecular Psychiatry*, 14(6):590-600.

Ban, H.J., Heo, J.Y., Oh, K.S. and Park, K.J. (2010) “Identification of type-2 diabetes-associated combination of SNPs using Support Vector Machine”. In *BMC Genetics*, 11:26-37.

Bateson, W. (1909) "Mendel's principles of heredity". Cambridge University Press.

Bertram, L., McQueen, M.B., Mullin, K., Blacker, D., Tanzi, R.E. (2007) “Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database”. In *Natural Genetics*, 39(1):17-23.

Boulesteix, A.L., Strobl, C., Weidinger, S., Wichmann, H.E. and Wagenpfeil, S. (2007) “Multiple testing for SNP-SNP interactions”. In *Institut für Statistik*, Technical Report no.004.

Breiman L. (2001) “Random Forests”. In *Machine Learn*, 45:5-32.

Briggs, F.B., Ramsay, P.P., Madden, E., Norris, J.M., Holers, V.M., Mikuls, T.R., Sokka, T., Seldin, M.F., Gregersen, P.K., Criswell, L.A. and Barcellos, L.F. (2010) “Supervised machine learning and logistic regression identifies novel epistatic risk factors with PTPN22 for rheumatoid arthritis”. In *Genes Immunology*, 11:199-208.

Brinza, D. and Zelikovsky, A. (2007) “Risk factor searching heuristics for SNP case-control studies”. In *Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine*, 282-287.

Buntine, W. and Niblett, T. (1992) "A further comparison of splitting rules for decision-tree induction". In *Machine Learning*, 8:75-85.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P. and Van Eerdewegh, P. (2005) “Identifying SNPs Predictive of Phenotype Using Random Forests”. In *Genetic Epidemiology*, 28(2):171-182.

Chakravarti, A. (1998) “It's raining SNPs, hallelujah?”. In *Nature Genetics*, 19:216-217.

Chakravarti, A. (2001) “Single Nucleotide Polymorphism ...to a future of genetic medicine”. In *Nature*, 409:822-823.

Chatterjee, N. and Carroll, R.J. (2005) "Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies". In *Biometrika*, 92(2):399-418.

Chu, L. and Chen, B. (2008) "Construction of a cancer perturbed protein-protein interaction network for discovery of apoptosis drug targets". In *BMC Systems Biology*, 2:56-73.

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

Culverhouse, R., Suarez, B.K., Lin, J. and Reich, T. (2002) "A perspective on epistasis: limits of models displaying no main effect". In *American Journal of Human Genetics*, 70(2):461-471.

Dong, C., Chu, X., Wang, Ying, Wang, Yi, Jin, L., Shi, T., Huang, W. and Li, Y. (2008) “Exploration of gene-gene interaction effects using entropy-based methods”. In *European Journal of Human Genetics*, 16(2):229-235.

Edwards, T.L., Bush, W.S., Turner, S.D., Dudek, S.M., Torstenson, E.S., Schmidt, M., Martin, E. and Ritchie, M.D. (2008) “Generating linkage disequilibrium patterns in data simulations using genomeSIMLA”. In *Lecture Notes on Computer Science*, 4973:24-35.

Fallin, M.D., Lasseter, V.K., Avramopoulos, D., Nicodemus, K.K., Wolyniec, P.S., McGrath, J.A., Steel, G., Nestadt, G., Liang, K.Y., Huganir, R.L., Valle, D. and Pulver, A.E. (2005) "Bipolar I disorder and schizophrenia: a 440-single-nucleotide polymorphism screen of 64 candidate genes among ashkenazi jewish case-parent trios". In *American Journal of Human Genetics*, 77(6):918-936.

Feng, I.J. and Radivoyevitch, T. (2009) “SNP-SNP interactions between dNTP supply enzymes and mismatch DNA repair in breast cancer”. In *Ohio Collaborative Conference on Bioinformatics 2009*, 123-128.

Ferrer-Costa, C., Orozco, M. and deLa Cruz, X. (2002) “Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties”, In *Journal of Molecular Biology*, 315(4):771-786.

Fisher, R.A. (1918) "The correlations between relatives on the supposition of Mendelian Inheritance". In *Transactions of the Royal Society of Edinburgh*, 52:399-433.

Freitas, A.A. (2001) "Understanding the crucial role of attribute interaction in data mining". In *Artificial Intelligence Review*, 16(3):177-199.

García-Margariños, M., López-de-Ullibarri, I., Cao, R. and Salas, A. (2009) "Evaluating the ability of three-based methods and logistic regression for the detection of SNP-SNP interaction". In *Annals of Human Genetics*, 73(Pt 3):360-369.

Gauderman, W.J., Murcray, C., Gilliland, F. and Conti, D.V. (2007) "Testing association between disease and multiple SNPs in a candidate gene". In *Genetic Epidemiology*, 31(5):383-395.

Goodman, J.E., Mechanic, L.E., Luke, B.T., Ambs, S., Chanock, S. and Harris, C.C. (2006) “Exploring SNP-SNP interactions and colon cancer risk using polymorphism interaction analysis”. In *Int Journal of Cancer*, 118(7):1790-1797.

Greene, C.S., Penrod, N.M., Kiralis, J. and Moore, J.H. (2009) "Spatially uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions". In *BioData Mining*, 2:5-13.

Günther, F., Wawro, N. and Bammann, K. (2009) "Neural networks for modeling gene-gene interactions in association studies". In *BMC Genetics*, 10:87-101.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) "Gene selection for cancer classification using support vector machines". In *Machine Learning*, 46(1-3):389-422.

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

He, H., Oetting, W.S., Brott, M.J. and Basu, S. (2009) “Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study”. In *BMC Medical Genetics*, 10:127-144.

Jakulin, A. (2005) "Machine learning based on attribute interactions". PhD Dissertation, Faculty of Computer and Information Science, University of Ljubljana.

Jakulin, A. and Bratko, I. (2004a) "Testing the significance of attribute interactions". In *Proceedings of the twenty-first international conference on Machine learning*, 52-60.

Jakulin, A. and Bratko, I. (2004b) "Quantifying and visualizing attribute interactions: an approach based on entropy". In [*http://arxiv.org/abs/cs.AI/0308002v3*](http://arxiv.org/abs/cs.AI/0308002v3).

Jakulin, A., Bratko, I., Smrke, D., Demšar, J. and Zupan, B. (2003) "Attribute interactions in medical data analysis". In *Ninth Conference on Artificial Intelligence in Medicine in Europe*.

Jiang, R., Tang, W., Wu, X. and Fu, W. (2009) “A random forest approach for the detection of epistatic interactions in case-control studies”. In *BMC Bioinformatics*, 10(Suppl 1):S65-S77.

Jones, D. (2007) “Steps on the road to personalized medicine”. In *Nature Reviews Drug Discovery*, 6:770-771.

Khoury, M.J., Evans, J. and Burke, W. (2010) “A reality check for personalized medicine”. In *Nature*, 464:680.

Kim, Y., Wojciechowski, R., Sung, H., Mathias, R.A., Wang, L., Klein, A.P., Lenroot, R.K., Malley, J. and Bailey-Wilson, J.E. (2009) "Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects". In *BMC Proceedings*, 3(Suppl 7):S64-S69.

Kooperberg, C. and Ruczinski, I. (2005) "Identifying interacting SNPs using monte carlo logic regression". In *Genetic Epidemiology*, 28(2):157-170.

Li, W. and Reich, J. (1999) "A complete enumeration and classification of two-locus disease models". In *Human Heredity*, 50(6):334-349.

Li, Q., Louis, T.A., Fallin, M.D. and Ruczinski, I. (2009) "Trio logic regression - Detection of SNP-SNP interactions in Case-Parent Trios". In *John Hopkins University, Dept. of Biostatistics Working Papers*, Working Paper 194.

Macdonald, S.J., Pastinen, T., Genissel, A., Cornforth, T.W. and Long, A.D. (2005) "A low-cost open-source SNP genotyping plataform for association mapping applications". In *Genome Biology*, 6(12):R105-R116.

Maestrini, E., Pagnamenta, A.T., Lamb, J.A., Bacchelli, E., Sykes, N.H., Sousa, I., Toma, C., Barnby, G., Butler, H., Winchester, L., Scerri, T.S., Minopoli, F., Reichert, J., Cai, G., Buxbaum, J.D., Korvatska, O., Schellenberg, G.D., Dawson, G., de Bildt, A., Minderaa, R.B., Mulder, E.J., Morris, A.P., Bailey, A.J. and Monaco, A.P. (2010) "High-density SNP association study and copy number variation of the AUTS1 and AUTS5 loci implicate the IMMP2L-DOCK4 gene region in autism susceptibility". In *Molecular Psychiatry*, 15(9):954-968.

Mao, W. (2009) "The application of Apriori-Gen algorithm on the interaction analysis of SNPs in genetic disease". In *2009 WRI World Congress on Computer Science and Information Engineering*, 626-630.

McKinney, B.A., Reif, D.M., White, B.C., Crowe, J.E. and Moore, J.H. (2007) "Evaporative cooling feature selection for genotypic data involving interactions". In *Bioinformatics*, 23(16):2113-2120.

McKinney, B.A., Crowe, J.E., Guo, J. and Tian, D. (2009) "Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis". In *PLoS Genetics*, 5(3):e1000432.

Mechanic, L.E., Luke, B.T., Goodman, J.E., Chanock, S.J. and Harris, C.C. (2008) “Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions”. In *BMC Bioinformatics*, 9:146-158.

Mingers, J. (1988) "An empirical comparison of selection measures for decision-tree induction". In *Machine Learning*, 3(4):319-342.

Moore, J.H. (2003) "The ubiquitous nature of epistasis in determining susceptibility to common human diseases". In *Human Heredity*, 56(1-3):73-82.

Moore, J.H. (2005) "A global view of epistasis". In *Nature Genetics*, 37:13-14.

Moore, J.H. (2007) "Bases, bits and disease: a mathematical theory of human genetics". In *European Journal of Human Genetics*, 16(2):143-144.

Moore, J.H. and Ritchie, M.D. (2010) "The challenges of whole-genome approaches to common diseases". In *Journal of the American Medical Association*, 291(13):1642-1643.

Moore, J.H. and White, B.C. (2007) "Genome-wide genetic analysis using genetic programming: the critical need for expert knowledge". In *Genetic Programming Theory and Practice IV*, Book Chapter, 11-28.

Moore, J.H. and Williams, S.M. (2005) "Transversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis". In *Bioessays*, 27(6):637-646.

Moore, J.H., Boczko, E.M. and Summar, M.L. (2005) "Connecting the dots between genes, biochemistry, and disease susceptibility: systems biology modeling in human genetics". In *Molecular Genetics and Metabolism*, 84(2):104-111.

Moore, J.H., Gilbert, J.C., Tsai, C., Chiang, F., Holden, T., Barney, N. and White, B.C. (2006) "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility". In *Journal of Theoretical Biology*, 241(2):252-261.

Moore, J.H., Asselbergs, F.W. and Williams, S.M. (2010) "Bioinformatics challenges for genome-wide association studies". In *Bioinformatics*, 26(4):445-455.

Mukherjee, B. and Chatterjee, N. (2008) "Exploiting gene-environment independence for analysis of case-control studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency". In *Biometrics*, 64(3):685-694.

Mukherjee, B., Ahn, J., Gruber, S.B., Rennert, G., Moreno, V. and Chatterjee, N. (2008) "Tests for gene-environment interaction from case-control data: a novel study of Type I error, power and designs". In *Genetic Epidemiology*, 32(7):615-626.

Nakamichi, R., Imoto, S. and Miyano, S. (2004) “Case-control Study of Binary Disease Trait Considering Interactions between SNPs and Environmental Effects using Logistic Regression”. In *Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering 2004*, 73-78.

Oh, S. and Park, T. (2007) "Association tests based on the principal-component analysis". In *BMC Proceedings*, 1(Suppl 1):S130-S134.

Onay, V.U., Briollais, L., Knight, J.A., Shi, E., Wang, Y., Wells, S., Li, H., Rajendram, I., Andrulis, I.L. and Ozcelik, H. (2006) “SNP-SNP interactions in breast cancer susceptibility”. In *BMC Cancer*, 6:114-130.

Packer, B.R., Yeager, M., Burdett, L., Welch, R., Beerman, M., Qi, L., Sicotte, H., Staats, B., Acharya, M., Crenshaw, A., Eckert, A., Puri, V., Gerhard, D.S. and Chanock, S.J. (2006) “SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes”. In *Nucleic Acids Research*, 32(Database issue):D528-D532.

Pauline, C.N., Murray, S.S., Levy, S. and Venter, J.C. (2009) “An agenda for personalized medicine”. In *Nature*, 461:724-726.

Peng, Q., Zhao, J. and Xue, F. (2010) "PCA-based bootstrap confidence interval tests for gene-disease association involving multiple SNPs". In *BMC Genetics*, 11:6-13.

Perrière, G., Lobry, J.R. and Thioulouse, J. (1996) "Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences". In *Bioinformatics*, 12(6):519-524.

Pharoah, P.D.P., Tyrer, J., Dunning, A.M., Easton, D.F. and Ponder, B.A.J. (2007) "Association between common variation in 120 candidate genes and breast cancer risk". In *PLoS Genetics*, 3(3):e42-e47.

Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F. and Moore, J.H. (2001) “Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer”. In *American Journal of Human Genetics*, 69(1):138-147.

Ritchie, M.D., White, B.C., Parker, J.S., Hahn, L.W. and Moore, J.H. (2003) "Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases". In *BMC Bioinformatics*, 4:28-41.

Robnik-Šikonja, M. and Kononenko, I. (2003) "Theoretical and empirical analysis of ReliefF and RReliefF". In *Machine Learning Journal*, 53(1-2):23-69.

Saangyong, U., Dong-Hoi, K., Young-Woong, K., Sungwon, C., Jaeyoun, C. and Jin, K. (2009) “A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis”. In *Expert Systems*, 26:60-69.

Schwender, H. and Ickstadt, K. (2007) “Identification of SNP interactions using logic regression”. In *Biostatistics*, 9(1):187-198.

Schwender, H. and Ickstadt, K. (2008) “Empirical Bayes analysis of single nucleotide polymorphisms”. In *BMC Bioinformatics*, 9:144-159.

Sirota, M., Schaub, M.A., Batzoglou, S., Robinson, W.H. and Butte, A.J. (2009) “Autoimmune disease classification by inverse association with SNP alleles”. In *PLoS Genetics*, 5(12):e1000792.

Stone, J.L., Merriman, B., Cantor, R.M., Geschwind, D.H. and Nelson, S.F. (2007) "High density SNP association study of a major autism linkage region on chromosome 17". In *Human Molecular Genetics*, 16(6):704-715.

Strobl, C., Boulesteix, A., Augustin, T. (2006) "Unbiased split selection for classification trees based on the gini index". In *Computational Statistics & Data Analysis*, 52(1):483-501.

The International HapMap Consortium (2003) “The International HapMap Project”. In *Nature*, 426:789-796.

Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M. and Moore, J.H. (2007) “A Balanced Accuracy Function for Epistasis Modeling in Imbalanced Datasets using Multifactor Dimensionality Reduction”. In *Genetic Epidemiology*, 31(4):306-315.

Wellcome Trust Case-Control Consortium (2010) “Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls”. In *Nature*, 464:713-720.

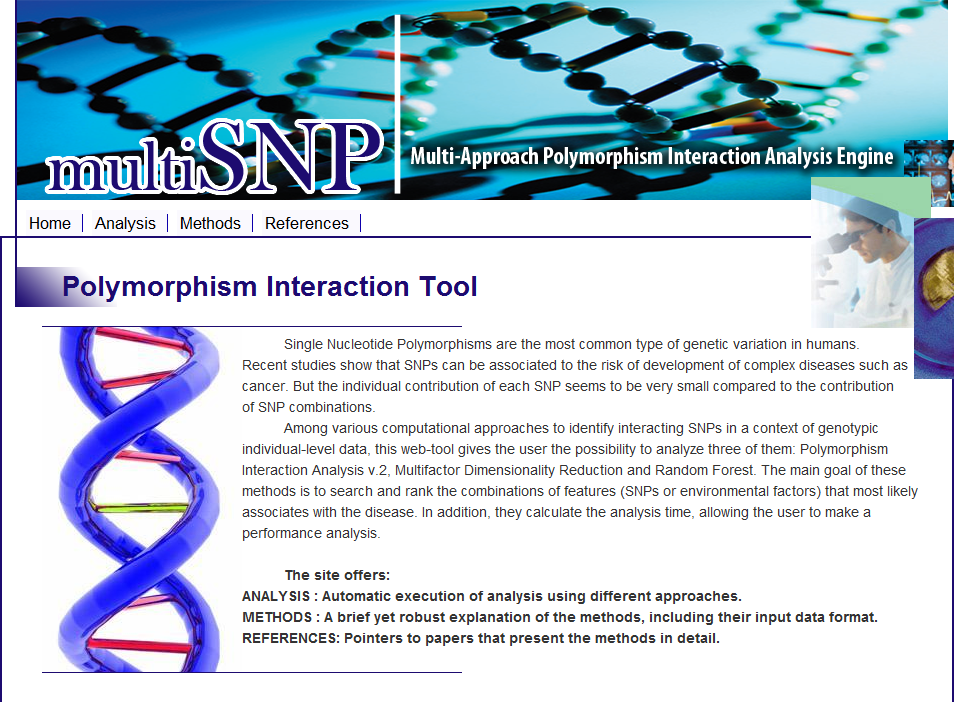
White, A.P. and Liu, W.Z. (1994) "Bias in information-based measures in decision tree induction". In *Machine Learning*, 15(1):25-41.

Williams, L.J., Abdi, H., French, R. and Orange, J.B. (2010) "A tutorial on multi-block discriminant correspondence analysis (MUDICA): A new method for analyzing discourse data from clinical populations". In *Journal of speech language and hearing research*, 53(5):1372-1393.

Wright, F.A., Huang, H., Guan, X., Gamiel, K., Jeffries, C., Barry, W.T., de Villena, F.P., Sullivan, P.F., Wilhelmsen, K.C. and Zou, F. (2007) “Simulating association studies: a data-based resampling method for candidate regions or whole genome scans”. In *Bioinformatics*, 23:2581-2588.

Apêndice A - Ferramenta Web

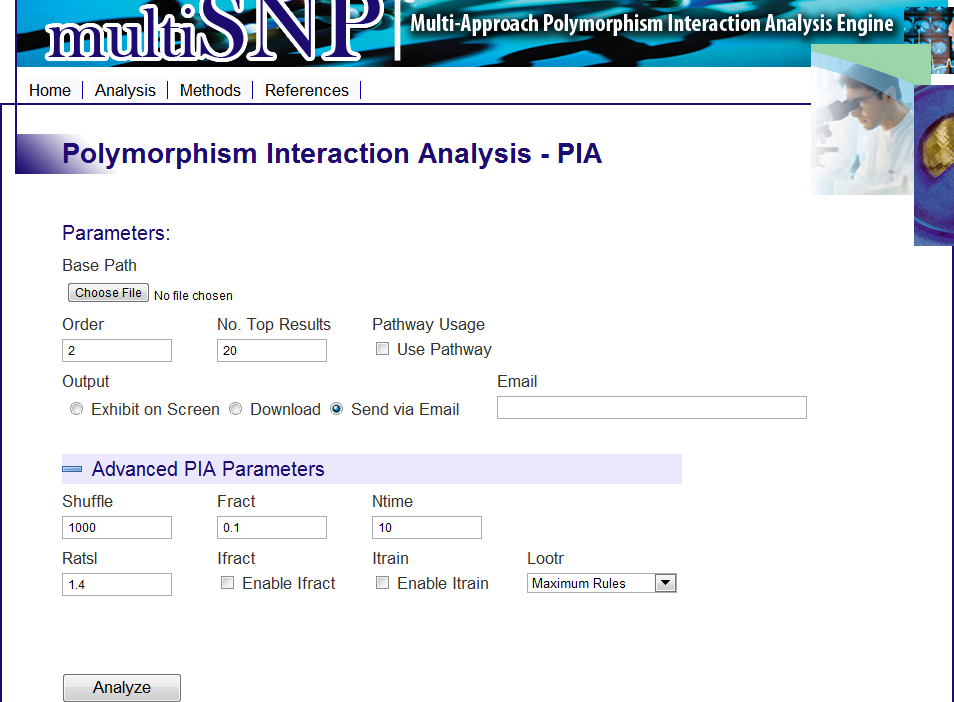
Este apêndice fornece uma visão geral do funcionamento da ferramenta Web desenvolvida neste trabalho, detalhada no tópico 5.3. Será mostrado um exemplo de análise com o método PIA, todos os outros métodos são executados de forma semelhante e possibilitam a visualização dos mesmos tipos de resultados sendo a configuração de parâmetros as únicas diferenças relevantes entre as diferentes telas de análise dos métodos. A legenda de cada imagem fornece as informações sobre cada etapa. O endereço de acesso ao site é <<https://jaqueira.cin.ufpe.br/pit/faces/index.jsp>>.



**Figura A.1.** Página Inicial

Fonte: Criação própria

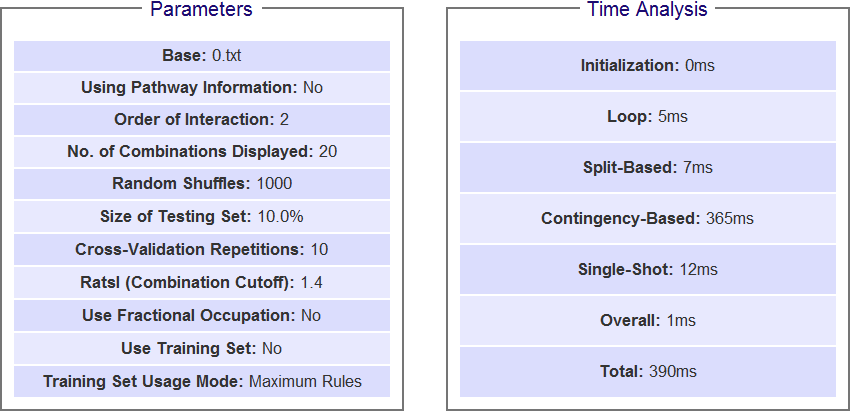
Tela inicial da Ferramenta Web. A partir do menu superior podem ser acessadas todas as opções: análise dos métodos (Analysis), detalhe dos métodos (Methods) e referências (References).



**Figura A.2.** Tela de Análise

Fonte: Criação própria

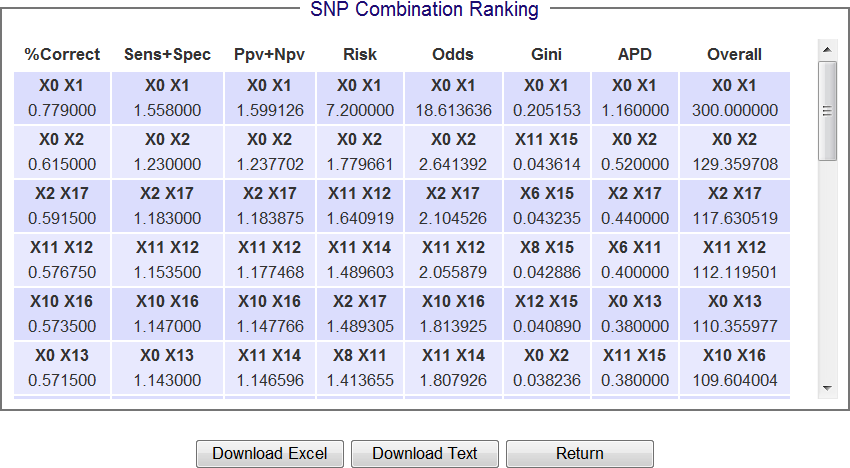
Tela típica de uma análise, neste caso com o método PIA. O primeiro parâmetro "Base Path" corresponde ao caminho onde está a base de dados no computador. Ao clicar no botão "Choose File" abre-se uma tela para que o arquivo que representa a base de dados seja escolhido. Os parâmetros "Order", "No. Top Results" e "Pathway Usage" representam, respectivamente, a ordem de interação, a quantidade de combinações que se pretende exibir e a opção de se realizar um pequeno teste com a informação das vias metabólicas nos quais os genes estão inseridos (caso as mesmas estejam contidas na base de dados, no caso entre o cabeçalho contendo o nome das variáveis e a primeira linha de dados). Essa opção consiste no cálculo de quanto cada via metabólica descrita contribui dadas as métricas finais [Mechanic et al. 2008]. O parâmetro "Output" corresponde à forma como os resultados serão exibidos: na tela, download dos resultados ou envio dos resultados por email. Nos últimos dois casos as opções de visualização são através de um arquivo excel ou de um arquivo de texto simples (mais informações nos próximos passos deste guia). O campo "Email" aparecerá apenas caso a opção de resultado por email seja marcada, o usuário deverá digitar o email para onde os resultados serão encaminhados. Ao clicar na barra "Advanced PIA Parameters" os parâmetros próprios deste método aparecerão (nesta figura este trecho já se encontra expandido). O usuário deverá configurar os parâmetros de acordo com sua preferência e então clicar no botão "Analyze" para iniciar a análise. Ao clicar em qualquer parâmetro geral ou próprio do método, será mostrada uma pequena tela descrevendo aquele parâmetro. Caso o usuário coloque um valor indevido no parâmetro também será exibida uma tela solicitando a mudança.



**Figura A.3.** Resultados na Tela (parte 1)

Fonte: Criação própria

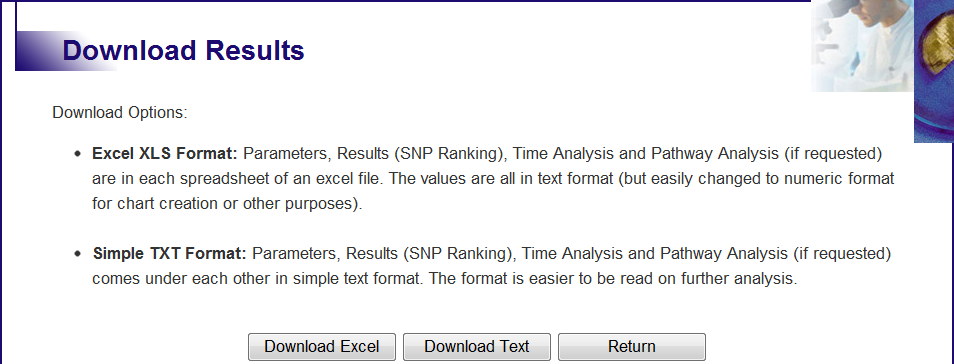
Caso a opção selecionada seja exibir os resultados na tela, na parte superior serão exibidas duas tabelas contendo os parâmetros utilizados e uma análise do tempo de execução em milissegundos.



**Figura A.4.** Resultados na Tela (parte 2)

Fonte: Criação própria

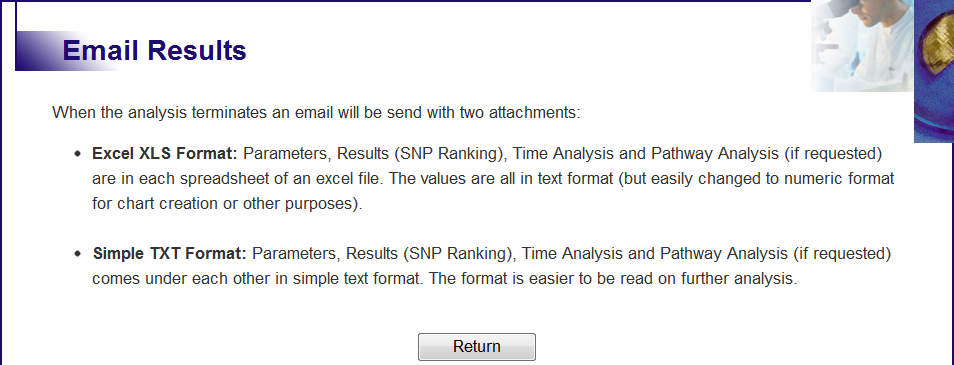
Na parte inferior são exibidos os resultados propriamente ditos, que consistem em uma lista contendo os polimorfismos para cada métrica que constitui o algoritmo com suas respectivas pontuações em ordem decrescente. Abaixo existem três botões: "Download Excel" permitindo ao usuário fazer o download no formato excel caso aquela análise o tenha interessado, "Download Text" para fazer o download no formato de texto e "Return" para retornar à tela de análise.



**Figura A.5.** Download dos Resultados

Fonte: Criação própria

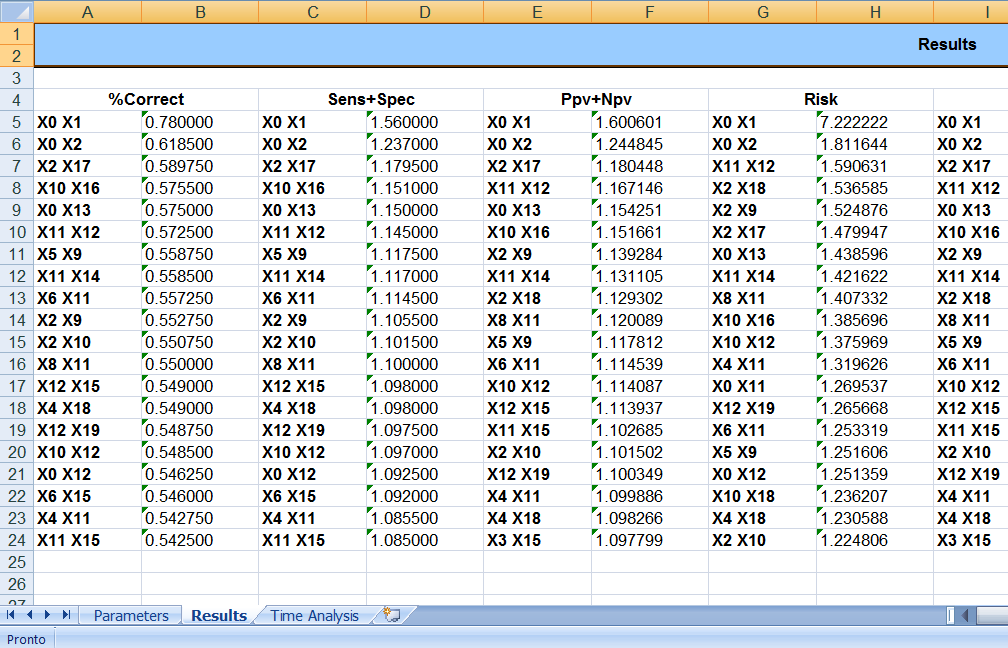
Caso a opção de retorno tenha sido a de realizar o download dos resultados, a tela exibida é esta. Os resultados não serão mostrados, existindo apenas os botões de download nos formatos excel e texto (como na tela exibida na Figura A.4) e de retorno para a análise anterior. São fornecidas informações sobre o conteúdo dos arquivos nos dois formatos possíveis para download.



**Figura A.6.** Resultados por Email

Fonte: Criação própria

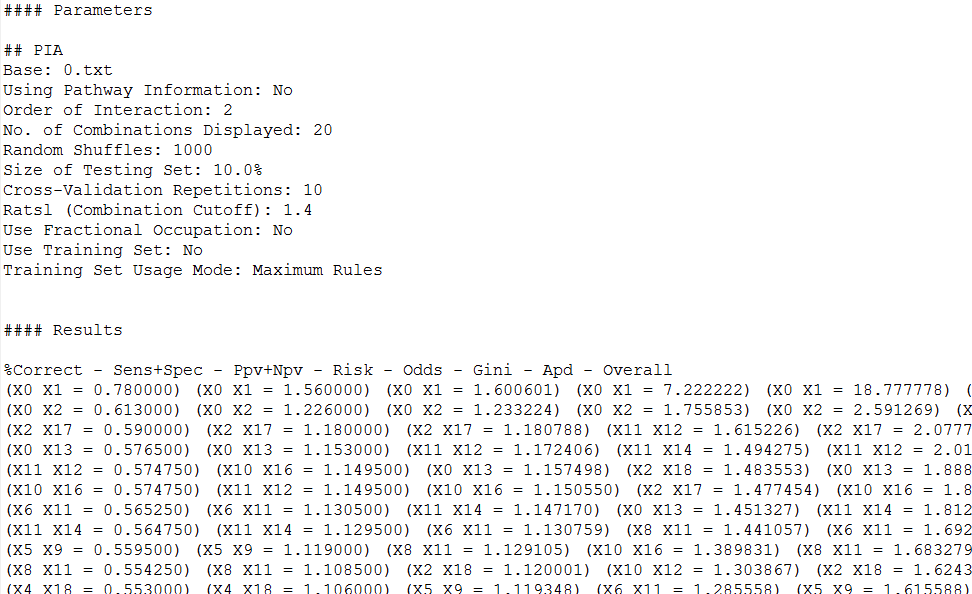
Caso a opção de retorno tenha sido a de receber os resultados por email, a tela exibida é esta. Os resultados não serão mostrados, sendo estes enviados para o email definido no campo "Email" da tela de análise. Serão enviados os resultados nos formatos excel e texto.



**Figura A.7.** Resultados no Formato Excel

Fonte: Criação própria

Esta figura mostra parte dos resultados no formato excel. O arquivo contém três planilhas: parâmetros, resultados e análise do tempo (caso a análise de via metabólica tenha sido selecionada também terá uma planilha adicional contendo tal análise).



**Figura A.8.** Resultados no Formato de Texto

Fonte: Criação própria

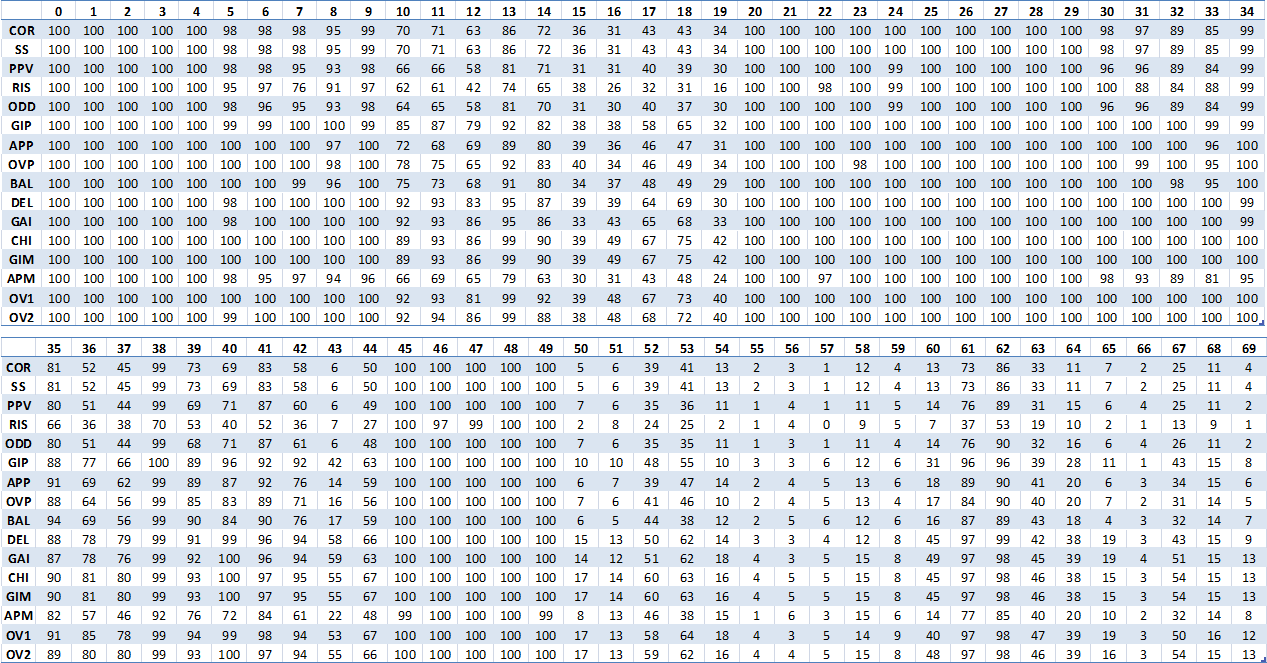
Esta figura mostra parte dos resultados no formato de texto. Este formato tem como objetivo a utilização destes resultados como entrada para algum outro algoritmo de análise por ser mais facilmente lido como texto.

Apêndice B - Demais Resultados

Esta seção exibe as taxas de acerto de todos os métodos aplicados a todos os modelos em todas as possíveis combinações de parâmetros que definem o tamanho da entrada. Cada tabela deste apêndice exibe as taxas de acerto para todos os modelos dada uma combinação entre número de SNPs e indivíduos. São 16 tabelas no total representando todas as combinações entre 20, 30, 40 e 50 SNPs com 200, 400, 800 e 1600 indivíduos.

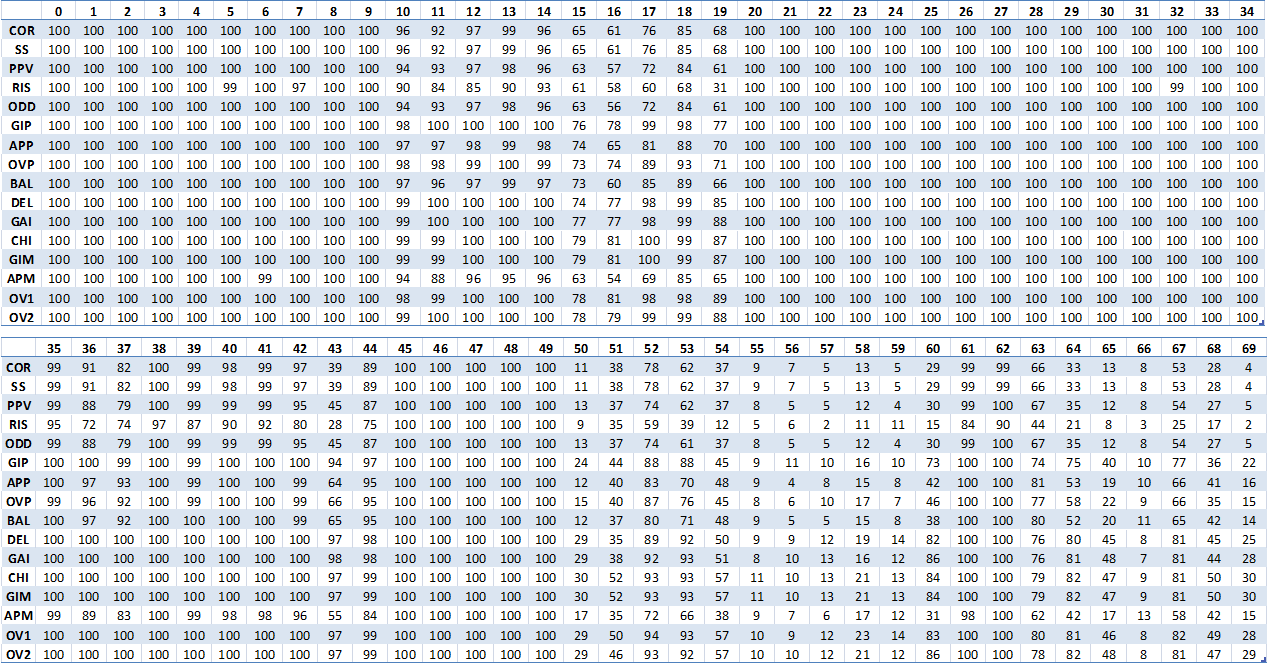
Para facilitar a visualização as métricas de cada método foram representadas por pequenas siglas, a legenda é a seguinte:

* COR = Taxa de Acerto (PIA)
* SS = Sensitividade + Especificidade (PIA)
* PPV = Precisão Positiva + Precisão Negativa (PIA)
* RIS = Razão de Risco (PIA)
* ODD = Razão de probabilidade (PIA)
* GIP = Índice Gini (PIA)
* APP = Diferença de Probabilidade Absoluta (PIA)
* OVP = Overall (PIA)
* BAL = Acurácia Balanceada (MDR)
* DEL = ΔR (ESNP2)
* GAI = Ganho de Informação (MASS)
* CHI = Chi-Quadrado (MASS)
* GIM = Índice Gini (MASS)
* APM = Diferença de Probabilidade Absoluta (MASS)
* OV1 = Overall com todas as métricas (MASS)
* OV2 = Overall com todas as métricas exceto APD (MASS)



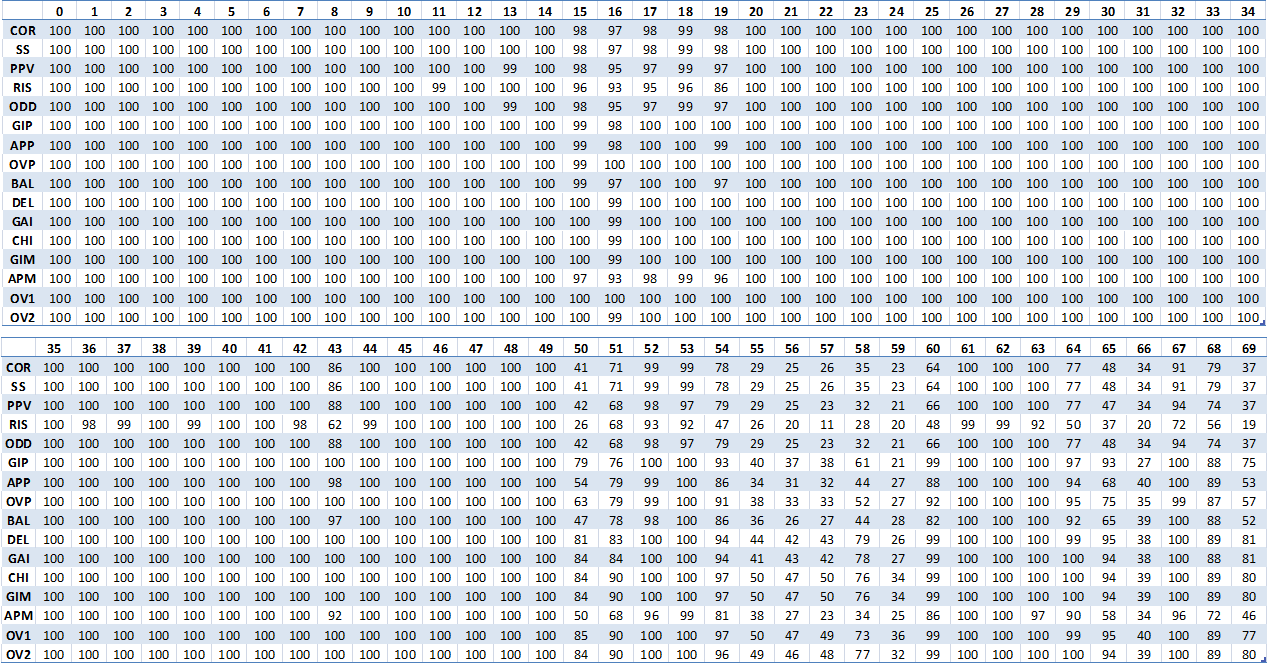
**Tabela B.1.** Taxas de Acerto (%) para 20 SNPs e 200 Indivíduos

Fonte: Criação própria



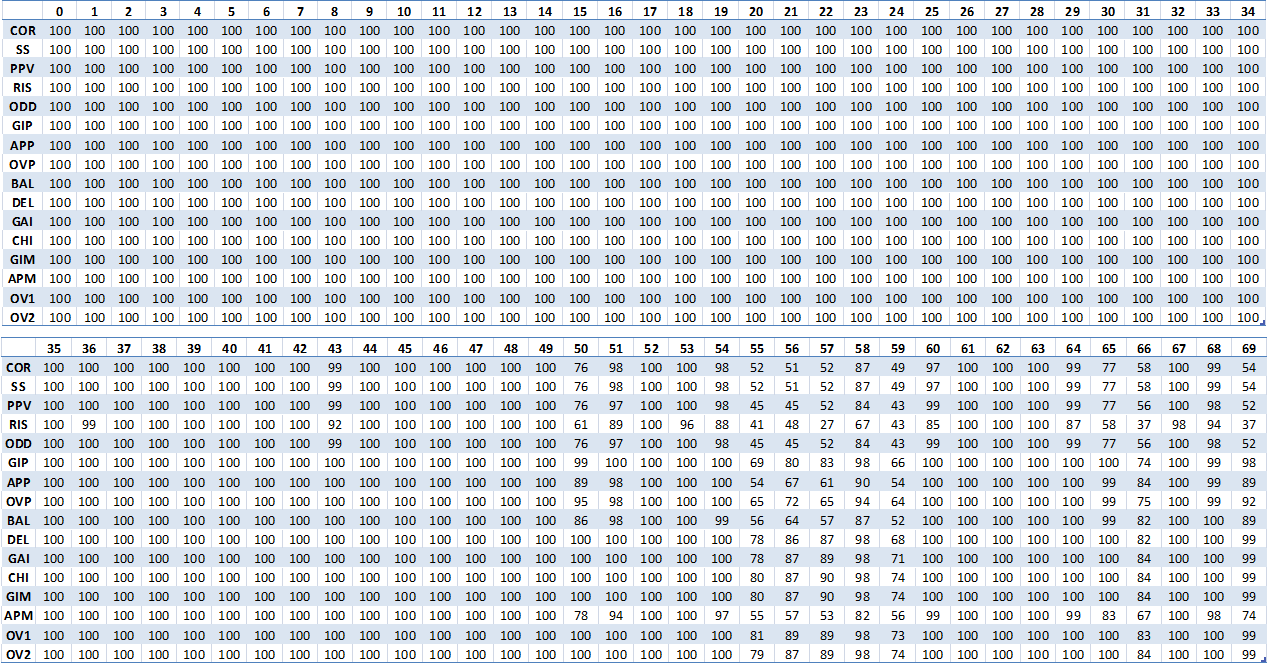
**Tabela B.2.** Taxas de Acerto (%) para 20 SNPs e 400 Indivíduos

Fonte: Criação própria



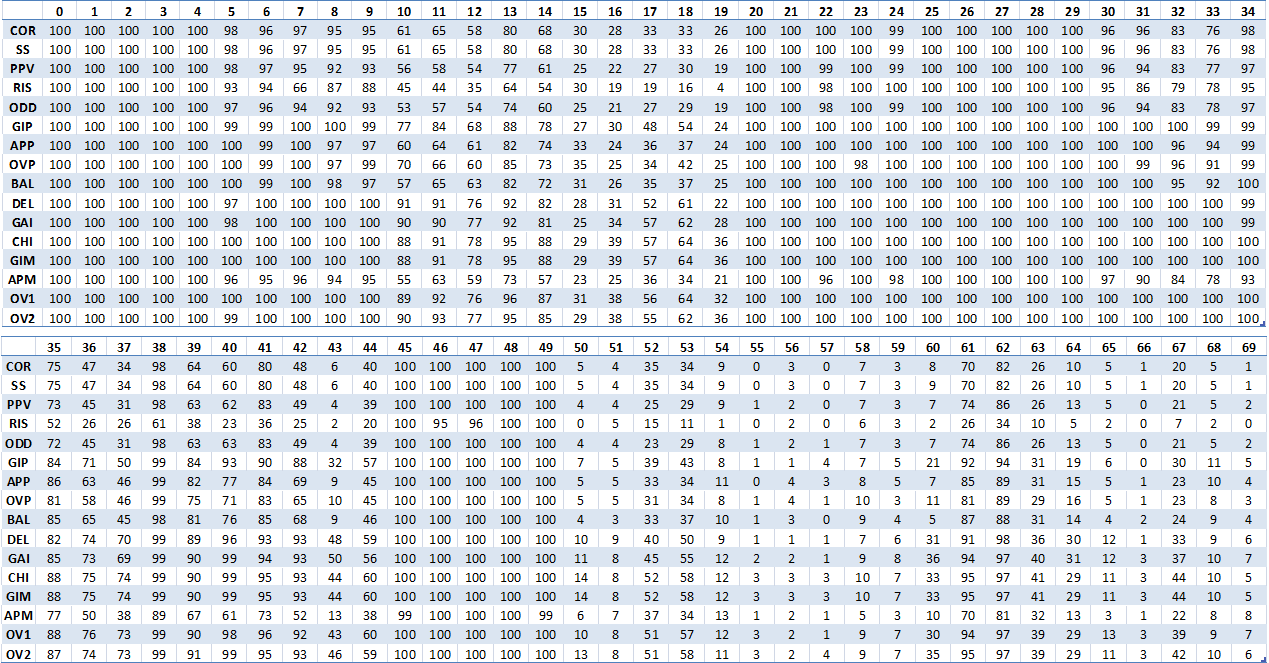
**Tabela B.3.** Taxas de Acerto (%) para 20 SNPs e 800 Indivíduos

Fonte: Criação própria



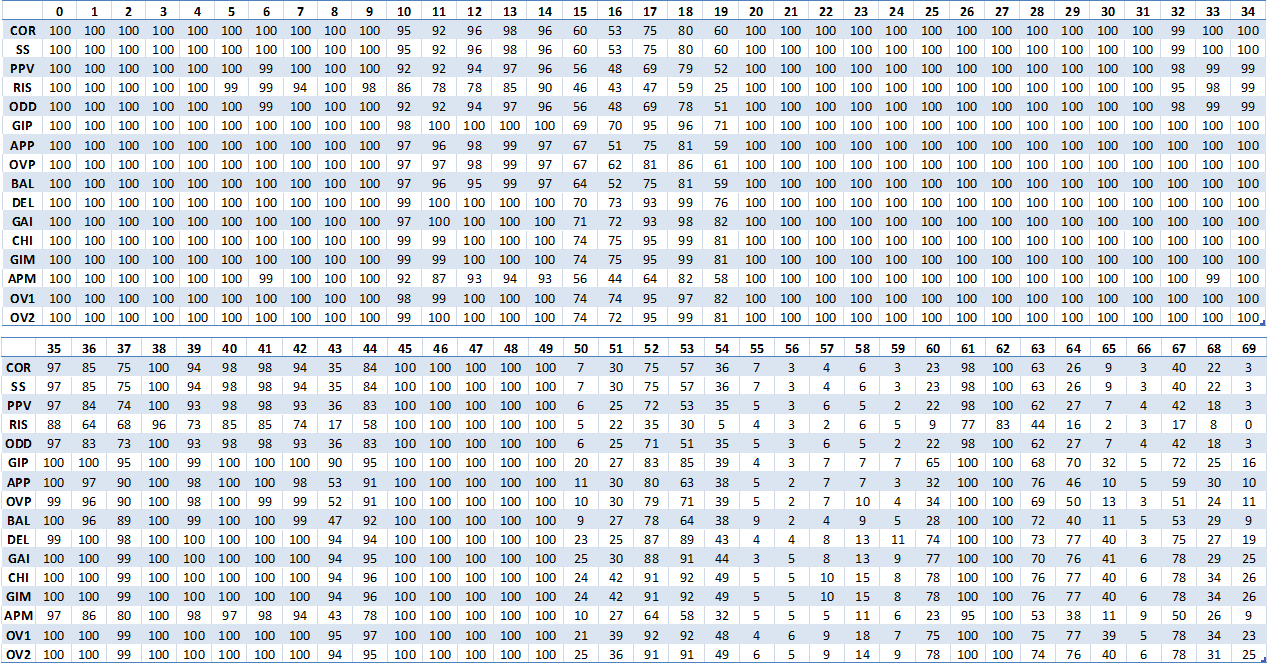
**Tabela B.4.** Taxas de Acerto (%) para 20 SNPs e 1600 Indivíduos

Fonte: Criação própria



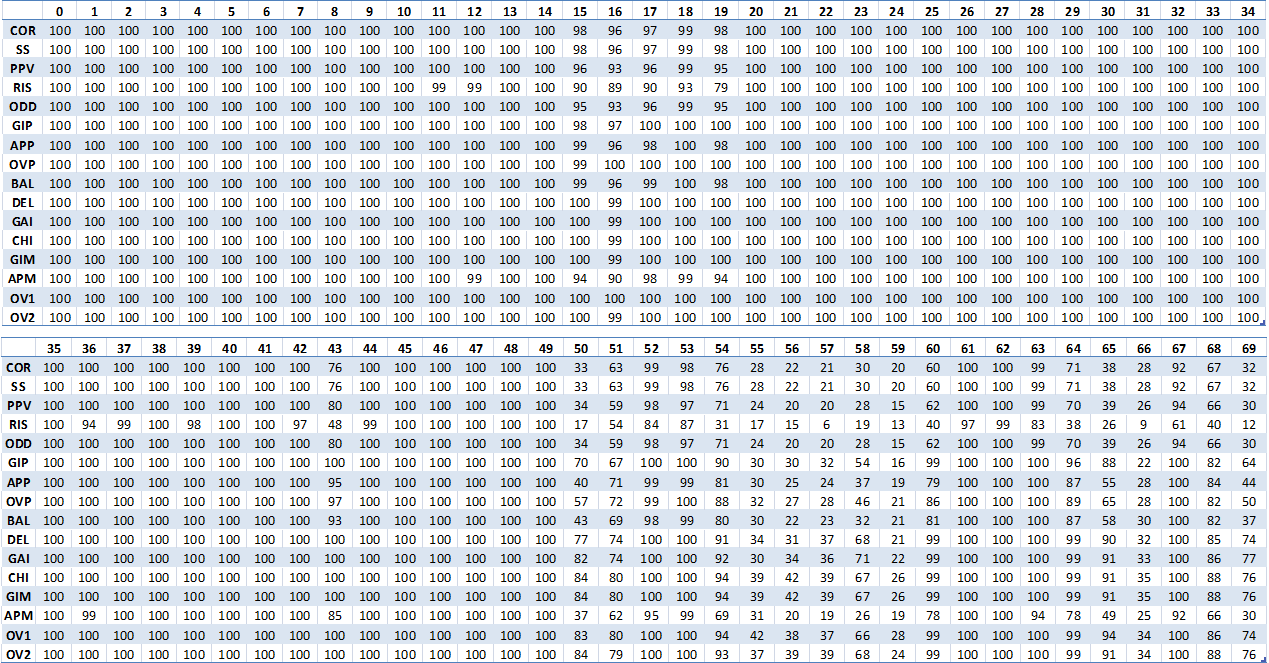
**Tabela B.5.** Taxas de Acerto (%) para 30 SNPs e 200 Indivíduos

Fonte: Criação própria



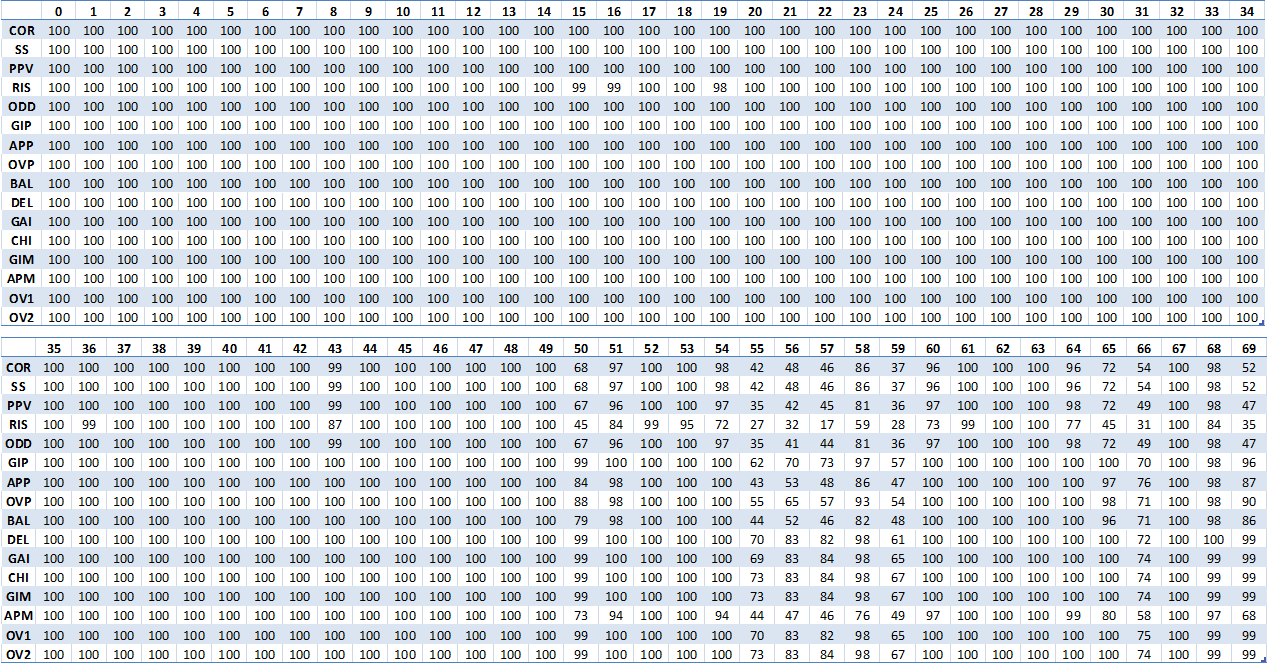
**Tabela B.6.** Taxas de Acerto (%) para 30 SNPs e 400 Indivíduos

Fonte: Criação própria



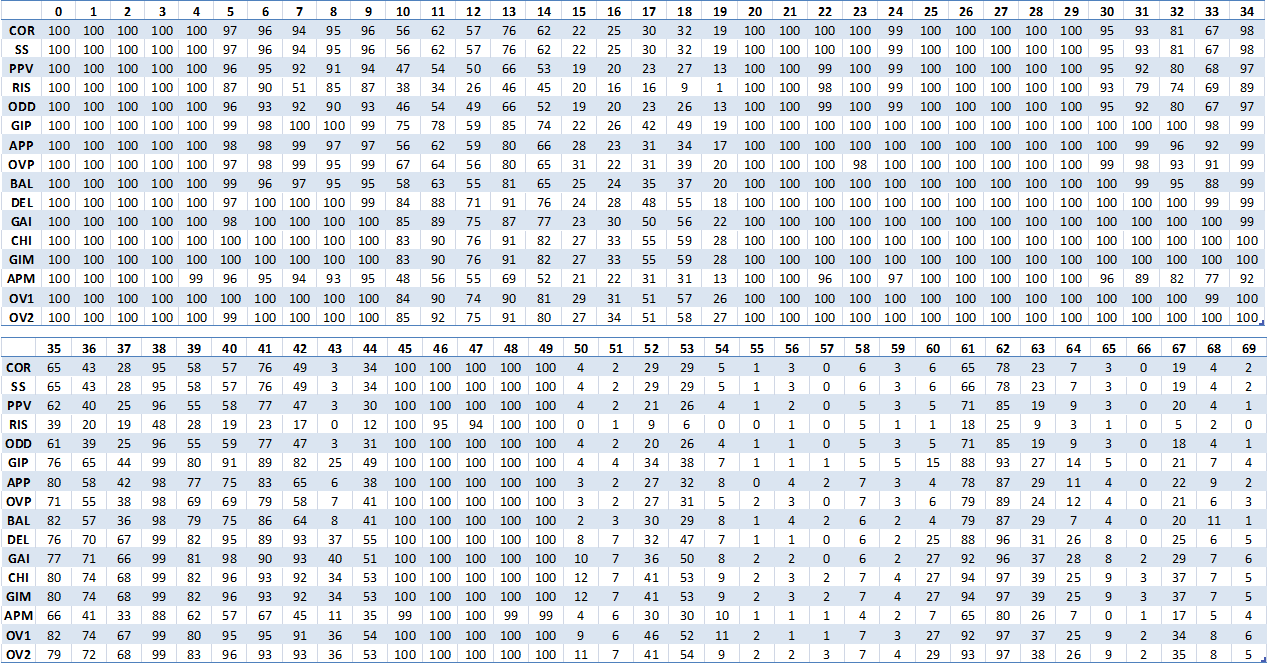
**Tabela B.7.** Taxas de Acerto (%) para 30 SNPs e 800 Indivíduos

Fonte: Criação própria



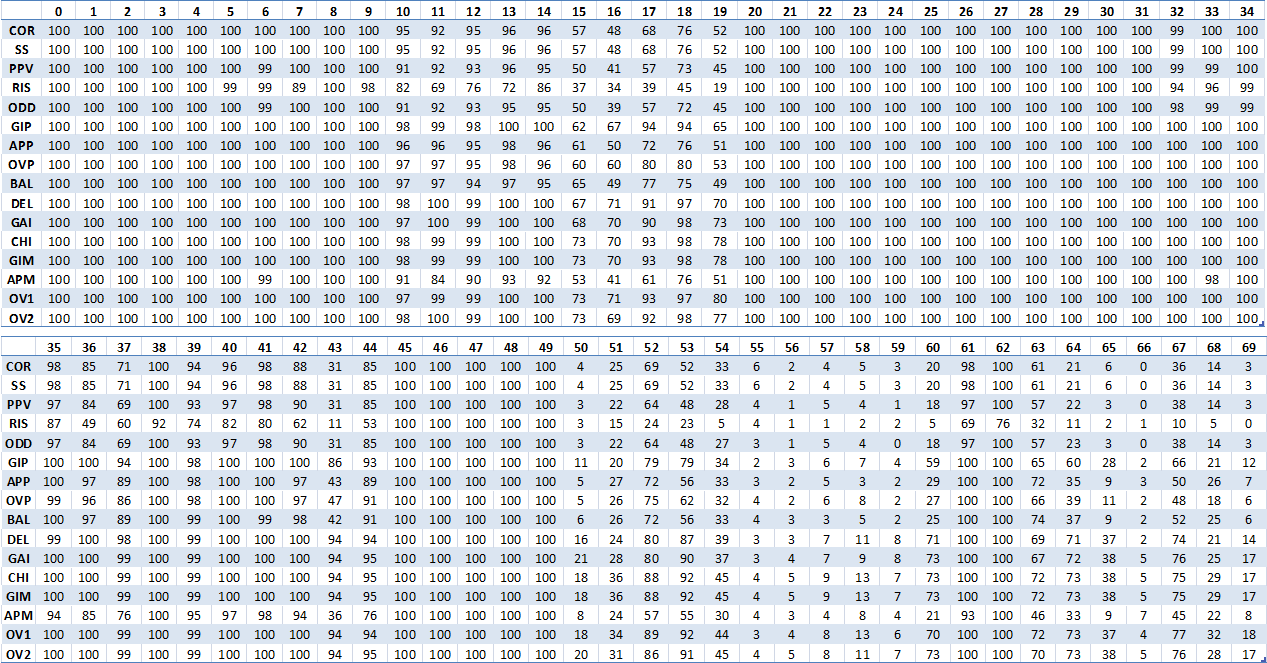
**Tabela B.8.** Taxas de Acerto (%) para 30 SNPs e 1600 Indivíduos

Fonte: Criação própria



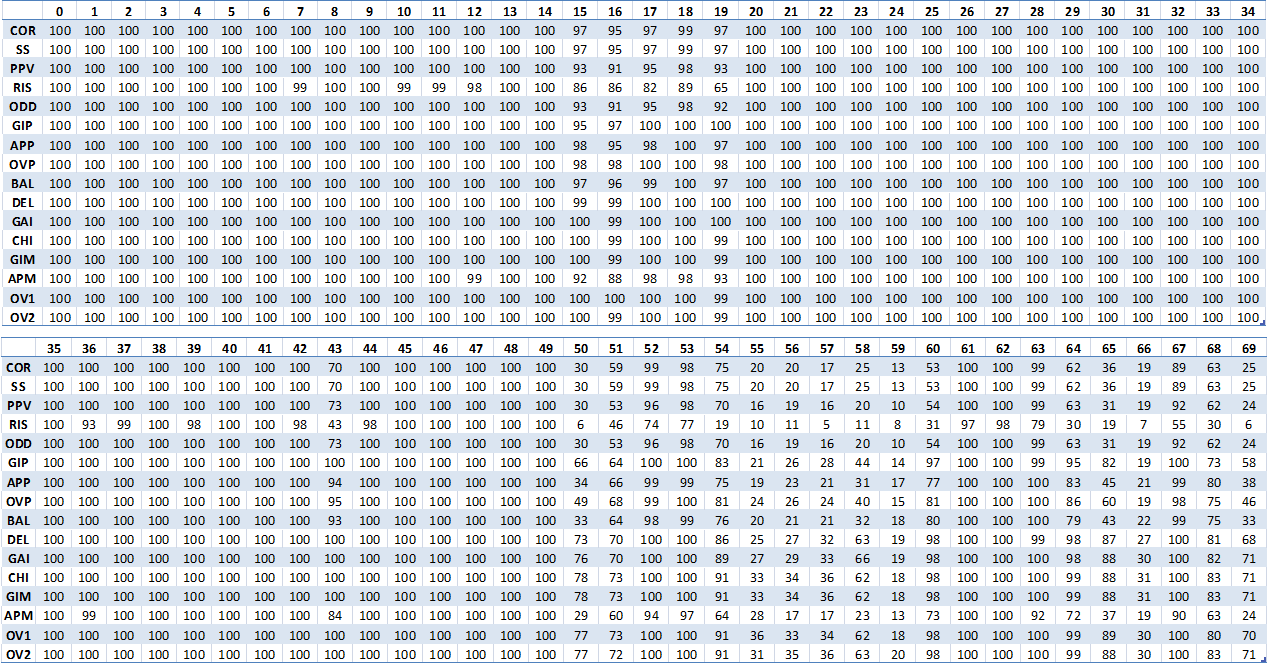
**Tabela B.9.** Taxas de Acerto (%) para 40 SNPs e 200 Indivíduos

Fonte: Criação própria



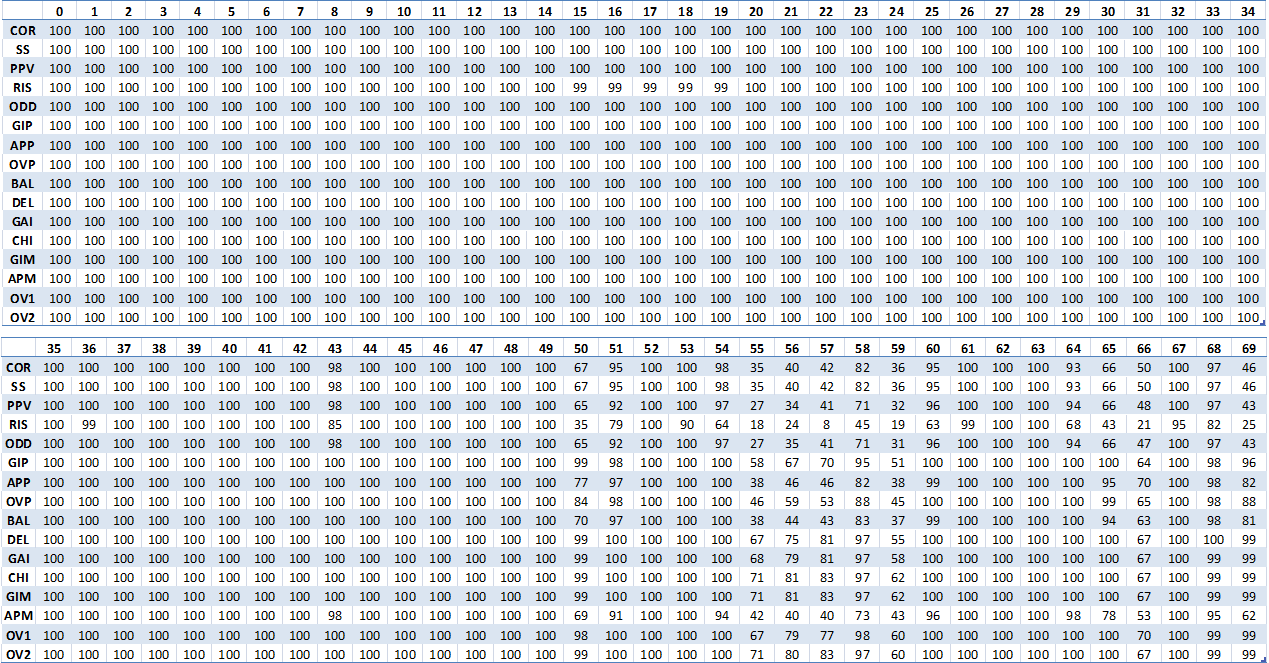
**Tabela B.10.** Taxas de Acerto (%) para 40 SNPs e 400 Indivíduos

Fonte: Criação própria



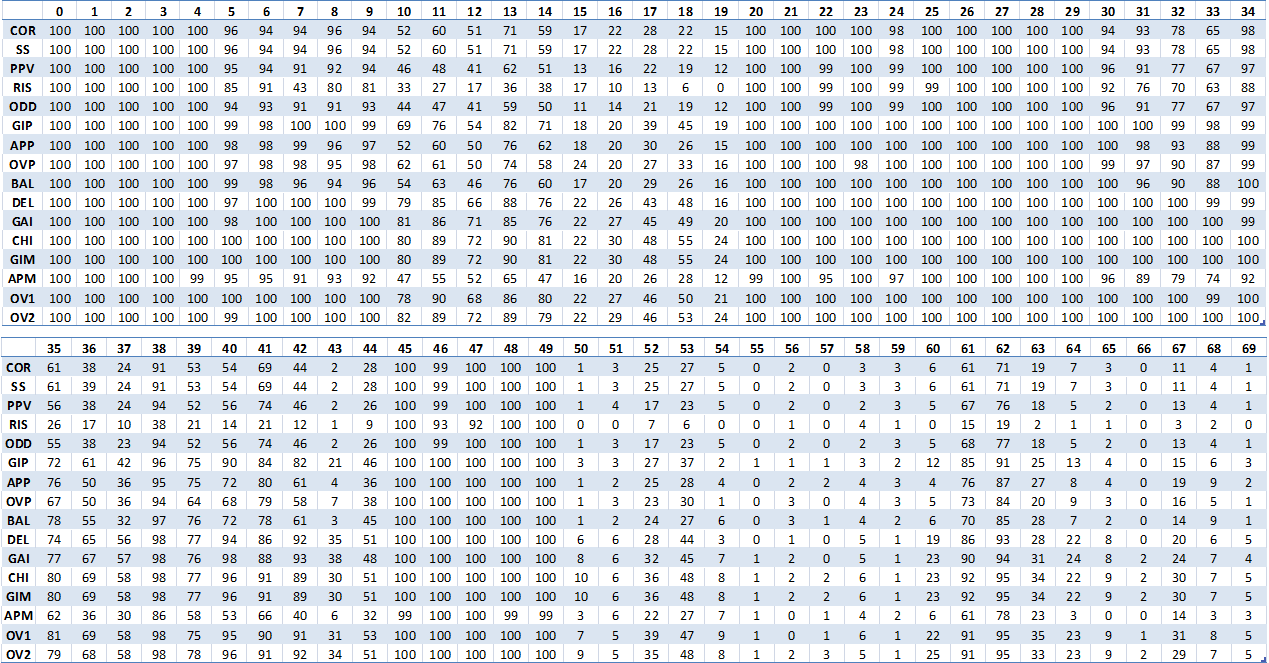
**Tabela B.11.** Taxas de Acerto (%) para 40 SNPs e 800 Indivíduos

Fonte: Criação própria



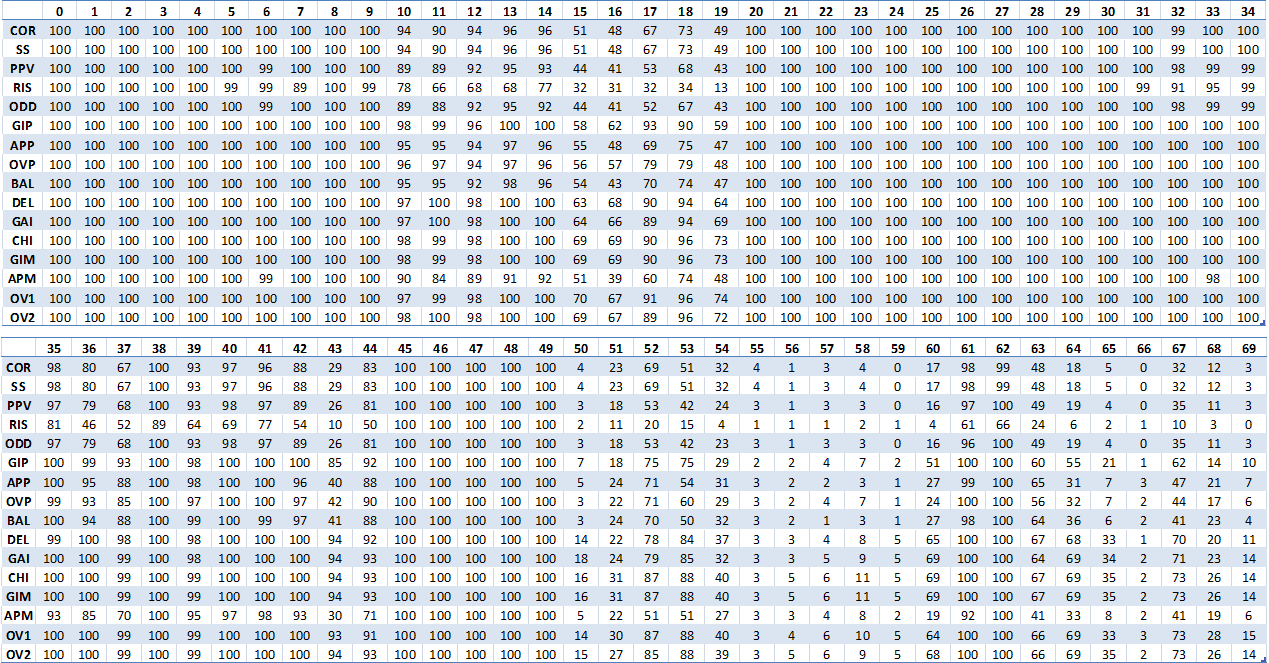
**Tabela B.12.** Taxas de Acerto (%) para 40 SNPs e 1600 Indivíduos

Fonte: Criação própria



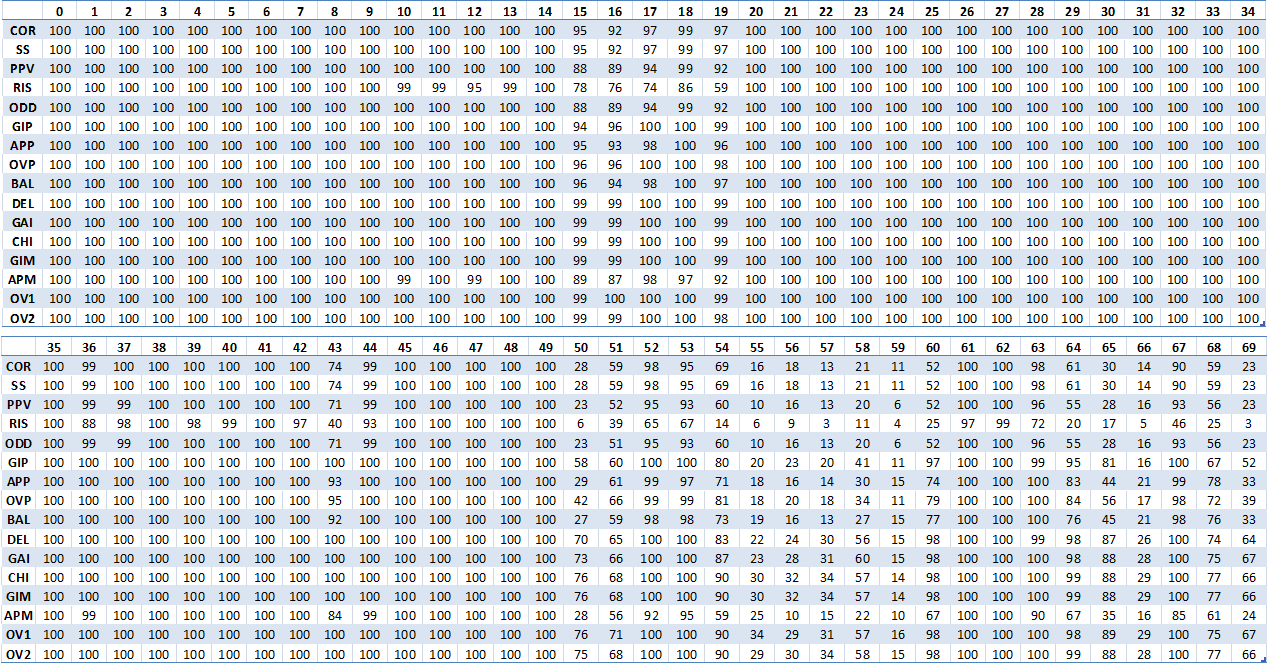
**Tabela B.13.** Taxas de Acerto (%) para 50 SNPs e 200 Indivíduos

Fonte: Criação própria



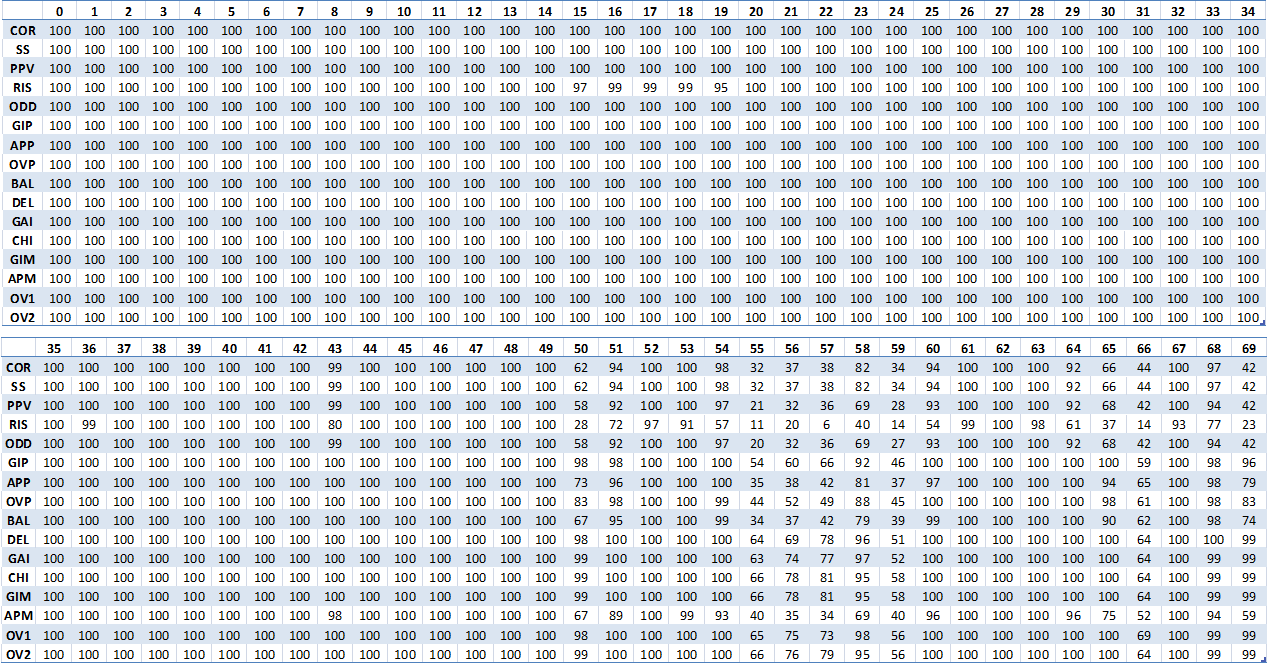
**Tabela B.14.** Taxas de Acerto (%) para 50 SNPs e 400 Indivíduos

Fonte: Criação própria



**Tabela B.15.** Taxas de Acerto (%) para 50 SNPs e 800 Indivíduos

Fonte: Criação própria



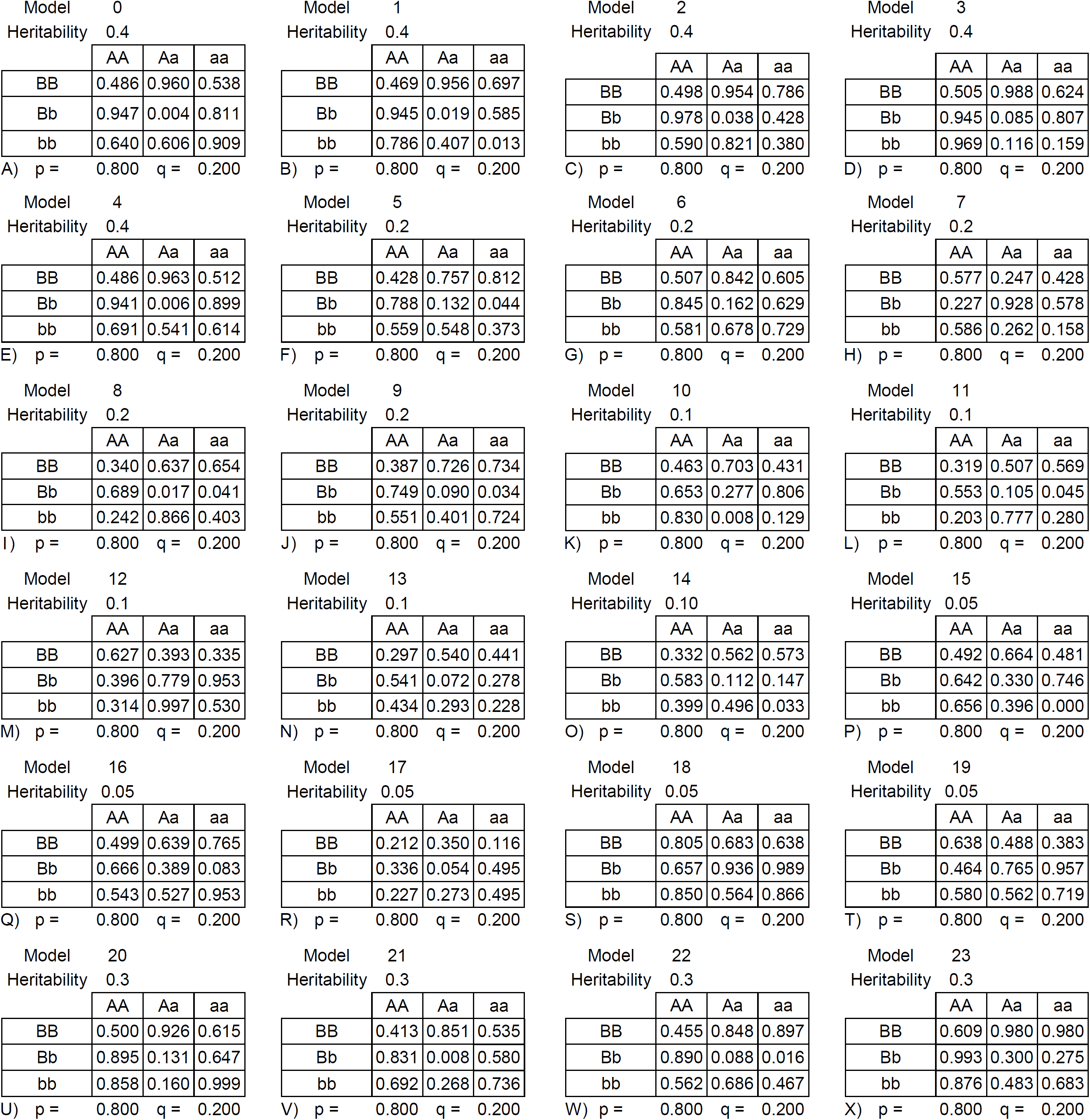
**Tabela B.16.** Taxas de Acerto (%) para 50 SNPs e 1600 Indivíduos

Fonte: Criação própria

Apêndice C - Modelos Epistáticos

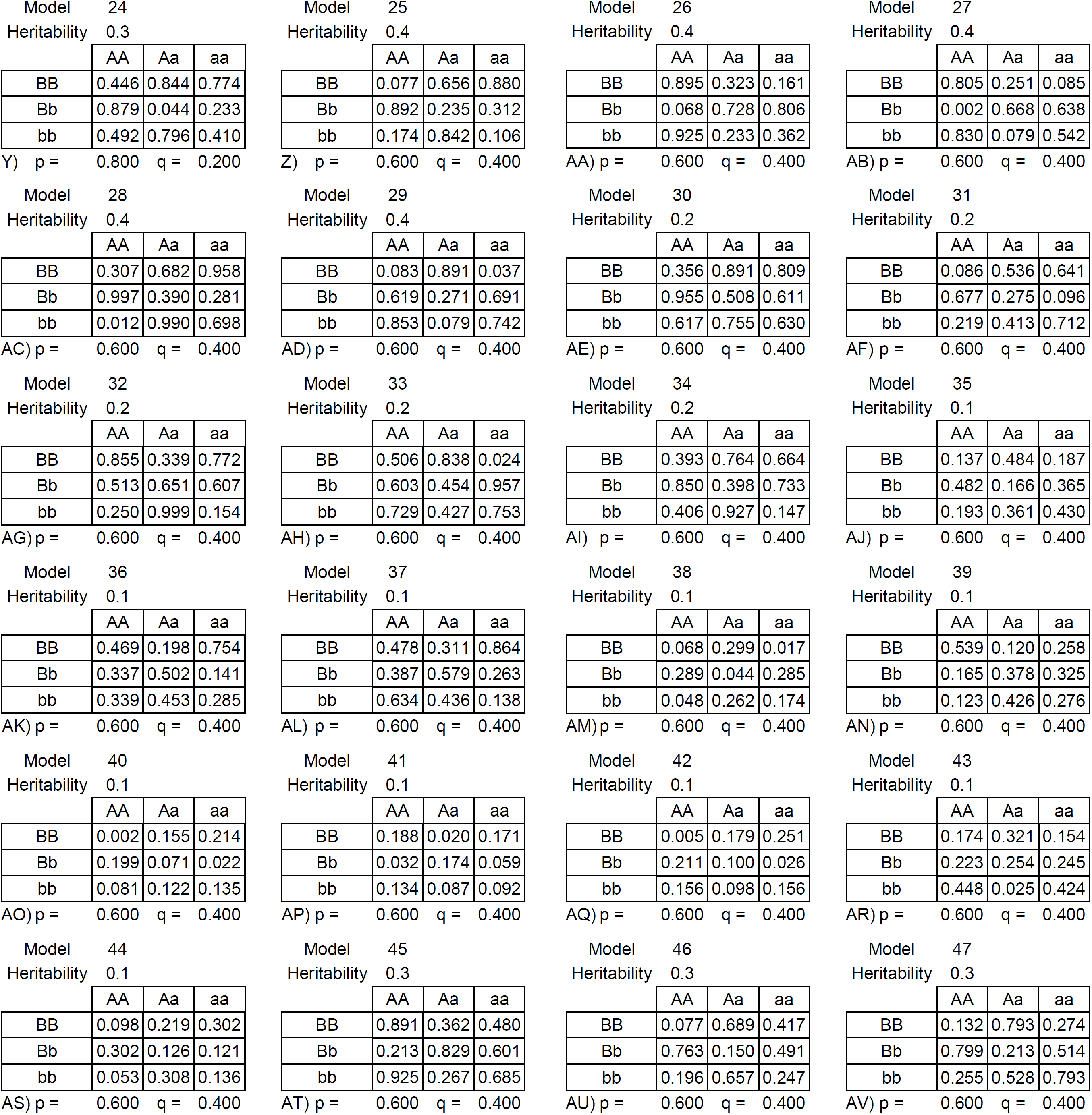
Nesta seção são exibidas as 70 tabelas de penetrância representando os 70 modelos epistáticos gerados no estudo [Velez et al. 2007]. As tabelas foram copiadas do material suplementar deste mesmo estudo e foram redimensionadas para facilitar a exibição.

As tabelas de penetrância foram utilizadas para gerar as bases de dados utilizadas nos experimentos deste trabalho. Os detalhes são discutidos no capítulo 4. As imagens começam na página seguinte pois tomam uma página inteira.



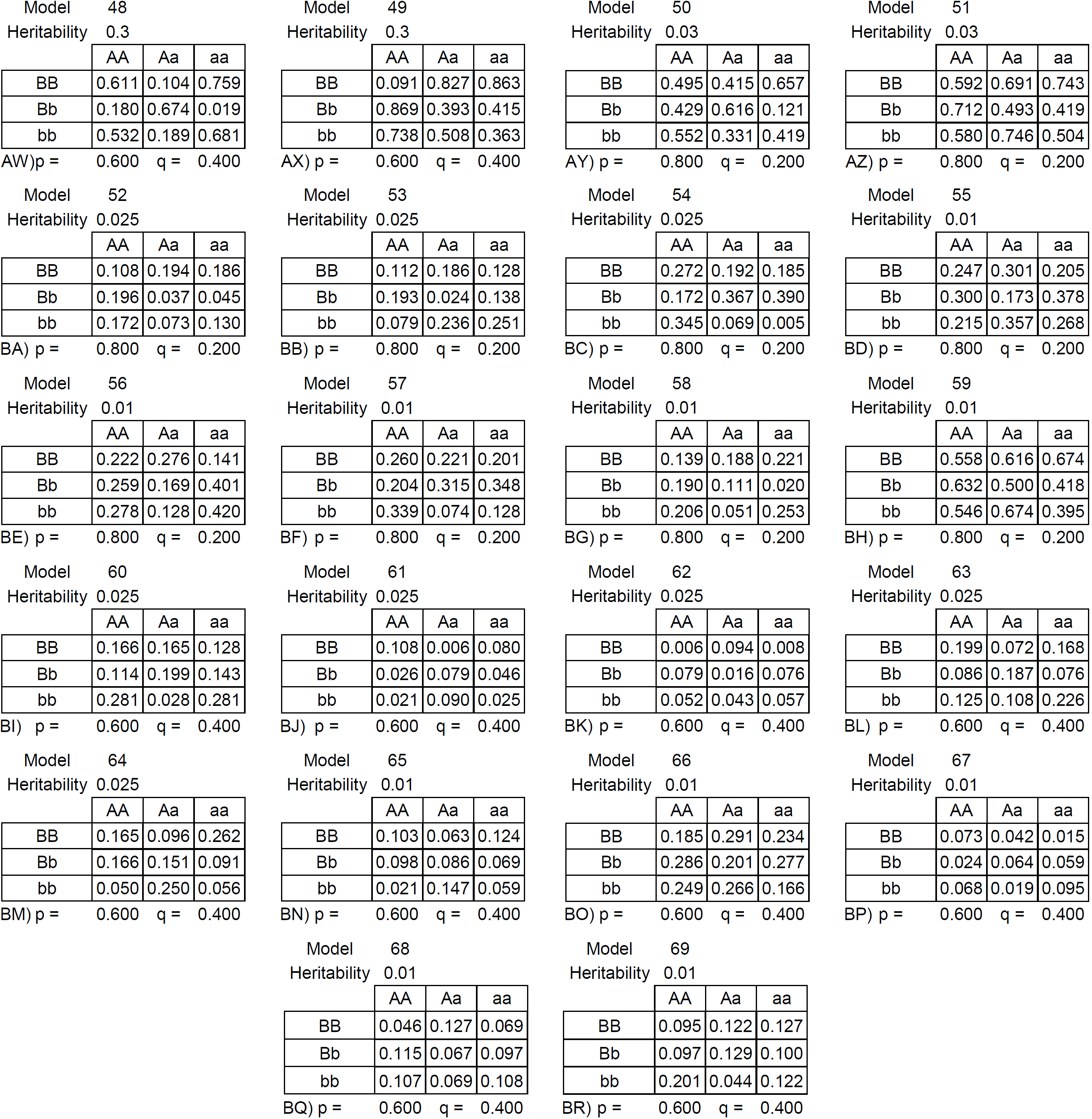
**Figura C.1.** Modelos Epistáticos 0 até 23

Fonte: [Velez et al. 2007]



**Figura C.2.** Modelos Epistáticos 24 até 47

Fonte: [Velez et al. 2007]



**Figura C.3.** Modelos Epistáticos 48 até 69

Fonte: [Velez et al. 2007]

1. Apesar da referência utilizada neste trabalho a respeito de árvores de decisão e árvores de classificação e regressão ser de 2001, estas estruturas, bem como sua teoria, foram propostas por L. Breiman e A. Cutler em 1984. [↑](#footnote-ref-1)
2. Nomeado, em inglês, de *environmentability* [↑](#footnote-ref-2)