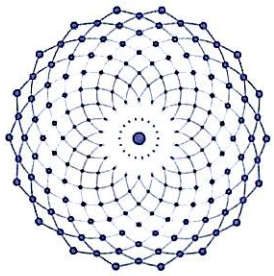

Dissertação de Mestrado

THÊMIS: Um sistema para análise forense de DNA utilizando Redes Bayesianas

José Tenório César Costa
tenoriocesar@gmail.com

Orientadores:

Profª. Drª. Eliana S. Almeida
Prof. Dr. Alejandro C. Frery

Maceió, Abril de 2009

Proc. 20040429671-7
DOCUMENTO RECEBIDO
Data: 24 / 07 / 2014
Mª Santos Barbosa

José Tenório César Costa

THÊMIS: Um sistema para análise forense de DNA utilizando Redes Bayesianas

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Curso de Mestrado em Modelagem Computacional de Conhecimento do Instituto de Computação da Universidade Federal de Alagoas.

Orientadores:

Profª. Drª. Eliana S. Almeida

Prof. Dr. Alejandro C. Frery

Maceió, Abril de 2009

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária Responsável: Helena Cristina Pimentel do Vale

C837t Costa, José Tenório César.
THÊMIS : um sistema para análise forense de DNA utilizando redes bayesianas / José Tenório César Costa, 2009.
124 f. : il.

Orientadora: Eliana Silva de Almeida.
Co-Orientador: Alejandro César Frery.
Dissertação (mestrado em Modelagem Computacional de Conhecimento) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2009.

Bibliografia: f. 120-124.

1. Bioinformática. 2. Genética forense. 3. DNA forense. 4. Modelagem computacional. 5. Redes bayesianas. I. Título.

CDU: 004.78:575.113.1

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Modelagem Computacional de Conhecimento pelo Programa Multidisciplinar de Pós-Graduação em Modelagem Computacional de Conhecimento, da Universidade Federal de Alagoas, aprovada pela comissão examinadora que abaixo assina:

Profa. Dra. Eliana Silva de Almeida
UFAL – Instituto de Computação
Orientadora

Prof. Dr. Alejandro César Frery
UFAL – Instituto de Computação
Co-orientador

Prof. Dr. Luiz Antônio Ferreira da Silva
UFAL – Instituto de Ciências Biológicas e da Saúde
Examinador

Prof. Dr. Maurício Marengoni
MACKENZIE – Faculdade de Computação e Informática
Examinador

Maceió, abril de 2009.

Resumo

Desde meados da década de 80, a tipagem do DNA (*DNA fingerprinting*) tem revolucionado a ciência forense, provendo uma poderosa ferramenta de investigação, sendo atualmente bastante utilizada em estudos de paternidade. Os laboratórios que trabalham com a análise forense de DNA realizam quantidades cada vez maiores de estudos desse tipo, incitando o uso de sistemas de software que auxiliem essa análise. Dentre as características essenciais para softwares dessa magnitude, está a confiabilidade, haja vista a minuciosidade do estudo. Dessa forma, é interessante o uso de métodos formais na execução de tais estudos. Neste trabalho, é construído um sistema de software, denominado THÊMIS, que utiliza o ferramental das Redes Bayesianas como meio de representação do conhecimento acerca de estudos de paternidade, obtendo por meio de inferências os resultados requeridos pela genética forense no que tange ao cálculo do Índice de Paternidade (IP).

Abstract

Since the mid 80, DNA fingerprinting has revolutionized forensic science, providing a powerful tool for research, currently being widely used in studies of paternity. Laboratories that work with forensic analysis of DNA carry increasing amounts of such studies and encourage the use of software systems that help with this type of analysis. One of the requirements for software of this magnitude is reliability, considering the level of detail of the study. Thus, it is interesting the use of formal methods. In this work, a software system called THÊMIS is built. THÊMIS uses Bayesian Networks as knowledge representation about studies of paternity, using inferences to obtain the results required by the forensic genetics regarding the calculation of the Index of Paternity (IP).

Agradecimentos

A Deus por tudo que me fora concedido, em especial por minha prodigiosa família e por minha maravilhosa noiva.

A meus queridos pais, César e Zelma, por terem me concebido a vida e por não terem medido esforços para tornar possível a concretização dos meus sonhos, bem como pelo amor e apoio incessantes.

A meus estimados irmãos, Tarcísio e Cynthia, por se fazerem presentes em todos os momentos de minha vida, apoiando-me e me repreendendo sempre que preciso.

Ao meu pequenino, porém gigante, sobrinho, Pedrinho, por toda alegria que tem proporcionado a mim e aos meus familiares, bem como pelo amor e carinho que nos tem dado.

A minha maravilhosa noiva, Arianna, pelo amor, carinho, compreensão, paciência e apoio ininterruptos, fazendo-se, pois, sempre presente em minha vida.

A todos os professores do Programa de Pós-Graduação em Modelagem Computacional de Conhecimento da Universidade Federal de Alagoas pela contribuição que têm dado a minha formação profissional, em especial a meus estimados orientadores, Profª. Eliana Silva de Almeida e Prof. Alejandro Frery, pela dedicação e paciência, pelos conselhos e incentivo e por não terem medido esforços no que tange à concretização desse sonho, acreditando em meu potencial e me incetivando sempre a superar-me.

Ao Prof. Luiz Antonio Ferreira da Silva e ao MSc. Dalmo Almeida de Azevedo do Laboratório de DNA Forense da Universidade Federal de Alagoas pelo suporte na validação do sistema *THÊMIS*.

À Fundação de Amparo à Pesquisa do Estado de Alagoas (FAPEAL) pelo apoio financeiro.

E, por fim, aos colegas de curso e às demais pessoas que participaram direta ou indiretamente na consolidação desse trabalho. Em especial, ao meu amigo João Roberto, pelo ótimo trabalho desenvolvido e pela imensa ajuda dada no que se refere a essa empreitada.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Pseudo-códigos

Capítulo 1

Introdução

A análise de dados biológicos pode ser relativamente complexa, haja vista que, em geral, a quantidade de dados a serem analisados é demasiadamente grande, bem como muitas dessas análises envolvem cálculos matemáticos e estatísticos, tornando, assim, a tarefa dos biólogos bastante árdua.

Os avanços da biologia molecular nas últimas décadas vêm favorecendo a geração de uma enorme quantidade de dados num tempo cada vez menor. Essa grande capacidade de geração de dados permite que os pesquisadores acelerem o ritmo de suas pesquisas, exigindo a utilização de estruturas mais robustas para o gerenciamento destes, bem como de ferramentas computacionais com capacidade de analisar e auxiliar na tarefa de dar um significado biológico a todos estes dados em tempo satisfatório (Setubal & Meidanis, 1997; Baldi & Brunak, 2001).

Buscando auxiliar os biólogos na análise forense de DNA, bem como dar continuidade aos estudos e às atividades desenvolvidas no Instituto de Computação da Universidade Federal de Alagoas na área de Bioinformática e no uso das Redes Bayesianas como uma abordagem estocástica para a incerteza, a presente dissertação visa à modelagem, implementação e validação de um sistema de software que utilize dados de DNA autossômico, também conhecido como nuclear, para obtenção dos resultados requeridos pela genética forense (Butler, 2005). Estes resultados dizem respeito à verificação de vínculo genético em estudos de paternidade.

Será mostrado que este sistema pode ser utilizado em qualquer laboratório de DNA Forense, haja vista sua confiabilidade. A validação desse software foi realizada no Laboratório de DNA Forense da Universidade Federal de Alagoas (`http://www.labdnaforense.org/`).

Nas seções e capítulos seguintes, serão apresentados de maneira objetiva os conceitos de maior relevância necessários ao entendimento da análise fo-

rense de DNA (no que tange à verificação de vínculo genético em estudos de paternidade), bem como das redes bayesianas (também conhecidas como redes de crença, redes probabilísticas, redes causais e mapas de conhecimento), que serão utilizadas na representação do conhecimento acerca de casos em que será verificada a paternidade de um indivíduo.

Por fim, serão apresentadas a modelagem, a implementação e a validação de um sistema de software, o sistema THÊMIS, que utiliza as redes bayesianas na obtenção dos resultados em estudos de paternidade.

1.1 Contexto

Por mais de dois mil anos a humanidade tenta desvendar os mistérios que há na inteligência. O sonho do desenvolvimento de sistemas computacionais inteligentes tem sido fonte de pesquisa desde meados dos anos 40, época em que surgiram os primeiros computadores (Marques & Dutra, n.d.).

A busca pela concretização desse sonho originou um novo campo de conhecimento denominado Inteligência Artificial (IA). Desde o seu surgimento, a IA tem focado seus estudos na busca por maneiras de se representar a inteligência humana por meio de computadores de diversos tipos, o que se pode ver claramente no seguinte texto de Luger (2004, p. 195):

> *A questão da representação, ou de como capturar, da melhor forma possível, os aspectos críticos da atividade inteligente para uso num computador, ou mesmo para a comunicação com os seres humanos, tem sido um tema constante ao longo dos cinqüenta anos da história da IA.*

A representação do conhecimento por meio de computadores é, na grande maioria dos casos, uma tarefa bastante complexa, uma vez que exige, além do conhecimento acurado do domínio a ser representado, o entendimento de diversos conceitos cuja compreensão é de fundamental importância para a construção de uma boa representação, ou seja, uma representação que retrate fielmente o domínio.

Há diversas abordagens que podem ser utilizadas para representar o conhecimento, dentre elas: métodos fracos para resolução de problemas, métodos fortes para resolução de problemas e métodos baseados em agentes (para maiores detalhes, ver Luger, 2004). A escolha da abordagem, por estar intimamente relacionada ao domínio a ser representado, é uma das etapas mais importantes na modelagem de um sistema para representação do conheci-

mento, podendo, pois, em caso de uma escolha "equivocada", comprometer todo o sistema.

Assim como as abordagens para representação, os domínios são muito diversos, uma vez que representam aspectos do mundo real, o qual é, por natureza, infinito. Devido à infinidade de domínios existentes, os quais possuem muitas características peculiares, deve haver grande cautela na escolha da abordagem a ser utilizada para representá-los.

Nos domínios gerados na análise forense de DNA em estudos de paternidade, há um certo grau de incerteza, haja vista que apesar de cada ser humano possuir um perfil genético único (como será discutido mais adiante), em estudos forenses é impossível se analisar todo o genótipo dos indivíduos envolvidos, havendo, pois, a necessidade de se utilizar inferências a partir da análise de uma amostra.

Uma abordagem geral para tratar a falta de informação (incerteza) é por meio de cálculos probabilísticos, uma vez que "a probabilidade proporciona um meio para resumir a incerteza" (Russell & Norvig, 2004, p. 451), permitindo, dessa forma, a obtenção de conclusões úteis, a partir de dados incompletos e imprecisos. Nesse tipo de abordagem, as informações *a posteriori* são obtidas por inferências a partir das informações *a priori* e das evidências disponíveis (ver capítulo 2).

1.2 Advento das Redes Bayesianas

Pode-se observar a evolução no tratamento da incerteza em sistemas computacionais fazendo uma breve analogia entre as abordagens utilizadas nas décadas de 60 a 80, as quais são explanadas sucitamente a seguir e discutidas com maiores detalhes em Flores et al. (2000).

A partir dos anos 60, foram construídos os primeiros sistemas computacionais de apoio à decisão que eram, em sua maioria, voltados para problemas de diagnóstico. Esses sistemas tratavam a incerteza de forma restritiva devido à falta de interesse no uso da teoria das probabilidades, vista como intratável em termos computacionais, obtendo, pois, em geral, resultados matematicamente incorretos em domínios maiores.

Nos anos 70, surgem os sistemas especialistas (SE) que utilizavam, para manipular incerteza, métodos *ad hoc*, os quais associavam fatores de certeza às regras constituintes da base de conhecimento como forma de tratamento para a incerteza. Por poderem manipular domínios maiores, esses sistemas são mais robustos e mais complexos que os anteriores, porém podem produzir

resultados inesperados.

No final dos anos 80, o interesse por abordagens que utilizavam a teoria das probabilidades para lidar com incerteza ganhou foco novamente, estando essa retomada intimamente ligada ao surgimento das redes probabilísticas — modelos que representam graficamente as dependências probabilísticas entre os objetos do domínio. Dentre as vantagens dessa nova abordagem, pode-se citar:

- representação e manipulação da incerteza baseadas em modelos matemáticos;
- modelagem do conhecimento de forma intuitiva acerca do domínio;
- permissão à realização de inferência causal, de diagnóstico, intercausal ou mista.

A utilização de modelos gráficos —grafos cujos nós são variáveis aleatórias e cujos arcos representam as relações entres essas variáveis— é uma abordagem interessante para representar domínios incertos com alta complexidade, uma vez que estes modelos unem a teoria dos grafos, que permite descrever graficamente as relações entre as variáveis, e a teoria das probabilidades, que atribui níveis de crença às variáveis.

Dentre os modelos gráficos, estão as redes bayesianas que são um formalismo poderoso para representar e raciocinar sob circunstâncias de incerteza (Cheng et al., 2002, p. 43).

Desde a década de 90, época em que teve início a maior parte das pesquisas em redes bayesianas, essa rede probabilística vem sendo utilizada para a solução de vários tipos de problemas pertinentes às mais diversas áreas, como por exemplo:

- classificação de tumores ovarianos (ver Antal et al., 2003);
- análise da seqüência de terremotos que ocorreram numa região (ver Agostinelli & Rotondi, 2003);
- investigação das causas que propiciam a queda na qualidade da água em sistemas de tratamento (ver Pike, 2004);
- previsão da taxa (quantidade) de macroinvertebrados em rios (ver Adriaenssens et al., 2004);
- previsão da confiabilidade de sistemas de software (ver Bai, 2005);

- previsão de manutenabilidade para sistema de software orientado a objetos (ver Koten & Gray, 2006);
- modelagem dos principais fatores envolvidos no processo de produção de software (ver Vieira et al., 2006);
- tratamento da incerteza existente na análise de dados biológicos (Egeland et al., 2000; Dawid et al., 2002; Cowell, 2003; Pena, 2006; Almudevar, 2007; Santos Júnior et al., 2008, 2009).

1.3 Advento da Análise Forense de DNA

A recombinação gênica provê um alto grau de variabilidade entre os seres humanos, possuindo cada indivíduo um perfil genético (genótipo) único, exceto os gêmeos monozigóticos.

Antigamente, a ciência forense utilizava-se apenas das análises sorológicas dos polimorfismos de grupos sangüíneos e proteínas e de alguns marcadores genéticos.

Por volta do início do século XX, teve início o exame forense de amostras biológicas, o qual fazia uso dos grupos sangüíneos ABO em evidências relacionadas à identificação de pessoas e a crimes. Todavia, atualmente os grupos sangüíneos eritricitários, como os sistemas ABO e Rh, foram substituídos por outras técnicas de análise forense, em especial a que utiliza como objeto de estudo o DNA. Dentre as vantagens no uso do DNA sobre a sorologia tradicional, pode-se citar: a possibilidade de sua aplicação sobre toda e qualquer fonte de material biológico, o seu potencial discriminatório, a sua resistência aos fatores ambientais, dentre outras (ver Henry, 2008).

Correspondendo à segunda fase na evolução da ciência forense, em 1954, foi demonstrada, na superfície dos leucócitos, a ocorrência de um sistema de histocompatibilidade mediado por antígenos, denominado complexo *Histocompatibility Leucocyte Antigen* (HLA), determinado por genes alélicos muito próximos localizados no braço curto do cromossomo 6, com grande poder de discriminação individual ou determinação da individualidade genética (Calabrez, 1999).

Desde meados da década de 80, a tipagem do DNA (*DNA fingerprinting*) tem revolucionado a ciência forense, provendo uma poderosa ferramenta de investigação no que tange à identificação de criminosos (Bianchi & Lio, 2007), à análise da paternidade de uma dada pessoa (Egeland et al., 2000; Santos Júnior et al., 2009), ao reconhecimento de corpos em desastres de grande

escala (Alonso et al., 2005; Leclair et al., 2004, 2007), dentre outros. Milhares de casos têm sido solucionados graças ao poder da genética forense.

A tipagem do DNA foi primeiro descrita em 1985 pelo geneticista inglês Alec Jeffreys, o qual descobriu que certas regiões do DNA contendo seqüências de DNA eram repetidas e o número de repetições numa amostra diferia de um indivíduo para o outro. Surge, assim, a terceira fase do desenvolvimento das ciências forenses voltadas à identificação humana.

Por desenvolver uma técnica para examinar a variação dessas seqüências repetidas de DNA, que produziam uma espécie de "impressão digital" do *DNA*, daí o termo *DNA fingerprinting*, o Dr. Jeffreys desenvolveu a habilidade de realizar teste de identidade em humanos.

Essas regiões de DNA repitido foram chamadas de *Variable Number of Tandem Repeats* (VNTRs) e eram analisadas por uma técnica conhecida por *Restriction Fragment Length Polymorphism* (RFLP), pois usava uma enzima de restrição para cortar as regiões do DNA que envolviam as VNTRs (Butler, 2005). O RFLP foi o primeiro método usado para a resolução de um problema de imigração na Inglaterra (Jeffreys et al., 1985). No ano seguinte, esta mesma técnica foi utilizada para solucionar um caso de duplo homicídio (maiores detalhes em http://www.forensic.gov.uk/html/media/case-studies/f-18.html acessado pela última vez em 25-02-2009).

Desde então, testes de identificação humana usando métodos de tipagem de DNA vêm sendo bastante difundidos e utilizados, o que se pode ver claramente em Leclair et al. (2004), Alonso et al. (2005), Bianchi & Lio (2007) e Bandyopadhyay et al. (2008).

Atualmente, mais de 150 laboratórios forenses públicos e dezenas de laboratórios de teste de paternidade particulares conduzem centenas de milhares de testes de DNA anualmente nos Estados Unidos. Além disso, muitos países europeus e asiáticos têm programas de DNA forense e o número de laboratórios desse tipo vem aumentando consideravelmente em todo mundo (Butler, 2005).

No Brasil, foi somente em 1994 que se criou o primeiro laboratório para análise forense de DNA. Tratava-se da *Divisão de Pesquisa de DNA Forense da Polícia Civil do Distrito Federal.* A partir daí muitos esforços têm sido feitos buscando a criação de novos centros para análise de material genético com rígidos padrões de qualidade, haja vista a minuciosidade desse tipo de estudo.

Maiores detalhes a respeito da genética forense no Brasil podem ser obtidos em Smarra et al. (2006).

1.4 Justificativas

Dentre os fatores preponderantes na escolha do tema desta dissertação, estão:

- a complexidade que há em se representar o conhecimento por meio de sistemas computacionais, em especial quando se trata de um domínio com incertezas, como é o caso dos domínios em que são realizados os estudos de paternidade, uma vez que apesar de cada ser humano possuir um perfil genético único, em estudos forenses é impossível se analisar todo o genótipo dos indivíduos envolvidos, havendo, pois, a necessidade de se utilizar inferências na análise dos dados;

- a imprecisão que há nos fatos reais, o que se pode ver claramente no seguinte texto de Albert Einstein:

  > *Na medida em que as leis matemáticas se referem à realidade, elas não são certas. E na medida em que elas são certas, elas não se referem à realidade. (apud Luger, 2004, p. 291)*

- a aplicação prática das redes bayesianas na resolução de uma infinidade de problemas do mundo contemporâneo, dentre os quais a análise de vínculo genético no que tange a estudos de paternidade (ver seção 1.2);

- a inexistência no Brasil de softwares tal qual o `familias` (ver Egeland et al., 2000) que utiliza o ferramental das redes bayesianas como meio de representação do conhecimento acerca de estudos de paternidade, obtendo por meio de inferências os resultados requeridos pela genética forense no que tange ao cálculo do Índice de Paternidade (IP). Sendo importante mencionar que nesse software, é necessário o usuário construir a rede que representa o estudo em questão, exigindo, pois, desse indivíduo um certo grau de conhecimento sobre as redes bayesianas. Além disso, o `familias` não possui uma base de dados para o armazenamento dos perfis genéticos, os quais, devido a isso, não poderão ser utilizados em estudos futuros.

1.5 Materiais

Os materiais necessários ao pleno desenvolvimento do presente trabalho, haja vista o seu cunho teórico e sua aplicabilidade prática, foram artigos científicos e demais publicações científicas, em especial aquelas que foram publicadas em revistas renomadas da área, livros e ferramentas computacionais

que corroboraram o conhecimento adquirido e contribuíram para a aquisição de novos conhecimentos. Dentre os softwares analisados que utilizam redes bayesianas como abordagem para representação do conhecimento (ver Apêndice A), o escolhido como motor de inferência na análise dos dados biológicos no que tange à verificação de vínculo genético em estudos de paternidade foi o `UnBBayes`, visto que este software se encontra disponível sob licença *GNU General Public License (GPL)*, possui *Application Programming Interface (API)*, bem como interface gráfica bastante simples e intuitiva, eficiência no processo de inferência e confiabilidade.

1.6 Objetivos e Contribuições

Dentre os objetivos e contribuições do trabalho estão:

- aquisição de um conhecimento acurado em redes bayesianas, mais especificamento no uso desse formalismo na análise forense de DNA;
- entendimento e avaliação de algumas plataformas computacionais disponíveis para o uso desse modelo (ver Apêndice A);
- modelagem, implementação e validação de um sistema de software que utilize o formalismo das redes bayesianas na análise de dados biológicos, armazenando numa base de dados as informações sobre os estudos e não exigindo do usuário a construção da rede que representa a genealogia em questão. Dessa forma, os dados inseridos no sistema podem ser utilizados em análises futuras e o usuário não necessita ter conhecimento algum sobre redes bayesianas, ou seja, o uso desse formalismo fica transparente ao usuário;
- concessão do referido sistema para uso no Laboratório de DNA Forense da Universidade Federal de Alagoas (`http://www.labdnaforense.org/`). Este laboratório realiza grande quantidade de estudos de paternidade e, para cada estudo realizado, os dados obtidos após a execução da tipagem de DNA são analisados com o auxílio de planilhas eletrônicas, nas quais são construídas funções para este fim.

1.7 Estrutura

O texto está estruturado da seguinte forma:

Capítulo 2 Aborda os conceitos da teoria das probabilidades necessários ao entendimento das redes de crença.

Capítulo 3 Apresenta os conceitos de maior relevância sobre redes bayesianas (topologia, inferência etc.).

Capítulo 4 Apresenta os conceitos de maior relevância sobre análise forense de DNA autossômico na verificação de vínculo genético em estudos de paternidade.

Capítulo 5 Mostra a aplicação das redes bayesianas na análise de dados biológicos voltados à verificação de vínculo genético em estudos de paternidade.

Capítulo 6 Mostra a modelagem, implementação e validação do sistema *THÊMIS* que utiliza redes bayesianas na representação do conhecimento acerca de estudos de paternidade.

Capítulo 7 Apresenta as considerações finais do trabalho (conclusões e trabalhos futuros).

Apêndice A Contém uma breve descrição de algumas ferramentas que provêem suporte à construção de redes bayesianas e à inferência nesse modelo probabilístico.

Apêndice B Contém os cálculos das probabilidades necessárias à construção das tabelas de probabilidade *a priori* das `v.a. cpg` e `cmg` apresentadas no Capítulo 5.

Apêndice C Contém os arquivos csv usados no Capítulo 6 (seção 6.5).

É importante ressaltar que os Capítulos 2 e 3 e o Apêndice A advêm de meu Trabalho de Conclusão de Curso (ver Costa, 2007), sendo incorporados no presente trabalho após uma acurada revisão.

Capítulo 2

Conceitos de Probabilidade e de Estatística

Neste capítulo, serão abordados alguns dos conceitos de probabilidade e de estatística que terão grande utilidade no entendimento das redes bayesianas, tais como: probabilidade *a priori*, probabilidade *a posteriori*, independência condicional e teorema de Bayes.

2.1 Teoria dos Conjuntos

O conhecimento de alguns conceitos da teoria dos conjuntos é importante para que haja um bom entendimento da teoria das probabilidades.

2.1.1 Definição

Pode-se definir conjunto como uma coleção de objetos denominados elementos. Esses elementos são indicados por letras minúsculas, ao passo que os conjuntos são representados por letras maiúsculas. As relações entre elementos e conjuntos são expressas pelos símbolos pertence ($\in$) e não pertence ($\notin$). Já as relações entre conjuntos são indicadas pelos símbolos está contido ($\subset$), não está contido ($\not\subset$), contém ($\supset$), não contém ($\not\supset$), igual ($=$) e diferente ($\neq$).

O conjunto que não possui elementos é denominado conjunto vazio e representado por $\emptyset$ ou $\{\}$, enquanto que o conjunto ao qual todos os objetos que estão sendo estudados pertecem é chamado conjunto fundamental ou conjunto universo, sendo representado geralmente pela letra maiúscula U ou, em se tratando de espaços amostrais, por Ω.

2.1.2 Representação

Um conjunto pode ser representado utilizando-se as três formas descritas a seguir.

Extensão Enumeram-se seus elementos, os quais são escritos entre chaves e separados por vírgulas.

Compreensão O conjunto é representado por meio de uma sentença a partir da qual se pode deduzir seus elementos.

Figuras O conjunto é representado por meio do chamado diagrama de Venn.

2.1.3 Operações

As operações primitivas entre conjuntos são descritas a seguir.

União $A \cup B$ (lê-se união de A e B) é o conjunto formado por todos os elementos que pertencem a A ou a B.

Intersecção $A \cap B$ (lê-se intersecção de A e B) é o conjunto formado por todos os elementos que pertencem a A e a B.

Complemento A^c (lê-se complemento de A) é o conjunto dos elementos que pertencem ao conjunto universo U, mas não pertencem a A.

Outra operação importante entre conjuntos é a **Diferença**. $A \setminus B$ (lê-se diferença de A e B) é o conjunto dos elementos que pertencem a A, mas não pertencem a B, ou seja, $A \setminus B = A \cap B^c$.

2.2 Modelos Matemáticos

Há situações em que não se pode prever com exatidão a ocorrência de um determinado fenômeno. A previsão do tempo é um exemplo disso, pois não é possível prever com certeza como será o clima num dado momento futuro. A situações como essa, dá-se o nome de fenômeno aleatório —denominação atribuída à situação ou acontecimento cujos resultados não podem ser previstos com certeza (Magalhães & de Lima, 2002). Este tipo de fenômeno produz resultados diferentes mesmo sendo submetido às mesmas condições.

Para quantificar as incertezas das várias ocorrências de um experimento aleatório (nome dado ao fenômeno que se pretende observar) utilizam-se os chamados modelos probabilísticos —modelos matemáticos não determinísticos que permitem prever a probabilidade de um dado resultado ocorrer sem

a necessidade de repetir a experiência. É importante mencionar que experimento aleatório é outra denominação para fenômeno aleatório.

Ao contrário dos modelos matemáticos não determinísticos que são usados para quantificar experimentos aleatórios, os modelos matemáticos determinísticos são utilizados para quantificar experimentos que geram resultados iguais quando executados sob as mesmas condições. Por exemplo, sabendo-se que a velocidade de um projétil é dada pela fórmula $V = \Delta S/\Delta T$, onde V é a velocidade, ΔS é a variação do espaço e ΔT é a variação do tempo, dadas as variações do espaço e do tempo, é obtido o valor da velocidade V. Se forem executados vários cálculos consecutivos com os mesmos valores para essas variáveis, obter-se-á sempre o mesmo valor para V.

2.3 Noções de Estatística

A Estatística pode ser vista como um conjunto de técnicas que provêem sistematicamente organização, descrição, análise e interpretação de um conjunto de valores (dados) obtidos por meio de estudos e experimentos realizados nas mais diversas áreas da atividade humana (Magalhães & de Lima, 2002). Pode-se dividi-la em três áreas:

Estatística Descritiva Essa área é utilizada na fase inicial da análise dos dados, tendo por finalidade extrair dos dados colhidos o maior número possível de informações relevantes referentes ao fenômeno em estudo. Essa tarefa, aparentemente simples, pode se tornar bastante complexa à medida que a quantidade de informações sobre o fenômeno aumenta. Em suma, citando Magalhães & de Lima (2002, p. 2):

> *a estatística descritiva pode ser definida como um conjunto de técnicas destinadas a descrever e resumir os dados, a fim de que possamos tirar conclusões a respeito de características de interesse.*

Probabilidade É a teoria matemática utilizada no estudo da incerteza proveniente de fenômenos aleatórios (ver seção 2.3.2).

Inferência Estatística É a obtenção de conclusões a partir de um subconjunto de valores, sendo este, em geral, bem menor que o conjunto inicial. Sua utilização só faz sentido quando não é possível ter acesso a todo o conjunto de dados que se pretende analisar. Isso ocorre, principalmente, por motivos econômicos, étnicos e físicos.

Dá-se o nome de população ao conjunto dos dados dos quais se pretende observar características de interesse. Esse vocábulo pode referir-se tanto a indivíduos quanto ao alvo cuja característica de interesse é pertinente. Dessa forma, tanto os alunos de uma dada instituição quanto os sistemas produzidos por uma fábrica de software podem compor uma população. Conforme mencionado anteriormente, na grande maioria dos casos, não é possível acessar toda a população que se pretende estudar, havendo, pois, a necessidade de se aplicar técnicas de inferência estatística para obter conclusões a partir de subconjuntos de valores, denominados amostras.

2.3.1 Freqüência

Antes de definir o que é "freqüência", será explicado o que são variáveis no contexto da Estatística e quais os seus tipos, haja vista que os conceitos de freqüência e de variável estão intimamente relacionados.

Ao se analisar uma população, em geral, é coletada uma enorme quantidade de dados, os quais devem ser "tratados", a fim de se extrair de maneira rápida e objetiva o maior número possível de informações consistentes e de interesse. Tabelas de freqüência e gráficos são procedimentos empregados para esse fim.

Suponha que se queira fazer uma pesquisa em uma empresa para analisar algumas características dos indivíduos que a compõem como, por exemplo, o sexo, o nível de escolaridade, a altura e o número de filhos. Uma das formas de se obter informações a respeito dessa população é por meio de questionários. Após a tabulação do questionário, o conjunto de informações disponíveis é chamado de tabela de dados brutos.

As características (sexo, nível de escolaridade, altura e número de filhos) perguntadas aos indivíduos são denominadas variáveis e são classificadas em dois tipos: quantitativas e qualitativas (Magalhães & de Lima, 2002). As variáveis númericas (quantitativas) podem ser subdivididas em discretas —são as que resultam de contagens, assumindo, em geral, valores inteiros— e contínuas —são, em geral, provenientes de mensuração, assumindo valores em intervalos reais. Como exemplo de variáveis quantitativas tem-se o número de filhos (quantitativa discreta) e altura (quantitativa contínua). As variáveis qualitativas são caracterizadas por representarem atributos e/ou qualidades, podendo ser subdivididas em nominais —são as que não possuem ordenação natural— e em ordinais —são as que têm uma ordenação natural. As variáveis sexo (masculino ou feminino) e nível de escolaridade (fundamental, médio ou superior) são qualitativas nominais e ordinais respectivamente.

A Figura 2.1 resume por meio de um esquema a classificação das variáveis.

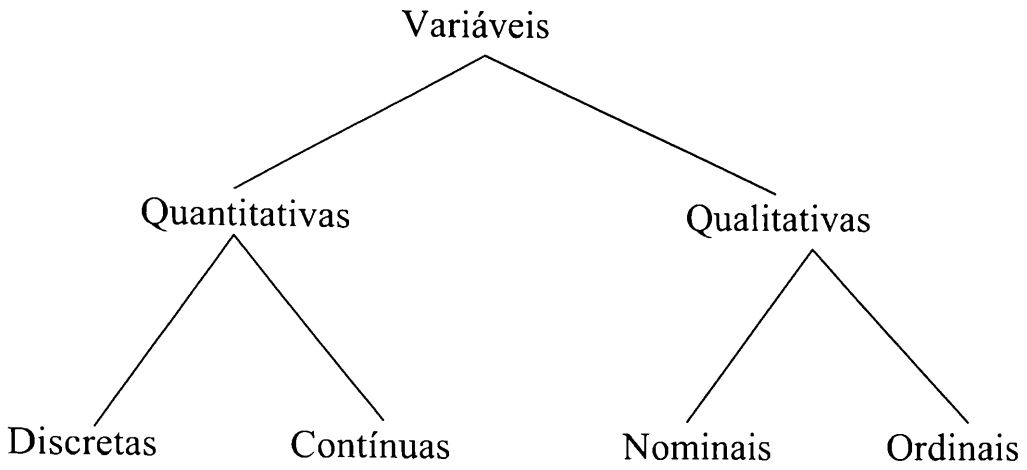

Figura 2.1: Classificação de variáveis

A tabela de dados brutos obtida após a tabulação do questionário, apesar de conter uma grande quantidade de informações, pode não ser funcional para se tirar conclusões sobre as variáveis de interesse. Para suprir essa deficiência é contruída uma nova tabela, que conterá os valores das variáveis e suas respectivas contagens, denominada *tabela de freqüência*. A essas contagens dá-se o nome de *freqüência absoluta* ou, simplesmente, *freqüência*.

Seja E um experimento (como, por exemplo, jogar uma moeda) e C um resultado possível desse experimento (como, por exemplo, ocorrer cara no lançamento da moeda). Supondo que E seja repetido n vezes e que C ocorreu n_C vezes, tem-se que:

I) n_C é a freqüencia absoluta de C, ou seja, a quantidade de vezes que o experimento E teve C como resultado, inclusive as repetições;

II) $\int C = n_C/n$ é a freqüência relativa de C nas n repetições do experimento E. A freqüência relativa de C possui as seguintes propriedades:

(P1) $0 \leq \int C \leq 1$;

(P2) $\int C = 1$ caso C ocorra em todas as repetições do experimento E;

(P3) $\int C = 0$ caso C não ocorra nas n repetições do experimento E.

Suponha que ao se jogar 5 vezes uma moeda comum, obtenham-se os seguintes resultados: C, K, K, C, C, onde C representa o resultado 'ocorrer cara' e K, 'ocorrer coroa'. Sendo n_C e f_C as freqüências absoluta e relativa de C respectivamente, tem-se que no final da quinta jogada $n_C = 3$ e $f_C = 3/5$.

2.3.2 Probabilidade

O conjunto de todos os resultados possíveis de um fenômeno aleatório é denominado *espaço amostral*, representado pela letra grega ômega (Ω). Os subconjuntos de Ω são denominados eventos e representados por letras latinas maiúsculas.

As operações mostradas na seção 2.1.3 são estendidas para os eventos de um espaço amostral, haja vista que esses eventos são conjuntos. Portanto, dados dois eventos A e B, tem-se que:

- $A \cup B$ representa a ocorrência de pelo menos um dos dois eventos;
- $A \cap B$ representa a ocorrência dos dois eventos. Quando $A \cap B = \{\}$, os eventos A e B são chamados disjuntos ou mutuamente exclusivos;
- A^c representa a não ocorrência do evento A.

Outra operação interessante é a *Diferença*. Dados dois eventos A e B, a diferença de A e B ($A \setminus B$) representa a ocorrência apenas do evento A.

Um conceito importante é o de eventos complementares. Sejam A e B dois eventos, diz-se que estes são complementares se eles são disjuntos ($A \cap B = \{\}$) e $A \cup B = \Omega$.

Definição 1 (Probabilidade) *A probabilidade* Pr *é uma função que atribui valores reais aos eventos do espaço amostral, devendo satisfazer as seguintes propriedades:*

(P1) $0 \leq \Pr(A) \leq 1, \forall A \subset \Omega$;

(P2) $\Pr(\Omega) = 1$;

(P3) $\Pr(\emptyset) = 0$;

(P4) $\Pr(\cup_{i=1}^{n}) = \sum_{i=1}^{n} \Pr(A_i)$, *com todos os* A_i *disjuntos.*

A atribuição de probabilidades aos elementos que constituem o espaço amostral pode ser feita de dois modos (ver Magalhães & de Lima, 2002, p. 38):

- a partir de características teóricas do fenômeno;
- a partir das freqüências de ocorrência (ver seção 2.3.1).

Uma discussão interessante sobre a origem das probabilidades pode ser encontrada em Russell & Norvig (2004, p. 459).

O conceito básico da linguagem empregada pela teoria das probabilidades é a variável aleatória (v.a.). Uma v.a. pode ser vista, informalmente, como um característico numérico que representa o resultado de um experimento (James, 1981, p. 35). Os valores que a variável pode assumir formam o seu domínio.

As variáveis aleatórias podem ser classificadas em booleanas, discretas, contínuas, mistas e singulares (para maiores detalhes, ver James, 1981).

Abaixo é dada uma definição matemática para uma variável aleatória segundo James (1981).

Definição 2 (Variável Aleatória) *Seja Ω um espaço amostral qualquer e* Pr *uma probabilidade sobre seus elementos. A transformação $X : \Omega \to \mathbb{R}$ é chamada variável aleatória real (v.a.$\mathbb{R}$) se sabe-se calcular*

$$\Pr(X \leq x) = \Pr(\{\omega \in \Omega : X(\omega) \leq x\})$$

para todo $x \in \mathbb{R}$.

A seguir são mostrados os chamados **axiomas de Kolmogorov**, os quais foram formulados pelo matemático russo Andrei Kolmogorov que mostrou como elaborar a teoria das probabilidades a partir dos princípios básicos dessa ciência (Russell & Norvig, 2004, p. 458).

Axioma 1 *Todas as probabilidades estão no intervalo fechado $[0,1]$, ou seja, $0 \leq \Pr(A) \leq 1, \forall A \subset \Omega$.*

Axioma 2 *Eventos certos têm probabilidade 1, ao passo que eventos impossíveis têm probabilidade 0, ou seja, $\forall A \subset \Omega$, se A é um evento certo, tem-se $\Pr(A) = 1$ e se A é um evento impossível, tem-se $\Pr(A) = 0$.*

Axioma 3 *A probabilidade da união de dois eventos quaisquer do espaço amostral é dada pela equação (2.1).*

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B), \forall A, B \subset \Omega \tag{2.1}$$

Obs.: No presente trabalho, a probabilidade da intersecção de n eventos $(A_1, A_2, \ldots, A_n)$ poderá ser representada de duas formas:

- $\Pr(A_1, A_2, \ldots, A_n)$ ou
- $\Pr(A_1 \cap A_2 \cap \cdots \cap A_n)$.

Definição 3 (Evento atômico) *Um evento atômico é uma especificação completa de um experimento, podendo ser considerado uma atribuição de valores aos possíveis resultados deste.*

Por exemplo, analise o experimento E que consiste em jogar uma moeda e um dado (não viciados) nessa mesma ordem. Sendo a moeda composta por duas faces (cara representado por C e coroa, por K) cada uma com probabilidade de ocorrência $\Pr = 1/2$ e o dado formado por seis faces com valores de 1 (um) a 6 (seis) e com probabilidade de ocorrência $\Pr = 1/6$ associada a cada face, tem-se a existência, nesse experimento, de doze eventos atômicos distintos, cada um com probabilidade de ocorrência $\Pr = 1/12$. $(C,1)$ e $(K,6)$ são alguns dos eventos atômicos do experimento E que representam a ocorrência das faces cara e 1 (um) e, coroa e 6 (seis) respectivamente.

Definição 4 (Probabilidade *a priori*) *A probabilidade* a priori, *também chamada de probabilidade incondicional, é o grau de crença associado a um evento na ausência de informações (evidências) sobre esse evento.*

Definição 5 (Probabilidade *a posteriori*) *A probabilidade* a posteriori, *também conhecida por probabilidade condicional, é o grau de crença associado a um evento dada alguma evidência sobre esse evento.*

Notação 4 (Probabilidade *a posteriori*) *A notação utilizada para probabilidade condicional é* $\Pr(A \mid B)$, *onde* A *e* B *são eventos. Essa expressão é lida da seguinte forma: probabilidade do evento* A *dada a ocorrência do evento* B.

A probabilidade condicional de um evento A dada a evidência do evento B é obtida por meio da equação (2.2), válida se, e somente se, $\Pr(B) \in]0,1]$.

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{2.2}$$

Um dos resultados mais úteis e instigantes da probabilidade decorre da definição de probabilidade condicional. A partir da equação (2.2) pode-se escrever $\Pr(A \cap B) = \Pr(A \mid B)\Pr(B) = \Pr(B \mid A)\Pr(A)$ que, generalizado, é conhecido como multiplicação de probabilidades.

Teorema 5 (Multiplicação de Probabilidades) *A probabilidade de ocorrência simultânea de* n *eventos pode ser calculada a partir das probabilidades condicionais:*

$$\Pr(A_1 \cap \cdots \cap A_n) = \Pr(A_1)\Pr(A_2 \mid A_1)\Pr(A_3 \mid A_1 \cap A_2)\cdots\Pr(A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Definição 6 (Partição do Espaço Amostral) *Os eventos* $B_1, \ldots, B_k$ *são uma partição do espaço amostral* Ω *se os mesmos são disjuntos dois a dois, se a união deles é o próprio espaço amostral e se todos possuem probabilidade estritamente positiva.*

Teorema 6 (Teorema da Probabilidade Total) *Sejam o espaço amostral* Ω*, o evento* A *e a partição* $B_1, \ldots, B_k$*, então*

$$\Pr(A) = \sum_{1 \leq i \leq k} \Pr(A \mid B_i) \Pr(B_i).$$

Esta relação é extremamente útil para calcular a probabilidade de ocorrência de um evento A a partir de evidências parciais.

Finalmente, está-se em condições de definir o conceito de independência entre eventos. Os eventos $A, B \subset \Omega$ são ditos independentes se

$$\Pr(A \cap B) = \Pr(A) \Pr(B).$$

Intuitivamente, dois eventos são independentes se a ocorrência de um não influi na probabilidade de ocorrência do outro.

Proposição 7 *Se* A *e* B *são eventos independentes, também o são os eventos* A *e* B^c*,* A^c *e* B*, e ainda* A^c *e* B^c*.*

Definição 7 (Independência coletiva) *Os eventos* $A_1, \ldots, A_n$ *são chamados 'coletivamente independentes' se*

$$\Pr(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}) = \Pr(A_{i_1}) \Pr(A_{i_2}) \cdots \Pr(A_{i_m}),$$

para todos os índices $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ *e todo* $2 \leq m \leq n$*.*

A independência coletiva de três eventos A, B e C requer não apenas que C seja independente de A, de B e que A e B sejam independentes entre si; a independência coletiva requer que C seja independente de $A \cap B$, de $A \cap B^c$ etc. A independência coletiva requer que a ocorrência do evento $A \cap B$ não afete a probabilidade de ocorrência de C etc.; por exemplo, requer que $\Pr(A \cap B \cap C) = \Pr(A \cap B) \Pr(C) = \Pr(A) \Pr(B) \Pr(C)$. A independência coletiva não decorre da independência dois-a-dois.

Considere a situação de se ter $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, eventos atômicos equiprováveis, e os eventos $A = \{\omega_1, \omega_4\}$, $B = \{\omega_2, \omega_4\}$ e $C = \{\omega_3, \omega_4\}$. Dado que $\Pr(\omega_i) = 1/4$ para todo $1 \leq i \leq 4$, temos que $\Pr(A) = \Pr(B) = \Pr(C) = 1/2$. Também

tem-se que $Pr(A \cap B) = Pr(A \cap C) = Pr(B \cap C) = 1/4$; logo A, B e C são independentes dois-a-dois. Por outro lado, $Pr(A \cap B \cap C) = 1/4 \neq 1/8 = Pr(A)Pr(B)Pr(C)$ e, portanto, os eventos A, B e C não são coletivamente independentes.

Definição 8 (Independência condicional) *Os eventos* $A_1, \ldots, A_n$ *são 'condicionalmente independentes' dado um evento* C *se*

$$Pr(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n} \mid C) = Pr(A_{i_1} \mid C) Pr(A_{i_2} \mid C) \cdots Pr(A_{i_n} \mid C).$$

Teorema 8 (Teorema de Bayes) *Sejam o espaço amostral* Ω, *o evento* A *e a partição* $B_1, \ldots, B_k$, *então para todo* $1 \leq i \leq k$ *vale que*

$$Pr(B_i \mid A) = \frac{Pr(A \mid B_i) Pr(B_i)}{\sum_{1 \leq i \leq k} Pr(A \mid B_i) Pr(B_i)}.$$

Uma forma um pouco mais geral deste teorema (ver Pearl, 1988) permite condicionar em algum outro evento E e afirmar

$$Pr(B_i \mid A, E) = \frac{Pr(A \mid B_i, E) Pr(B_i \mid E)}{\sum_{1 \leq i \leq k} Pr(A \mid B_i, E) Pr(B_i \mid E)}.$$

É freqüente encontrar a seguinte terminologia: 'H' denota a hipótese a ser verificada enquanto '*e*' denota a evidência coletada. Pretende-se conhecer $Pr(H \mid e)$ a partir da informação disponível, e para isso usa-se

$$Pr(H \mid e) = \frac{Pr(e \mid H) Pr(H)}{Pr(e)}. \tag{2.3}$$

Muito freqüentemente não há interesse em trabalhar com $Pr(e)$ e, para tanto, buscam-se outras formas de trabalhar com esta probabilidade conforme será visto a seguir.

É muito comum querer comparar as probabilidades a posteriori de uma hipótese e de sua antítese, isto é, dada a evidência *e* deseja-se comparar as probabilidades $Pr(H \mid e)$ e $Pr(H^c \mid e)$. Note que a evidência aconteceu, tem-se interesse em saber o que é mais provável, se a ocorrência de H ou a ocorrência de H^c. Para fazer isto, pode-se calcular a razão

$$\frac{Pr(H \mid e)}{Pr(H^c \mid e)} = \frac{Pr(e \mid H) Pr(H)}{Pr(e \mid H^c) Pr(H^c)}, \tag{2.4}$$

e com isso elimina-se a necessidade de computar $Pr(e)$.

A razão

$$\frac{Pr(H)}{Pr(H^c)} = \frac{Pr(H)}{1 - Pr(H)} = O(H)$$

é conhecida como 'chances relativas da priori' (*prior odds*), enquanto que a razão

$$\frac{\Pr(e \mid H)}{\Pr(e \mid H^c)} = \mathcal{L}(e \mid H)$$

recebe o nome de 'razão de verossimilhança' (*likelihood ratio*).

Pode-se escrever a equação (2.4) como

$$\mathcal{O}(H \mid e) = \mathcal{L}(e \mid H)\mathcal{O}(H),$$

que recebe o nome de 'chances a posteriori' da hipótese H tendo sido observada a evidência e (*posterior odds*). Note que este objeto só compara as chances da hipótese versus as chances da antítese, dada a ocorrência de uma única evidência, e.

Um conceito muito importante é o de 'momento de ordem k da variável aleatória X' que é definido pela integral

$$E(X^k) = \int_{\mathbb{R}} x^k f(x)\,dx$$

se existir e se a variável aleatória X for contínua, ou pelo somatório

$$E(X^k) = \sum_{i \in \mathbb{Z}} x_i^k \Pr(X = x_i)$$

se existir e se a variável aleatória X for discreta. Sendo importante informar que $f(x)$ é a função densidade de probabilidade da v.a. X e que $\Pr(X = x_i)$ é a função de probabilidade desta variável. O primeiro momento, se existir, chama-se 'média' ou 'esperança matemática', e a diferença entre o segundo momento e o quadrado da média, se existir, chama-se 'variância'. Se a variância existir, a sua raíz quadrada recebe o nome de 'desvio padrão'.

O exemplo a seguir ilustra os conceitos de distribuições conjunta, marginal e condicional, de esperança, de variância e de covariância.

Considere um domínio simples, composto por apenas duas variáveis aleatórias binárias: X e Y. A atribuição do valor zero (um, respectivamente) a essas variáveis tem significado booleano de falso (verdadeiro, respectivamente).

Suponha que a distribuição conjunta dessas duas variáveis aleatórias binárias está definida sobre quatro eventos atômicos de acordo com a tabela a seguir.

(X,Y)	Pr(X,Y)
(0,0)	0,12
(0,1)	0,08
(1,0)	0,16
(1,1)	0,64

Outra tabela por meio da qual se pode representar a distribuição conjunta das variáveis aleatórias binárias X e Y é mostrada abaixo.

X\Y	0	1
0	0,12	0,08
1	0,16	0,64

A distribuição conjunta permite calcular a probabilidade de qualquer evento (simples ou composto). Por conseguinte, a partir da tabela acima, pode-se calcular a probabilidade de qualquer evento, como, por exemplo, $\Pr(Y=1 \mid X=0)$. Essa probabilidade é calculada como segue:

$$\begin{aligned}
\Pr(Y=1 \mid X=0) &= \frac{\Pr(Y=1,X=0)}{\Pr(X=0)} \\
&= \frac{\Pr(Y=1,X=0)}{\sum_{j=0,1}\Pr(X=0,Y=j)} \\
&= \frac{\Pr(Y=1,X=0)}{\Pr(X=0,Y=0)+\Pr(X=0,Y=1)} \\
&= \frac{0,08}{0,12+0,08} = 0,40.
\end{aligned}$$

A distribuição marginal de cada variável aleatória é a lei que a governa sem considerar as outras. Logo, a distribuição de marginal de X é dada pela função de probabilidade $((0,2/10),(1,8/10))$, ao passo que a de Y é dada pela função de probabilidade $((0,28/100),(1,72/100))$.

As variáveis aleatórias X e Y serão independentes se para todo par de valores i,j valer que $\Pr((X,Y)=(i,j)) = \Pr(X=i)\Pr(Y=j)$. Verifica-se se vale a independência com, por exemplo,

$$\begin{aligned}
\Pr((X,Y)=(1,1)) &= 0,64 \\
\Pr(X=1)\cdot\Pr(Y=1) &= 0,80\cdot 0,72 = 0,576
\end{aligned}$$

Como $\Pr((X,Y)=(1,1)) \neq \Pr(X=1)\cdot\Pr(Y=1)$, X e Y não são independentes.

A distribuição condicional de uma variável aleatória é a lei que a governa, dado o conhecimento da ocorrência de outras variáveis aleatórias. Por exem-

plo, $Pr(Y = i \mid X(\omega) = 0)$. Definindo $W = (Y \mid X(\omega) = 0)$, a lei que governa W é dada por $((0,6/10),(1,4/10))$. Analogamente, definindo-se $Z = (X \mid Y(\omega) = 1)$, a lei que descreve o comportamento de Z é $((0,1/9),(1,8/9))$.

As esperanças e variâncias dessas duas últimas variáveis aleatórias podem ser facilmente calculadas. Para W tem-se:

$$\begin{aligned} E(W) &= 0 \cdot \frac{6}{10} + 1 \cdot \frac{4}{10} = \frac{4}{10}, \\ E(W^2) &= 0^2 \cdot \frac{6}{10} + 1^2 \cdot \frac{4}{10} = \frac{4}{10}, \\ Var(W) &= E(W^2) - E^2(W) = \frac{4}{10} - \frac{16}{100} = \frac{6}{25}. \end{aligned}$$

A covariância e a correlação entre duas variáveis aleatórias descrevem a associação entre elas:

$$Cov(X,Y) = E(XY) - E(X)E(Y) \quad \text{e} \quad \rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

No caso em questão, tem-se que a distribuição de XY é caracterizada pela função de probabilidade $((0,9/25),(1,16/25))$, logo $E(XY) = 0,64$. Com isso, $Cov(X,Y) = 0,64 - 0,80 \cdot 0,72 = 0,064$.

2.4 Comentários

No presente capítulo, foram abordados alguns dos conceitos de probabilidade e de estatística que serão úteis no entendimento das redes bayesianas explanadas no próximo capítulo. Dentre os conceitos aqui discutidos, estão: variáveis aleatórias, eventos complementares e mutuamente exclusivos, probabilidade condicional e incondicional, independência condicional, dentre outros.

Capítulo 3

Redes Bayesianas: Fundamentação Teórica

Neste capítulo, serão enfocados os conceitos de maior relevância sobre redes bayesianas cujo conhecimento se faz necessário aos que pretendem utilizar essa abordagem como meio de representação do conhecimento. Dentre os tópicos abordados, estão a construção de uma rede bayesiana e a inferência nesse tipo de rede probabilística.

3.1 Incerteza

Os seres humanos têm grande habilidade de tirar conclusões úteis a partir de informações incompletas e mal formadas. Isso é observado cotidianamente e, na grande maioria das vezes, essa atividade é realizada com sucesso. Analise, por exemplo, como é feito um diagnóstico médico, visto que qualquer tipo de diagnóstico está intimamente ligado à incerteza.

O paciente, inicialmente, procura um cardiologista, haja vista que está sentido dores próximo ao coração e bastante cansaço. Ao se dirigir ao consultório médico, o cardiologista irá solicitar ao paciente que lhe informe os sintomas (evidências). Após isso, o médico provavelmente irá pedir que o paciente faça alguns exames que o auxiliarão no diagnóstico da doença (causa). Ao receber os laudos dos exames (sinais), o cardiologista irá analisá-los e, juntamente com as informações que lhe foram passadas pelo paciente (sintomas) e com a experiência que adquiriu com os anos de trabalho (paradigma), tentará diagnosticar a doença. Apesar de, em grande parte dos casos, o resultado obtido ser correto, não há garantia alguma. Isso ocorre porque o médico não possui informações precisas e suficientes para descobrir a enfermidade do paciente, tendo, pois, que lidar com incerteza.

3.2 Notação e Definições

A **rede bayesiana** —uma das abordagens estocásticas para a incerteza— é uma ferramenta que provê o cálculo de distribuições de probabilidades (conjuntas, marginais e condicionais) de conjuntos de variáveis aleatórias (Lucke, 1995). Essa informação é relevante e, tipicamente, não é evidente a partir do modelo do fenômeno.

Essa rede probabilística pode ser vista como um grafo orientado e acíclico cujos nós são identificados como variáveis aleatórias com distribuições caracterizadas por tabelas de probabilidade ou leis condicionais. A estrutura do grafo descreve a dependência entre as variáveis aleatórias. Esse tipo de rede pode ser especificado como segue:

1. Os nós da rede representam variáveis aleatórias.
2. Os nós são conectados por meio de setas. Se houver uma seta do nó P até o nó F, P será denominado pai de F (seu filho).
3. Cada nó X_i tem uma probabilidade condicional $Pr(X_i \mid Pais(X_i))$ que quantifica o efeito dos pais sobre o nó filho.
4. A distribuição da variável aleatória X_i, dados todos os nós que a precedem, só depende dos seus pais.

Dentre as vantagens das redes bayesianas, tem-se:

- representação e manipulação da incerteza baseadas em modelos matemáticos;
- modelagem do conhecimento de forma intuitiva acerca do domínio;
- permissão à realização de inferência causal, de diagnóstico, intercausal ou mista.

3.3 Tratamento do Conhecimento Incerto

Na seção seguinte, será usada uma rede bayesiana para se representar o processo que ocorre na efetuação de uma compra num sistema de compras on-line.

3.3.1 Sistema de Compras

Ao efetuar uma compra pela Internet utilizando um sistema de compras, como o Mercado Pago (MP, ver http://www.mercadolivre.com.br), o sistema encaminha uma notificação ao comprador (CNMP) e uma ao vendedor (VNMP), as quais podem perder-se. Sendo o vendedor notificado, a venda será processada (VP), havendo uma probabilidade de ocorrência associada a essa operação. Após isso, o vendedor entrará em contato com o comprador (CNV) e postará a mercadoria (PM); ambos os eventos têm probabilidades de ocorrência associadas. Executadas essas duas etapas, o comprador poderá conferir a postagem (CCP) através das informações que os correios (CDIP) irão, provavelmente, disponibilizar via *Web*. A partir da postagem, a mercadoria será entregue (ME), havendo uma probabilidade de ocorrência associada à essa operação; o vendedor será provavelmente qualificado (CQP), seguindo a (provável) liberação do dinheiro (PL) e o comprador, por sua vez, será qualificado pelo vendedor (CQ). A Figura 3.1 mostra esta rede de causas e efeitos.

Para a especificação de uma rede estar completa, é necessário fornecer a probabilidade de cada variável aleatória dadas todas as configurações possíveis das variáveis da qual depende, gerando a chamada *tabela de probabilidade condicional*, ou TPC. A Figura 3.1 mostra apenas uma TPC, a partir da qual se sabe que a probabilidade de PM ocorrer dado que VP ocorreu é 80%, ou seja, $\Pr(\texttt{PM} = 1 \mid \texttt{VP} = 1) = 80\%$, e que a probabilidade de PM ocorrer dado que VP não ocorreu é 0, isto é, $\Pr(\texttt{PM} = 0 \mid \texttt{VP} = 0) = 0$. Nela e no restante do texto, a dependência é denotada apenas como "$\Pr(\texttt{PM} \mid \cdot)$" sem fazer menção explícita às variáves aleatórias condicionantes.

Uma relação indireta, a partir dessas premissas, pode ser observada: a relação entre os correios disponibilizarem as informações de postagem e o dinheiro ser liberado ao vendedor. Essa questão é de difícil solução, já que, por não receber nenhuma confirmação de envio por parte dos correios, resta ao comprador a confiança no vendedor; por várias vezes a encomenda demora a chegar, ficando o comprador alheio ao que ocorre. Sabendo-se que o comprador pode desconfiar do vendedor, isso pode aparentar uma tentativa deste enganar aquele, o qual reterá o pagamento.

Modelar esse problema em uma rede probabilística não é difícil, já que este formalismo requer apenas a probabilidade das relações causais diretas. Porém, como resultado, obtém-se uma ferramenta poderosa, capaz de efetuar inferências indiretas, como a proposta. Nesse âmbito, muitas questões podem ser respondidas, sendo possível analisar qualquer relação, por mais indireta que pareça.

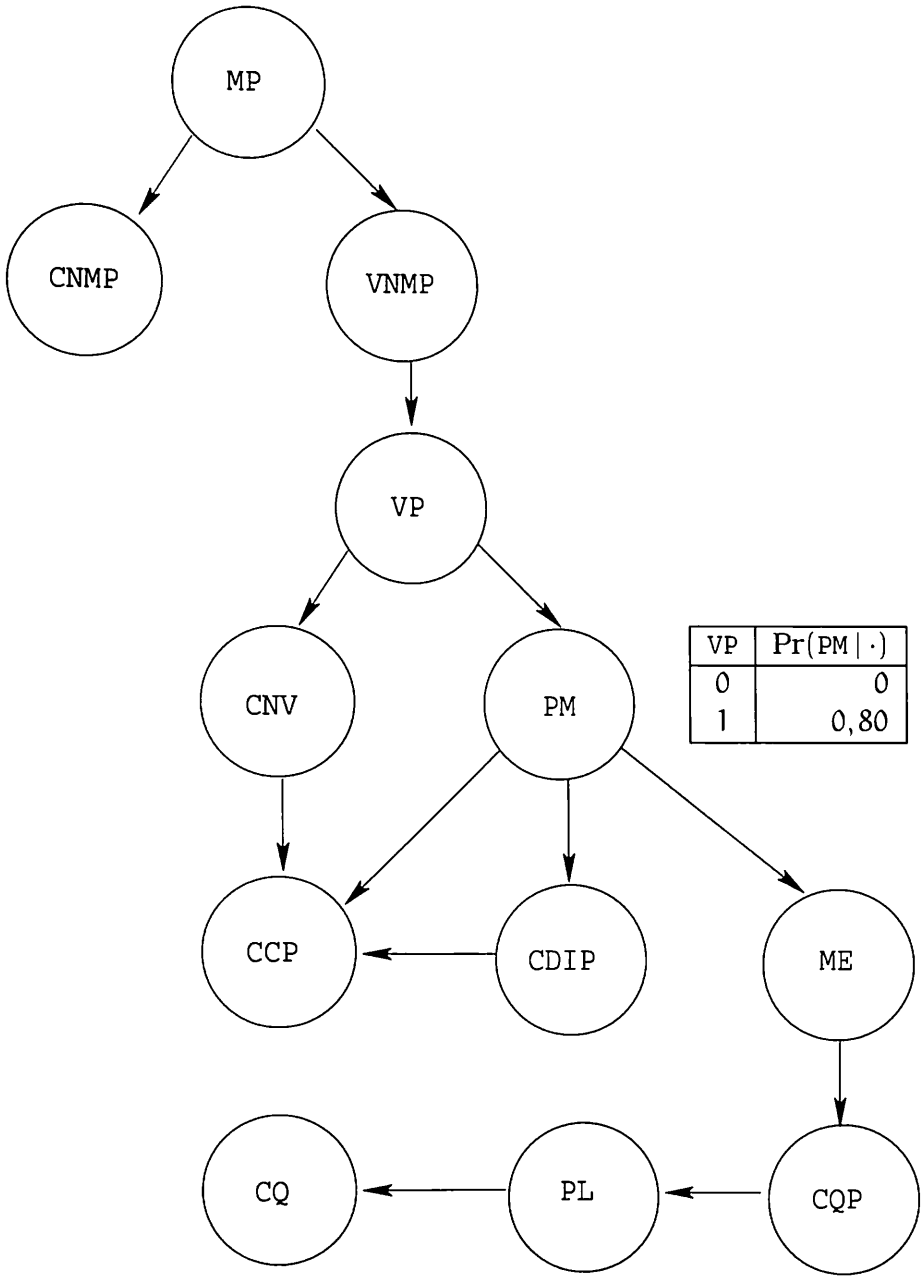

Figura 3.1: Modelagem do Mercado Pago com uma rede probabilística

3.4 Exemplo

Um exemplo clássico de abordagem bayesiana é mostrado no texto de Russell & Norvig (2004), onde é utilizada uma rede bayesiana para descrever as relações de causalidade.

A Figura 3.2 apresenta um exemplo simples de rede bayesiana com apenas dois nós. A variável aleatória R modela o fato de haver um roubo em um dia qualquer em uma determinada região de uma certa cidade. A variável aleatória B descreve se haverá boletim de ocorrência associado ao evento.

Pelo modelo descrito com esta rede bayesiana, pode-se concluir que trata-se de uma cidade relativamente perigosa (uma em cada cem pessoas é rou-

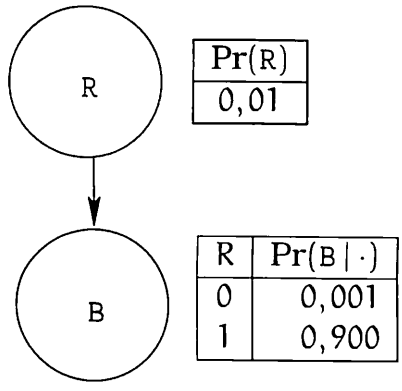

Figura 3.2: Rede bayesiana com dois nós

bada) e nove em cada dez pessoas roubadas cumprem com o seu dever de cidadãos reportando os roubos. Note que foi modelado o evento "não houve roubo, porém foi feito um boletim de ocorrência"; este evento possui probabilidade bem baixa.

A distribuição conjunta dessas duas variáveis aleatórias binárias está definida sobre quatro eventos atômicos, e pode ser descrita na forma de uma tabela como a seguinte.

(R,B)	$\Pr(R,B) = \Pr(B \mid R)\Pr(R)$
(0,0)	$(1-0,001)(1-0,01) = 0,98901$
(0,1)	$0,001 \cdot (1-0,01) = 0,00099$
(1,0)	$(1-0,900) \cdot 0,01 = 0.00100$
(1,1)	$0,900 \cdot 0,01 = 0,00900$

As probabilidades conjuntas foram calculadas utilizando o teorema do produto das probabilidades e as informações fornecidas pela topologia da rede bayesiana.

É interessante contrastar a intuição que se tem a respeito do problema com o modelo derivado da rede bayesiana. Note que o evento mais provável é não haver roubo nem haver boletim de ocorrência ($\Pr((R,B) = (0,0)) = 0,98901$); ele é quase cento e dez vezes mais provável do que o outro evento que intuitivamente é visto como freqüente: haver roubo e boletim de ocorrência:

$$\frac{\Pr((R,B) = (0,0))}{\Pr((R,B) = (1,1))} = \frac{0,98901}{0,00900} = 109,89.$$

Constata-se, assim, a influência que a distribuição *a priori* tem sobre o modelo.

A Figura 3.3 mostra um acréscimo de complexidade em relação à rede anterior. Além do roubo e do boletim de ocorrência, a variável aleatória N modela o fato de ser noticiado o roubo. Constata-se que o veículo de comunicação em

questão é um pouco sensacionalista, pois mesmo não havendo roubo ele dá a notícia com probabilidade 0,05 e dificilmente deixa de dar a notícia quando ela é verdadeira.

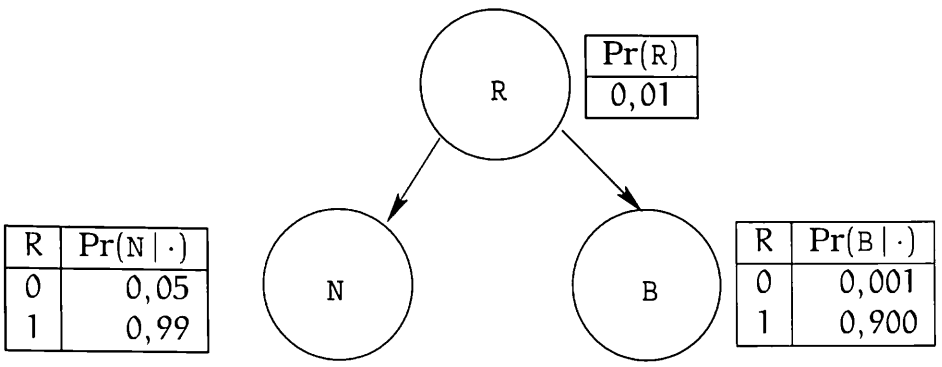

Figura 3.3: Rede bayesiana com três nós

A distribuição conjunta requer a especificação da probabilidade de oito eventos atômicos, conforme descrito na seguinte tabela. Para calculá-la utilizou-se novamente o teorema do produto das probabilidades que, neste caso, diz que $\Pr(N,B,R) = \Pr(R)\Pr(B \mid R)\Pr(N \mid B,R)$. Em princípio, não é conhecida a $\Pr(N \mid B,R)$, mas a topologia da rede informa que $\Pr(N \mid B,R) = \Pr(N \mid R)$, e esta informação está disponível.

(N,B,R)	Pr(R) Pr(B | R) Pr(N | R)	Pr(N,B,R)
(0,0,0)	0,99 · 0,999 · 0,95	0,9395595
(0,0,1)	0,01 · 0,100 · 0,01	0,0000100
(0,1,0)	0,99 · 0,001 · 0,95	0,0009405
(0,1,1)	0,01 · 0,900 · 0,01	0,0000900
(1,0,0)	0,99 · 0,999 · 0,05	0,0494505
(1,0,1)	0,01 · 0,100 · 0,99	0,0009900
(1,1,0)	0,99 · 0,001 · 0,05	0,0000495
(1,1,1)	0,01 · 0,900 · 0,99	0,0089100

O evento mais provável para este modelo é não haver roubo, nem boletim de ocorrência, nem notícia, mas não deixa de ser surpreendente que o segundo evento mais provável seja não haver roubo, nem boletim de ocorrência mas sim notícia. Novamente, é interessante avaliar a razão entre as probabilidades de dois eventos, por exemplo

$$\frac{\Pr((N,B,R) = (1,0,0))}{\Pr((N,B,R) = (1,1,1))} = \frac{0,99 \cdot 0,999 \cdot 0,05}{0,01 \cdot 0,900 \cdot 0,99} = 5,55.$$

Com a distribuição conjunta pode-se calcular, tal como já fora visto, diversas quantidades de interesse. Em particular, pode-se calcular $\mathrm{Cov}(B,N)$ e uma

medida da credibilidade do veículo de comunicação em questão, através de $\Pr(R \mid N)$.

3.5 Topologia

A topologia de uma rede bayesiana tem por função retratar a estrutura dos processos causais do domínio que se pretende representar. Após a especificação da topologia da rede, é necessário construir uma *tabela de probabilidade condicional – TPC* para cada nó.

Por construção em uma rede bayesiana, a distribuição de probabilidade conjunta sobre um conjunto de variáveis $(X_1, X_2, \cdots, X_n)$ é dada pela equação

$$\Pr(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} \Pr(X_i \mid \mathrm{Pais}(X_i)). \tag{3.1}$$

Na equação (3.1), está implícito que quando há independência condicional entre cada nó e seus antecessores na ordenação dos nós dados os seus pais, a rede de crença bayesiana é a representação correta para o domínio. Portanto, para se modelar uma rede causal topologicamente correta, é necessário que a escolha dos pais de cada nó seja feita mantendo-se essa propriedade, ou seja, a construção da topologia de uma rede de crença deve ser guiada pelas relações de independência condicional existentes entre as variáveis do domínio que se pretende representar.

O Pseudo-código 3.1 mostra um algoritmo para se construir a topologia de uma rede bayesiana mantendo-se a propriedade acima descrita.

Pseudo-código 3.1: Pseudo-código para construção de uma rede bayesiana

```
1  /*
2   * Passo 1
3   */
4    Escolher o conjunto de variáveis relevantes que descrevem
5    o domínio.
6  /*
7   * Passo 2
8   */
9    Escolher uma ordenação para as variáveis.
10 /*
11  * Passo 3
12  */
```

```
Enquanto existirem variáveis:
  (1) selecionar uma variável e adicionar um nó à rede
      para esta;
  (2) definir os pais de cada variável como o conjunto
      mínimo de nós já existentes na rede para os quais
      a propriedade de independência condicional se
      verifique;
  (3) definir a tabela de probabilidade condicional para
      cada variável.
```

Como cada nó se liga apenas aos nós definidos previamente, este algoritmo para construção de redes bayesianas garante que a rede é acíclica.

Quanto a ordenação dos nós da rede, esta deve ser feita adicionando-se primeiramente as causas (raízes da rede) e em seguida as variáveis que estas influenciam (efeitos) até que sejam atingidas as folhas da rede (variáveis que não influenciam nenhuma outra). Isso não significa que esta é a única forma de se ordenar os nós de uma rede probabilística, todavia esta ordenação (de causa para efeitos) propicia, na maioria das vezes, a criação de redes mais compactas cujas tabelas de probabilidades são mais fáceis de serem construídas.

A rede da Figura 3.4, obtida a partir da tabela que representa a distribuição conjunta das variáveis da rede da Figura 3.3, mostra que ao se construir uma rede bayesiana com ligações dos efeitos para as causas há a necessidade de serem especificadas dependências adicionais que requerem a definição de probabilidades de difícil obtenção, resultando, pois, em redes mais complexas e menos intuitivas.

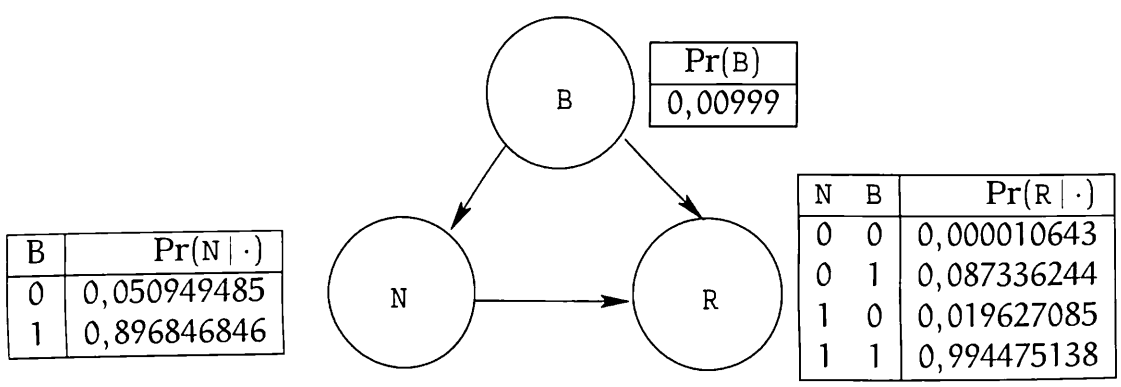

Figura 3.4: Rede bayesiana com três nós modificada

A seguir, são mostrados os cálculos das probabilidades para cada caso de condicionamento–*CC* das tabelas de probabilidades condicionais–*TPC* da rede da Figura 3.4. Como as variáveis da rede bayesiana são booleanas, foram

omitidos os cálculos das probabilidades dos valores falsos.

(*TPC–1*) $\Pr(\mathtt{B})$

(*CC–1*) $\Pr(\mathtt{B}=1)$

$$\begin{aligned}\Pr(\mathtt{B}=1) &= \sum_{\mathtt{N}=\{0,1\},\mathtt{R}=\{0,1\}} \Pr(\mathtt{N},\mathtt{B}=1,\mathtt{R}) \\ &= \Pr(\mathtt{N}=0,\mathtt{B}=1,\mathtt{R}=0)+\Pr(\mathtt{N}=0,\mathtt{B}=1,\mathtt{R}=1)+ \\ &\Pr(\mathtt{N}=1,\mathtt{B}=1,\mathtt{R}=0)+\Pr(\mathtt{N}=1,\mathtt{B}=1,\mathtt{R}=1) \\ &= 0,00999\end{aligned}$$

(*TPC–2*) $\Pr(\mathtt{N}\mid\mathtt{B})$

(*CC–1*) $\Pr(\mathtt{N}=1\mid\mathtt{B}=0)$

$$\begin{aligned}\Pr(\mathtt{N}=1\mid\mathtt{B}=0) &= \frac{\Pr(\mathtt{N}=1,\mathtt{B}=0)}{\Pr(\mathtt{B}=0)} \\ &= \frac{\sum_{\mathtt{R}=\{0,1\}}\Pr(\mathtt{N}=1,\mathtt{B}=0,\mathtt{R})}{\Pr(\mathtt{B}=0)} \\ &= \frac{\Pr(\mathtt{N}=1,\mathtt{B}=0,\mathtt{R}=0)+\Pr(\mathtt{N}=1,\mathtt{B}=0,\mathtt{R}=1)}{1-\Pr(\mathtt{B}=1)} \\ &= 0,050949485\end{aligned}$$

(*CC–2*) $\Pr(\mathtt{N}=1\mid\mathtt{B}=1)$

$$\begin{aligned}\Pr(\mathtt{N}=1\mid\mathtt{B}=1) &= \frac{\Pr(\mathtt{N}=1,\mathtt{B}=1)}{\Pr(\mathtt{B}=1)} \\ &= \frac{\sum_{\mathtt{R}=\{0,1\}}\Pr(\mathtt{N}=1,\mathtt{B}=1,\mathtt{R})}{\Pr(\mathtt{B}=1)} \\ &= \frac{\Pr(\mathtt{N}=1,\mathtt{B}=1,\mathtt{R}=0)+\Pr(\mathtt{N}=1,\mathtt{B}=1,\mathtt{R}=1)}{\Pr(\mathtt{B}=1)} \\ &= 0,896846846\end{aligned}$$

(*TPC–3*) $\Pr(\mathtt{R}\mid\mathtt{N},\mathtt{B})$

(CC–1) $\Pr(R = 1 \mid N = 0, B = 0)$

$$\begin{aligned}\Pr(R = 1 \mid N = 0, B = 0) &= \frac{\Pr(R = 1, N = 0, B = 0)}{\Pr(N = 0, B = 0)} \\ &= \frac{\Pr(R = 1, N = 0, B = 0)}{\sum_{R=\{0,1\}} \Pr(N = 0, B = 0, R)} \\ &= \frac{\Pr(R = 1, N = 0, B = 0)}{\Pr(N = 0, B = 0, R = 0) + \Pr(N = 0, B = 0, R = 1)} \\ &= 0,000010643\end{aligned}$$

(CC–2) $\Pr(R = 1 \mid N = 0, B = 1)$

$$\begin{aligned}\Pr(R = 1 \mid N = 0, B = 1) &= \frac{\Pr(R = 1, N = 0, B = 1)}{\Pr(N = 0, B = 1)} \\ &= \frac{\Pr(R = 1, N = 0, B = 1)}{\sum_{R=\{0,1\}} \Pr(N = 0, B = 1, R)} \\ &= \frac{\Pr(R = 1, N = 0, B = 1)}{\Pr(N = 0, B = 1, R = 0) + \Pr(N = 0, B = 1, R = 1)} \\ &= 0,087336244\end{aligned}$$

(CC–3) $\Pr(R = 1 \mid N = 1, B = 0)$

$$\begin{aligned}\Pr(R = 1 \mid N = 1, B = 0) &= \frac{\Pr(R = 1, N = 1, B = 0)}{\Pr(N = 1, B = 0)} \\ &= \frac{\Pr(R = 1, N = 1, B = 0)}{\sum_{R=\{0,1\}} \Pr(N = 1, B = 0, R)} \\ &= \frac{\Pr(R = 1, N = 1, B = 0)}{\Pr(N = 1, B = 0, R = 0) + \Pr(N = 1, B = 0, R = 1)} \\ &= 0,019627085\end{aligned}$$

(*CC–4*) $\Pr(\mathtt{R}=1 \mid \mathtt{N}=1,\mathtt{B}=1)$

$$\begin{aligned}\Pr(\mathtt{R}=1 \mid \mathtt{N}=1,\mathtt{B}=1) &= \frac{\Pr(\mathtt{R}=1,\mathtt{N}=1,\mathtt{B}=1)}{\Pr(\mathtt{N}=1,\mathtt{B}=1)} \\ &= \frac{\Pr(\mathtt{R}=1,\mathtt{N}=1,\mathtt{B}=1)}{\sum_{\mathtt{R}=\{0,1\}}\Pr(\mathtt{N}=1,\mathtt{B}=1,\mathtt{R})} \\ &= \frac{\Pr(\mathtt{R}=1,\mathtt{N}=1,\mathtt{B}=1)}{\Pr(\mathtt{N}=1,\mathtt{B}=1,\mathtt{R}=0)+\Pr(\mathtt{N}=1,\mathtt{B}=1,\mathtt{R}=1)} \\ &= 0,994475138\end{aligned}$$

A saber que as variáveis N, B são condicionalmente independentes dada alguma evidência sobre a variável R, outra forma de se calcular $\Pr(\mathtt{R}=1 \mid \mathtt{N}=1,\mathtt{B}=1)$ é utilizando o conceito de independência condicional (ver definição 8). Esse cálculo é mostrado abaixo.

$$\begin{aligned}\Pr(\mathtt{R}=1 \mid \mathtt{N}=1,\mathtt{B}=1) &= \frac{\Pr(\mathtt{N}=1,\mathtt{B}=1 \mid \mathtt{R}=1)\Pr(\mathtt{R}=1)}{\Pr(\mathtt{N}=1,\mathtt{B}=1)} \\ &= \frac{\Pr(\mathtt{N}=1 \mid \mathtt{R}=1)\Pr(\mathtt{B}=1 \mid \mathtt{R}=1)\Pr(\mathtt{R}=1)}{\sum_{\mathtt{R}=\{0,1\}}\Pr(\mathtt{N}=1,\mathtt{B}=1,\mathtt{R})} \\ &= \frac{0,99\cdot 0,9\cdot 0,01}{0,0089595} \\ &= 0,994475138\end{aligned}$$

3.6 Comentários

No presente capítulo, foram abordados os conceitos de maior relevância sobre redes bayesianas necessários ao entendimento de como se aplicar esse formalismo na obtenção de conclusões úteis, a partir de dados incompletos e imprecisos, visando a uma inteira compreensão das redes bayesianas presentes no Capítulo 5. Dentre os conceitos aqui discutidos, estão: incerteza e topologia.

Capítulo 4

Análise Forense de DNA: Fundamentação Teórica

Neste capítulo, serão abordados os conceitos de maior relevância sobre análise forense de DNA autossômico na verificação de vínculo genético em estudos de paternidade, sendo importante ressaltar que grande parte do que será exposto tem por base Goodwin et al. (2007) e Butler (2005)

4.1 Considerações Preliminares

Nos últimos 20 anos, as aplicações focadas em análise genética têm revolucionado a ciência forense. A análise de regiões de DNA que apresentavam polimorfismos fez surgir em 1984 o termo *DNA fingerprint*. No ano seguinte, a análise de perfis de DNA foi aplicada com sucesso na resolução de disputas de imigração. Em 1986, o DNA foi usado pela primeira vez em um caso criminal, permitindo a identificação do assassino de duas estudantes em Leicestershire, Reino Unido (maiores detalhes em `http://www.forensic.gov.uk/html/media/case-studies/f-18.html` acessado pela última vez em 25-02-2009).

O uso da análise de DNA foi rapidamente adotado pela comunidade forense e hoje constitui uma importante ferramenta de investigação criminal.

4.2 A Estrutura do DNA e o Genoma Humano

O genoma de cada indivíduo possui uma grande quantidade de DNA, o que é de fundamental importância para a geração de perfis genéticos.

O DNA (ácido desoxiribonucléico) contém toda informação que um organismo necessita em termos de função e reprodução. Estando presente no

núcleo das células, ele é o responsável pela especificação da função de cada uma das células constituintes do organismo.

O componente básico da molécula de DNA é o nucleotídeo, composto por um grupo trifosfato, um açúcar (desoxiribose) e uma base nitrogenada. O açúcar e o grupo trifosfato integram a estrutura da molécula de DNA, ao passo que as bases nitrogenadas: adenina (A), citosina (C), guanina (G) e timina (T), contêm as informações sobre a síntese protéica (ver Figura 4.1 obtida em (da Silva Júnior & Sasson, 2002)).

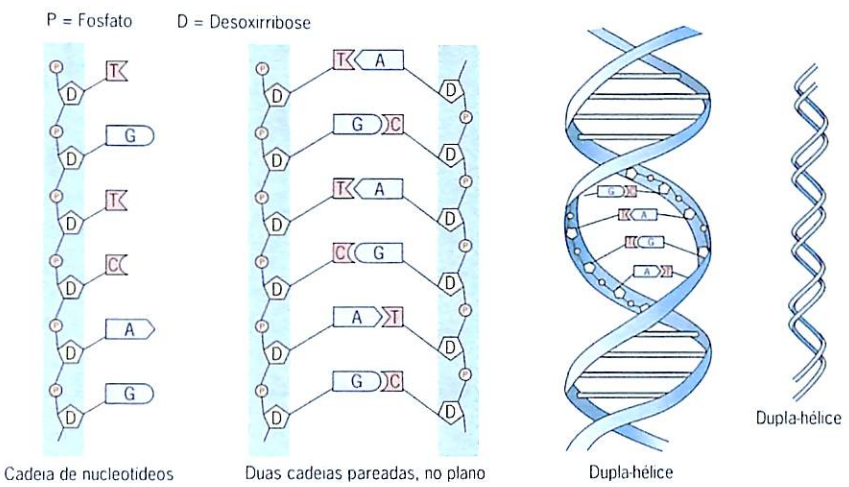

Figura 4.1: Estrutura do DNA

Os nucleotídeos se organizam em duas cadeias que se entrelaçam formando uma dupla hélice. Essa estrutura se assemelha a uma escada enrolada sobre si mesma em forma de espiral, onde os corrimões constituem-se do grupo trifosfato e da desoxirobose (parte invariante da molécula) e cada degrau é formado por duas bases nitrogendas, cuja ligação é feita por pontes de hidrogênio. Cada base é atraída por sua base complementar: adenina sempre forma par com timina e citosina com guanina. Sendo assim, se em uma cadeia de nucleotídeos há a seqüência ACGT, na outra haverá a seqüência TGCA, haja vista que as cadeias que constituem a molécula de DNA são complementares.

Em cada célula nuclear humana há duas cópias completas do genoma, complemento genético haplóide de um organismo vivo, que contém cerca de 3.200.000.000 pares de base (bp) de informação, os quais estão organizados em 23 cromossomos, cujos comprimentos variam de 73 mm a 14 mm.

Há na espécie humana 46 cromossomos, formando, 23 pares, dos quais 22 pares são autossômicos e 1 par está envolvido na determinação do sexo. X, Y são os cromossomos sexuais masculinos, ao passo que X, X correspondem

aos cromossomos sexuais femininos (ver Figura 4.2 obtida em Goodwin et al. (2007)). Cada cromossomo contém quantidades contínuas de DNA, o maior (cromossomo 1) tem aproximadamente 250.000.000 bp ao passo que o menor (cromossomo 22) tem apenas cerca de 50.000.000 bp (Goodwin et al., 2007).

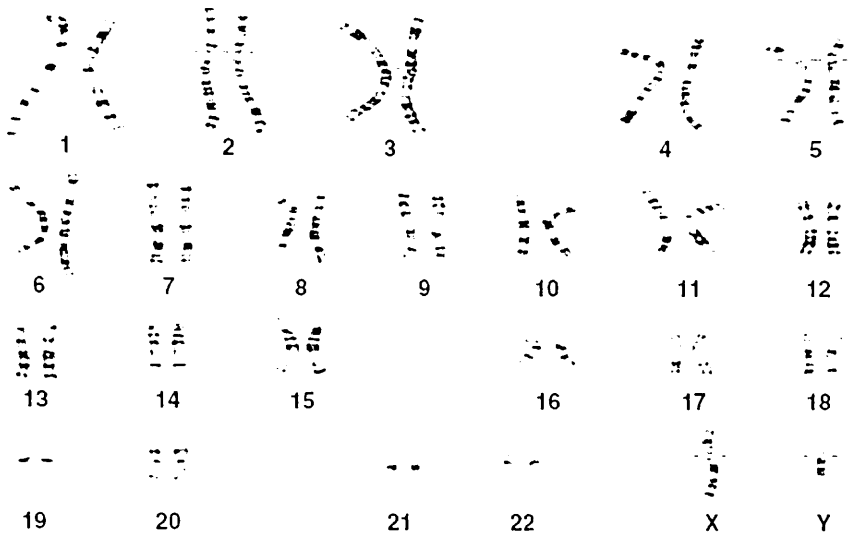

Figura 4.2: Cromossomos na espécie humana

Existem nos cromossomos regiões nas quais é possível encontrar um gene ou uma seqüência de nucleotídeos não-codificadores. A essas regiões é atribuído o nome de *locus*. Em cromossomos homólogos, cromossomos que se alinham durante a divisão celular meiótica, os genes localizados no mesmo *locus*, denominados genes alelos ou alelos, são responsáveis pela mesma característica genética.

Indivíduos heterozigóticos para uma dada característica possuem alelos diferentes no *locus* em questão, ao passo que indíviduos homozigóticos possuem alelos iguais. A configuração desses alelos corresponde ao genótipo do *locus*. Desse forma, pode-se definir o genótipo de um indivíduo como sendo o conjunto dos genótipos de seus *loci*[1]. A expressão desse genótipo, juntamente com a influência exercida pelo meio ambiente, constitui o fenótipo do indivíduo.

A reprodução humana é sexuada e, por conseguinte, está intimamente ligada a dois processos:

Meiose Processo de divisão celular por meio do qual o número de cromossomos de uma célula é reduzido à sua metade. Os gametas são formados por este processo.

Fecundação Fusão do gameta masculino com o gameta feminino.

[1] Plural de *locus*.

Em outras palavras, por meiose, o número diplóide de cromossomos (2n) é reduzido à metade (n – haplóide), e pela fecundação restabelece-se o número 2n (diplóide) típico da espécie. Dessa forma, em cada par de cromossomos homólogos encontrado em um indivíduo, o genótipo de cada *locus* é constituído por um alelo proveniente do pai e o outro herdado da mãe. Essa troca de material genético entre indivíduos de uma população são os responsáveis pela manutenção da variabilidade genética.

É graças a essa variabiliade genética, juntamente com os avanços ocorridos na área de genética forense, que é possível distinguir um indivíduo de outro por meio da análise dos perfis genéticos destes.

4.2.1 O Genoma

As regiões do DNA que codificam e regulam a síntese protéica são denominadas genes. Estima-se que o genoma humano contenha entre 20.000 e 25.000 genes e que somente 1,5% do genoma está diretamente envolvido com a codificação de proteínas. Aproximadamente 23,5% do genoma é classificado como seqüência gênica, mas não codificam proteínas (Goodwin et al., 2007).

Cerca de 75% do genoma é extragênico, do qual 21% é uma cópia simples de DNA, cuja função em muitos casos é desconhecida, e pouco mais de 50% é composto de DNA repetido. 45% do DNA repetido é intercalado com elementos completamente dispersos no genoma. A outra classe de elemento repetido é o DNA repetido em tandem (9%). Esta classe pode ainda ser dividida em três diferentes tipos: DNA satélite (5%), minissatélite (3%) e microssatélite (1%).

A Figura 4.3 obtida em Goodwin et al. (2007) mostra graficamente a estrutura do genoma apresentada acima.

Duas importantes categorias de DNA repetido em tandem têm sido amplamente utilizadas em genética forense: minissatélites ou *Variable Number Tandem Repeats* (VNTRs) e microssatélites ou *Short Tandem Repeats* (STRs). A variação entre diferentes alelos é causada pela diferença no número de unidades repetidas das quais resultam alelos com comprimentos diferentes. Devido a isso, polimorfismos na repetição em tandem são classificados como polimorfismos de comprimento.

Através do estudo das seqüências de DNA microssatélites, pode-se obter o perfil genético (*DNA Fingerprinting*) de qualquer indivíduo a partir de uma amostra de um de seus tecidos. Isso só é possível porque o DNA de um indivíduo é sempre igual em qualquer célula de seu organismo. Como não há indivíduos com o mesmo genótipo, exceto em casos de gêmeos idênticos, os perfis de DNA humano fornecem uma poderosa ferramenta na análise de vínculo

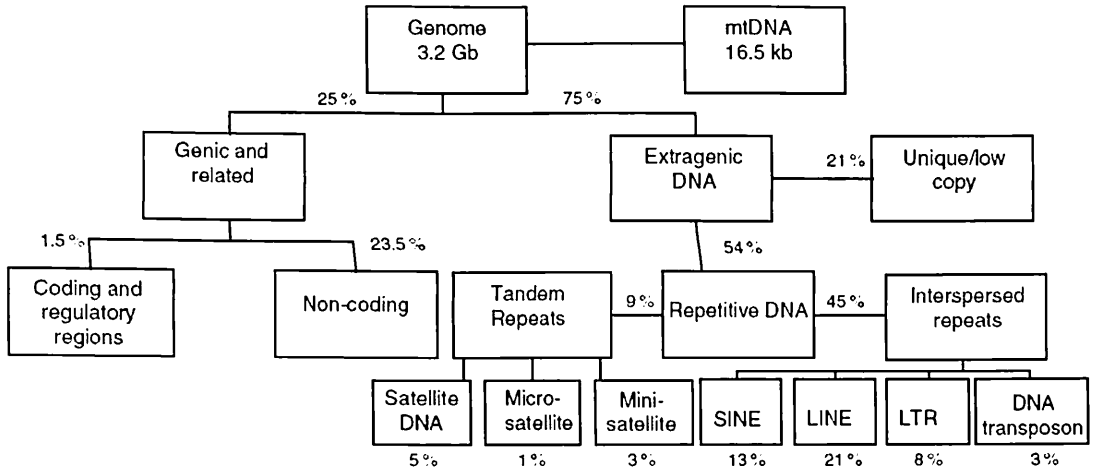

Figura 4.3: Genoma Humano

genético no que tange ao estudo de paternidade, bem como na identificação de criminosos a partir de vestígios deixados no local do crime.

4.3 A Reação em Cadeia de Polimerase e Marcadores STRs

A Reação em Cadeia de Polimerase (PCR) propicia a amplificação seletiva de uma região escolhida de uma molécula de DNA. Com essa técnica, qualquer região da molécula de DNA pode ser selecionada, desde que as seqüências nas extremidades dessa região sejam conhecidas. Para realizar uma PCR, dois pequenos oligonucleotídeos, moléculas sintéticas de DNA que delimitam a região a ser amplificada e atuam como iniciadores para as reações de síntese de DNA, devem hibridizar com a molécula de DNA, um com cada uma das fitas da hélice dupla. A amplificação é realizada pela enzima *Taq DNA Polimerase*, que por ser termo-estável, tem resistência à desnaturação pelo calor.

Para iniciar uma amplificação por PCR, a enzima é adicionada ao DNA-molde anelado aos iniciadores e incubada para que sintetize as novas fitas complementares. Para que as fitas recém-sintetizadas separem-se do molde, a mistura é aquecida a 94°C e posteriormente resfriada, permitindo que mais iniciadores hibridizem com suas respectivas posições. A seguir, a *Taq DNA Polimerase* realiza pela segunda vez a síntese de DNA.

O ciclo desnaturação-hibridização-síntese é repetido várias vezes, resultando na síntese de centenas de milhões de cópias do fragmento de DNA amplificado. A amostra resultante é geralmente analisada por eletroforese em gel

de agarose, processo de separação de moléculas com base na relação entre carga e massa das mesmas realizado numa matriz gelatinosa que permite que moléculas de cargas elétricas similares possam ser separadas com base em seus tamanhos.

O resultado da PCR é usado na análise dos polimorfismos STRs. O amplo uso dos STRs em laboratórios de genética forense se deve à facilidade em encontrá-los em pequenas quantidades de DNA, bem como em restos de DNA de baixa qualidade.

Inicialmente o estudo dos marcadores STRs mediante a técnica de PCR requeria uma análise individual para cada marcador genético, através de reações simples de amplificação. Atualmente as reações de amplificação são realizadas simultaneamente, com vários *loci* de uma única vez. Os trechos amplificados são detectados por meio de uma ferramenta denominada *seqüênciador de eletroforese capilar.*

A Figura 4.4 cedida pelo Laboratório de DNA Forense da Universidade Federal de Alagoas mostra a saída de um seqüenciador de eletroforese capilar para alguns marcadores STRs aplicados a um estudo de paternidade caso padrão, no qual se tem infomações sobre os perfis genéticos da criança, de sua mãe e de seu suposto pai (ver seção 4.5.2).

Na figura supramencionada, pode-se observar que para cada marcador (`D3S1358`, `D13S317`, `D7S820`, `D16S539` e `FGA`) há dois alelos associados à mãe (sigla `44-A-M`), à criança (sigla `44-A-C`) e ao suposto pai (sigla `44-A-SP`). Para o marcador `FGA` por exemplo, os alelos associados à mãe são $19-19$, à criança, $19-22$ e, ao suposto, $21-22$.

4.4 Genética de Populações

4.4.1 Equilíbrio de Hardy-Weinberg

O matemático inglês Godfrey Harold Hardy e o médico alemão Wilhelm Weinberg demonstraram que, no caso de dois alelos, quando os cruzamentos ocorrem de forma aleatória, sem que haja seleção natural, mutação e migração, as freqüências alélicas e genotípicas seguem uma distribuição binomial nas populações de organismos diplóides. Diz-se que a população que satisfaz essas condições encontra-se em *Equilíbrio de Hardy-Weinberg.*

Assumindo que uma dada população se encontra em equilíbrio de Hardy-Weinberg; dados dois alelos A_1 e A_2, estes terão freqüências p_{A_1} e $p_{A_2} = 1 - p_{A_1}$ respectivamente. Dessa forma, tem-se que:

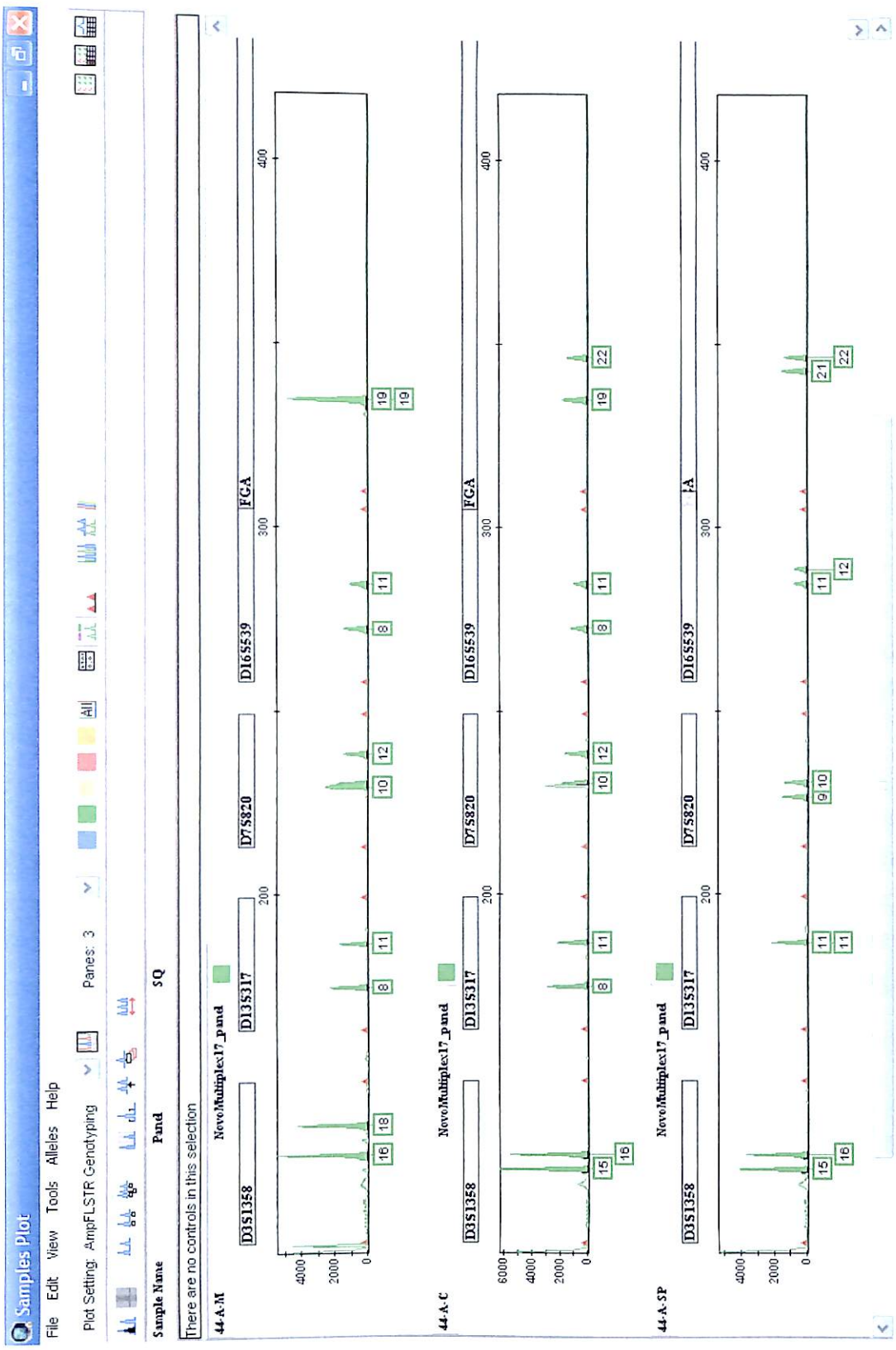

Figura 4.4: Saída de um seqüenciador de eletroforese capilar

- a proporção do genótipo homozigoto A_1A_1 é igual a $p_{A_1}^2$, ou seja, $\Pr(A_1A_1) = p_{A_1}^2$;

- a proporção do genótipo heterozigoto A_1A_2 é igual a $2p_{A_1}p_{A_2}$, ou seja, $Pr(A_1A_2) = 2p_{A_1}p_{A_2}$;
- a proporção do genótipo homozigoto A_2A_2 é igual a $p^2_{A_2}$, ou seja, $Pr(A_2A_2) = p^2_{A_2}$.

Na seção seguinte, é mostrado o procedimento para se chegar às probalibidades acima.

Cálculo das Proporções Genotípicas

Haja vista a necessidade do conhecimento de dois modelos probabilísticos discretos para os cálculos das proporções genotípicas, abaixo seguem as definições dos mesmos obtidas em Magalhães & de Lima (2002).

> ***Definição 9 (Modelo Bernoulli)*** *Uma variável aleatória X segue o modelo de Bernoulli se atribui 0 ou 1 à ocorrência de fracasso ou sucesso respectivamente. Com p representando a probabilidade de sucesso, a função discreta de probabilidade dessa variável é dada por*
>
> $$Pr(X = x) = p^x(1-p)^{1-x},\ x = 0,1.$$

> ***Definição 10 (Modelo binomial)*** *Considere a repetição de n ensaios de Bernoulli independentes e todos com a mesma probabilidade de sucesso p. A distribuição da variável aleatória que conta o número total de sucessos é denominada binomial com parâmetros n e p e sua função de probabilidade é dada por*
>
> $$Pr(X = k) = \binom{n}{k} p^k(1-p)^{n-k},\ k = 0,1,2,\ldots,n,$$
>
> *com $\binom{n}{k}$ representando o coeficiente binomial calculado por*
>
> $$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Considere dois alelos A_1 e A_2 com probabilidades de ocorrência $Pr(A_1) = p_{A_1}$ e $Pr(A_2) = p_{A_2}$ respectivamente e suponha que o sucesso é a ocorrência do alelo A_1 e o fracasso a ocorrência do alelo A_2, tem-se que a variável aleatória X que segue o modelo de Bernoulli tem:

- probabilidade de sucesso dada por $Pr(X = A_1) = p_{A_1}$ e

- probabilidade de fracasso dada por $\Pr(X = A_2) = p_{A_2} = 1 - p_{A_1}$.

Supondo a repetição de 2 ensaios de Bernoulli independentes da variável aleatória X definida acima, constrói-se uma nova variável aleatória W que segue o modelo Binomial, servindo, pois, para contar o número de sucessos nessa repetição. Com essa repetição dos $n = 2$ ensaios, há três configurações (genótipos) possíveis: A_1A_1 (ocorrência de dois sucessos $\rightarrowtail k = 2$), A_1A_2 (ocorrência de um sucesso $\rightarrowtail k = 1$) e A_2A_2 (não ocorre sucesso $\rightarrowtail k = 0$). Dessa forma, tem-se que o modelo Binomial da variável W é dado por

$$\Pr(W = k) = \binom{2}{k} p_{A_1}^k (1 - p_{A_1})^{2-k},\ k = 0, 1, 2$$

ou

$$\Pr(W = k) = \binom{2}{k} p_{A_1}^k p_{A_2}^{2-k},\ k = 0, 1, 2.$$

Com o modelo construído, efetuam-se os cálculos.

1. Proporção do genótipo homozigoto $A_1A_1 \rightarrowtail \Pr(W = 2)$

$$\Pr(W = 2) = \binom{2}{2} p_{A_1}^2 p_{A_2}^{2-2} = 1 p_{A_1}^2 p_{A_2}^0 = 1 p_{A_1}^2 1 = p_{A_1}^2$$

2. Proporção do genótipo heterozigoto $A_1A_2 \rightarrowtail \Pr(W = 1)$

$$\Pr(W = 1) = \binom{2}{1} p_{A_1}^1 p_{A_2}^{2-1} = 2 p_{A_1} p_{A_2}^1 = 2 p_{A_1} p_{A_2}$$

3. Proporção do genótipo homozigoto $A_2A_2 \rightarrowtail \Pr(W = 0)$

$$\Pr(W = 0) = \binom{2}{0} p_{A_1}^0 p_{A_2}^{2-0} = 1 p_{A_1}^0 p_{A_2}^2 = p_{A_2}^2$$

4.4.2 Freqüências Alélicas

Nesta seção, será mostrado que as proporções alélicas da prole numa população em Equilíbrio de Hardy-Weinberg são $(p_{A_1}^2 : 2p_{A_1}p_{A_2} : p_{A_2}^2)$, considerando os mesmos alelos apresentados anteriormente: A_1 com freqüência p_{A_1} e A_2 com freqüência p_{A_2}. Observe que o cruzamento resulta em nove tipos possíveis de acasalamento, cujas probabilidades são mostradas na Tabela 4.1.

Quando o pai e a mãe são heterozigotos (A_1A_2), a prole pode ter genótipo A_1A_1 (com proporção 1/4), A_1A_2 (com proporção 1/2) e A_2A_2 (com proporção 1/4) conforme mostra o diagrama da Figura 4.5.

a seguir são mostrados os cálculos das probabilidades de G_C ser igual a cada um dos eventos de Ω.

1. Proporção do genótipo homozigoto $A_1A_1 \rightarrowtail \Pr(G_C = A_1A_1)$

$$
\begin{aligned}
\Pr(G_C = A_1A_1) &= \sum_{\omega_1=\Omega}\sum_{\omega_2=\Omega} \Pr(G_C = A_1A_1 \cap G_P = \omega_1 \cap G_M = \omega_2) \\
&= \Pr(G_C = A_1A_1 \cap G_P = A_1A_1 \cap G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_1A_1 \cap G_P = A_1A_1 \cap G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_1A_1 \cap G_P = A_1A_1 \cap G_M = A_2A_2) + \\
&\quad \Pr(G_C = A_1A_1 \cap G_P = A_1A_2 \cap G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_1A_1 \cap G_P = A_1A_2 \cap G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_1A_1 \cap G_P = A_1A_2 \cap G_M = A_2A_2) + \\
&\quad \Pr(G_C = A_1A_1 \cap G_P = A_2A_2 \cap G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_1A_1 \cap G_P = A_2A_2 \cap G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_1A_1 \cap G_P = A_2A_2 \cap G_M = A_2A_2) \\
&= \Pr(G_C = A_1A_1 \mid G_P = A_1A_1 \cap G_M = A_1A_1) \times \Pr(G_P = A_1A_1) \times \Pr(G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_1A_1 \mid G_P = A_1A_1 \cap G_M = A_1A_2) \times \Pr(G_P = A_1A_1) \times \Pr(G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_1A_1 \mid G_P = A_1A_1 \cap G_M = A_2A_2) \times \Pr(G_P = A_1A_1) \times \Pr(G_M = A_2A_2) + \\
&\quad \Pr(G_C = A_1A_1 \mid G_P = A_1A_2 \cap G_M = A_1A_1) \times \Pr(G_P = A_1A_2) \times \Pr(G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_1A_1 \mid G_P = A_1A_2 \cap G_M = A_1A_2) \times \Pr(G_P = A_1A_2) \times \Pr(G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_1A_1 \mid G_P = A_1A_2 \cap G_M = A_2A_2) \times \Pr(G_P = A_1A_2) \times \Pr(G_M = A_2A_2) + \\
&\quad \Pr(G_C = A_1A_1 \mid G_P = A_2A_2 \cap G_M = A_1A_1) \times \Pr(G_P = A_2A_2) \times \Pr(G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_1A_1 \mid G_P = A_2A_2 \cap G_M = A_1A_2) \times \Pr(G_P = A_2A_2) \times \Pr(G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_1A_1 \mid G_P = A_2A_2 \cap G_M = A_2A_2) \times \Pr(G_P = A_2A_2) \times \Pr(G_M = A_2A_2) \\
&= (1 \times p_{A_1}^2 \times p_{A_1}^2) + (1/2 \times p_{A_1}^2 \times 2p_{A_1}p_{A_2}) + (0 \times p_{A_1}^2 \times p_{A_2}^2) + \\
&\quad (1/2 \times 2p_{A_1}p_{A_2} \times p_{A_1}^2) + (1/4 \times 2p_{A_1}p_{A_2} \times 2p_{A_1}p_{A_2}) + (0 \times 2p_{A_1}p_{A_2} \times p_{A_2}^2) + \\
&\quad (0 \times p_{A_2}^2 \times p_{A_1}^2) + (0 \times p_{A_2}^2 \times 2p_{A_1}p_{A_2}) + (0 \times p_{A_2}^2 \times p_{A_2}^2) \\
&= p_{A_1}^4 + p_{A_1}^3 p_{A_2} + p_{A_1}^3 p_{A_2} + p_{A_1}^2 p_{A_2}^2 \\
&= p_{A_1}^4 + 2p_{A_1}^3 p_{A_2} + p_{A_1}^2 p_{A_2}^2 \\
&= p_{A_1}^2 (p_{A_1}^2 + 2p_{A_1}p_{A_2} + p_{A_2}^2) \\
&= p_{A_1}^2 (p_{A_1} + p_{A_2})^2 \\
&= p_{A_1}^2 [p_{A_1} + (1 - p_{A_1})]^2 \\
&= p_{A_1}^2 (1)^2 \\
&= p_{A_1}^2
\end{aligned}
$$

Mãe	Pai	Probabilidade	A_1A_1	A_1A_2	A_2A_2
A_1A_1	A_1A_1	$p_{A_1}^2 \cdot p_{A_1}^2 = p_{A_1}^4$	$p_{A_1}^4$	0	0
A_1A_1	A_1A_2	$p_{A_1}^2 \cdot 2p_{A_1}p_{A_2} = 2p_{A_1}^3 p_{A_2}$	$\frac{1}{2}(2p_{A_1}^3 p_{A_2})$	$\frac{1}{2}(2p_{A_1}^3 p_{A_2})$	0
A_1A_2	A_1A_1	$2p_{A_1}p_{A_2} \cdot p_{A_1}^2 = 2p_{A_1}^3 p_{A_2}$	$\frac{1}{2}(2p_{A_1}^3 p_{A_2})$	$\frac{1}{2}(2p_{A_1}^3 p_{A_2})$	0
A_1A_1	A_2A_2	$p_{A_1}^2 \cdot p_{A_2}^2 = p_{A_1}^2 p_{A_2}^2$	0	$p_{A_1}^2 p_{A_2}^2$	0
A_2A_2	A_1A_1	$p_{A_2}^2 \cdot p_{A_1}^2 = p_{A_1}^2 p_{A_2}^2$	0	$p_{A_1}^2 p_{A_2}^2$	0
A_1A_2	A_1A_2	$2p_{A_1}p_{A_2} \cdot 2p_{A_1}p_{A_2} = 4p_{A_1}^2 p_{A_2}^2$	$\frac{1}{4}(4p_{A_1}^2 p_{A_2}^2)$	$\frac{1}{2}(4p_{A_1}^2 p_{A_2}^2)$	$\frac{1}{4}(4p_{A_1}^2 p_{A_2}^2)$
A_1A_2	A_2A_2	$2p_{A_1}p_{A_2} \cdot p_{A_2}^2 = 2p_{A_1}p_{A_2}^3$	0	$\frac{1}{2}(2p_{A_1}p_{A_2}^3)$	$\frac{1}{2}(2p_{A_1}p_{A_2}^3)$
A_2A_2	A_1A_2	$p_{A_2}^2 \cdot 2p_{A_1}p_{A_2} = 2p_{A_1}p_{A_2}^3$	0	$\frac{1}{2}(2p_{A_1}p_{A_2}^3)$	$\frac{1}{2}(2p_{A_1}p_{A_2}^3)$
A_2A_2	A_2A_2	$p_{A_2}^2 \cdot p_{A_2}^2 = p_{A_2}^4$	0	0	$p_{A_2}^4$
		Soma da Prole	$p_{A_1}^2$	$2p_{A_1}p_{A_2}$	$p_{A_2}^2$

Tabela 4.1: Proporções alélicas da prole numa população em Equilíbrio de Hardy-Weinberg

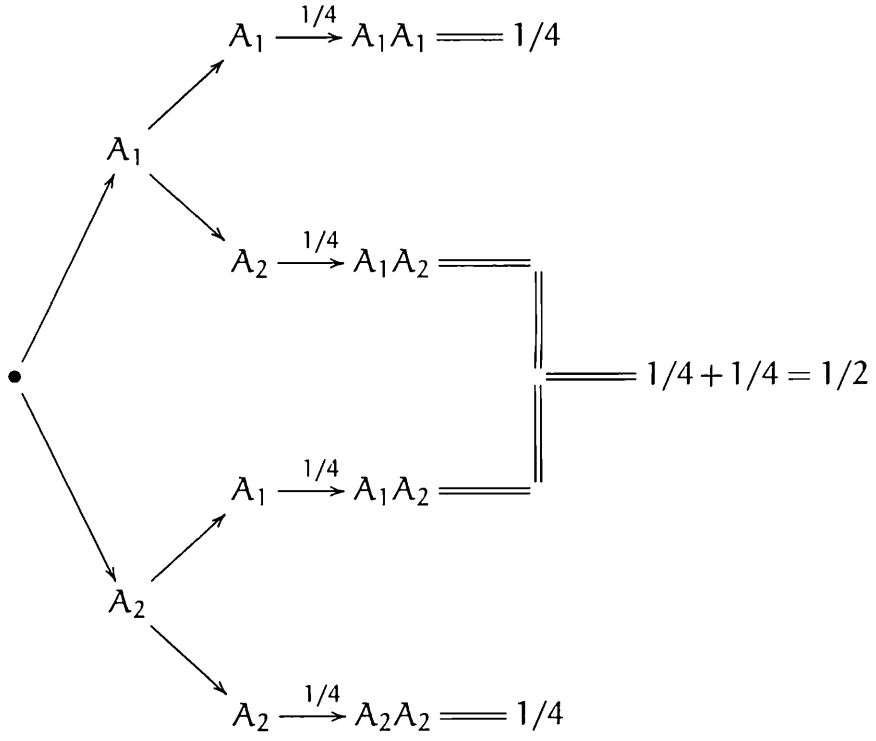

Figura 4.5: Proporções genotípicas

Cálculo das Proporções Alélicas da Prole

Sendo G_C, G_P, e G_M as variáveis aleatórias que representam os genótipos da prole, pai e mãe respectivamente, tem-se o seguinte espaço amostral $\Omega = \{A_1A_1, A_1A_2, A_2A_2\}$ para elas.

Sabendo que

$$\begin{aligned} \Pr(G_C \cap G_P \cap G_M) &= \Pr(G_C \mid G_P \cap G_M) \times \Pr(G_P \cap G_M) \\ &= \Pr(G_C \mid G_P \cap G_M) \times \Pr(G_P) \times \Pr(G_M), \end{aligned}$$

2. Proporção do genótipo heterozigoto $A_1A_2 \rightarrowtail \Pr(G_C = A_1A_2)$

$$
\begin{aligned}
\Pr(G_C = A_1A_2) &= \sum_{\omega_1=\Omega}\sum_{\omega_2=\Omega} \Pr(G_C = A_1A_2 \cap G_P = \omega_1 \cap G_M = \omega_2) \\
&= \Pr(G_C = A_1A_2 \cap G_P = A_1A_1 \cap G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_1A_2 \cap G_P = A_1A_1 \cap G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_1A_2 \cap G_P = A_1A_1 \cap G_M = A_2A_2) + \\
&\quad \Pr(G_C = A_1A_2 \cap G_P = A_1A_2 \cap G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_1A_2 \cap G_P = A_1A_2 \cap G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_1A_2 \cap G_P = A_1A_2 \cap G_M = A_2A_2) + \\
&\quad \Pr(G_C = A_1A_2 \cap G_P = A_2A_2 \cap G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_1A_2 \cap G_P = A_2A_2 \cap G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_1A_2 \cap G_P = A_2A_2 \cap G_M = A_2A_2) \\
&= \Pr(G_C = A_1A_2 \mid G_P = A_1A_1 \cap G_M = A_1A_1) \times \Pr(G_P = A_1A_1) \times \Pr(G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_1A_2 \mid G_P = A_1A_1 \cap G_M = A_1A_2) \times \Pr(G_P = A_1A_1) \times \Pr(G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_1A_2 \mid G_P = A_1A_1 \cap G_M = A_2A_2) \times \Pr(G_P = A_1A_1) \times \Pr(G_M = A_2A_2) + \\
&\quad \Pr(G_C = A_1A_2 \mid G_P = A_1A_2 \cap G_M = A_1A_1) \times \Pr(G_P = A_1A_2) \times \Pr(G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_1A_2 \mid G_P = A_1A_2 \cap G_M = A_1A_2) \times \Pr(G_P = A_1A_2) \times \Pr(G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_1A_2 \mid G_P = A_1A_2 \cap G_M = A_2A_2) \times \Pr(G_P = A_1A_2) \times \Pr(G_M = A_2A_2) + \\
&\quad \Pr(G_C = A_1A_2 \mid G_P = A_2A_2 \cap G_M = A_1A_1) \times \Pr(G_P = A_2A_2) \times \Pr(G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_1A_2 \mid G_P = A_2A_2 \cap G_M = A_1A_2) \times \Pr(G_P = A_2A_2) \times \Pr(G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_1A_2 \mid G_P = A_2A_2 \cap G_M = A_2A_2) \times \Pr(G_P = A_2A_2) \times \Pr(G_M = A_2A_2) \\
&= (0 \times p_{A_1}^2 \times p_{A_1}^2) + (1/2 \times p_{A_1}^2 \times 2p_{A_1}p_{A_2}) + (1 \times p_{A_1}^2 \times p_{A_2}^2) + \\
&\quad (1/2 \times 2p_{A_1}p_{A_2} \times p_{A_1}^2) + (1/2 \times 2p_{A_1}p_{A_2} \times 2p_{A_1}p_{A_2}) + (1/2 \times 2p_{A_1}p_{A_2} \times p_{A_2}^2) + \\
&\quad (1 \times p_{A_2}^2 \times p_{A_1}^2) + (1/2 \times p_{A_2}^2 \times 2p_{A_1}p_{A_2}) + (0 \times p_{A_2}^2 \times p_{A_2}^2) \\
&= p_{A_1}^3 p_{A_2} + p_{A_1}^2 p_{A_2}^2 + p_{A_1}^3 p_{A_2} + 2p_{A_1}^2 p_{A_2}^2 + p_{A_1} p_{A_2}^3 + p_{A_1}^2 p_{A_2}^2 + p_{A_1} p_{A_2}^3 \\
&= 2p_{A_1}^3 p_{A_2} + 4p_{A_1}^2 p_{A_2}^2 + 2p_{A_1} p_{A_2}^3 \\
&= 2p_{A_1}p_{A_2}(p_{A_1}^2 + 2p_{A_1}p_{A_2} + p_{A_2}^2) \\
&= 2p_{A_1}p_{A_2}(p_{A_1} + p_{A_2})^2 \\
&= 2p_{A_1}p_{A_2}[p_{A_1} + (1 - p_{A_1})]^2 \\
&= 2p_{A_1}p_{A_2}(1)^2 \\
&= 2p_{A_1}p_{A_2}
\end{aligned}
$$

3. Proporção do genótipo homozigoto $A_2A_2 \rightarrowtail \Pr(G_C = A_2A_2)$

$$
\begin{aligned}
\Pr(G_C = A_2A_2) &= \sum_{\omega_1=\Omega}\sum_{\omega_2=\Omega} \Pr(G_C = A_2A_2 \cap G_P = \omega_1 \cap G_M = \omega_2) \\
&= \Pr(G_C = A_2A_2 \cap G_P = A_1A_1 \cap G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_2A_2 \cap G_P = A_1A_1 \cap G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_2A_2 \cap G_P = A_1A_1 \cap G_M = A_2A_2) + \\
&\quad \Pr(G_C = A_2A_2 \cap G_P = A_1A_2 \cap G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_2A_2 \cap G_P = A_1A_2 \cap G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_2A_2 \cap G_P = A_1A_2 \cap G_M = A_2A_2) + \\
&\quad \Pr(G_C = A_2A_2 \cap G_P = A_2A_2 \cap G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_2A_2 \cap G_P = A_2A_2 \cap G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_2A_2 \cap G_P = A_2A_2 \cap G_M = A_2A_2) \\
&= \Pr(G_C = A_2A_2 \mid G_P = A_1A_1 \cap G_M = A_1A_1) \times \Pr(G_P = A_1A_1) \times \Pr(G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_2A_2 \mid G_P = A_1A_1 \cap G_M = A_1A_2) \times \Pr(G_P = A_1A_1) \times \Pr(G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_2A_2 \mid G_P = A_1A_1 \cap G_M = A_2A_2) \times \Pr(G_P = A_1A_1) \times \Pr(G_M = A_2A_2) + \\
&\quad \Pr(G_C = A_2A_2 \mid G_P = A_1A_2 \cap G_M = A_1A_1) \times \Pr(G_P = A_1A_2) \times \Pr(G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_2A_2 \mid G_P = A_1A_2 \cap G_M = A_1A_2) \times \Pr(G_P = A_1A_2) \times \Pr(G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_2A_2 \mid G_P = A_1A_2 \cap G_M = A_2A_2) \times \Pr(G_P = A_1A_2) \times \Pr(G_M = A_2A_2) + \\
&\quad \Pr(G_C = A_2A_2 \mid G_P = A_2A_2 \cap G_M = A_1A_1) \times \Pr(G_P = A_2A_2) \times \Pr(G_M = A_1A_1) + \\
&\quad \Pr(G_C = A_2A_2 \mid G_P = A_2A_2 \cap G_M = A_1A_2) \times \Pr(G_P = A_2A_2) \times \Pr(G_M = A_1A_2) + \\
&\quad \Pr(G_C = A_2A_2 \mid G_P = A_2A_2 \cap G_M = A_2A_2) \times \Pr(G_P = A_2A_2) \times \Pr(G_M = A_2A_2) \\
&= (0 \times p_{A_1}^2 \times p_{A_1}^2) + (0 \times p_{A_1}^2 \times 2p_{A_1}p_{A_2}) + (0 \times p_{A_1}^2 \times p_{A_2}^2) + \\
&\quad (0 \times 2p_{A_1}p_{A_2} \times p_{A_1}^2) + (1/4 \times 2p_{A_1}p_{A_2} \times 2p_{A_1}p_{A_2}) + (1/2 \times 2p_{A_1}p_{A_2} \times p_{A_2}^2) + \\
&\quad (0 \times p_{A_2}^2 \times p_{A_1}^2) + (1/2 \times p_{A_2}^2 \times 2p_{A_1}p_{A_2}) + (1 \times p_{A_2}^2 \times p_{A_2}^2) \\
&= p_{A_1}^2 p_{A_2}^2 + p_{A_1}p_{A_2}^3 + p_{A_1}p_{A_2}^3 + p_{A_2}^4 \\
&= p_{A_4}^4 + 2p_{A_1}p_{A_2}^3 + p_{A_1}^2 p_{A_2}^2 \\
&= p_{A_2}^2(p_{A_2}^2 + 2p_{A_2}p_{A_1} + p_{A_1}^2) \\
&= p_{A_2}^2(p_{A_2} + p_{A_1})^2 \\
&= p_{A_2}^2[(1 - p_{A_1}) + p_{A_1}]^2 \\
&= p_{A_2}^2(1)^2 \\
&= p_{A_2}^2
\end{aligned}
$$

4.5 Análise de Vínculo Genético

Nesta seção, será mostrado como é feita a análise de vínculo genético no que tange ao estudo de paternidade.

4.5.1 Considerações Preliminares

A aplicação de perfis de DNA para estudos de paternidade oferece um modo fácil de estabelecer vínculos genéticos. Desde o primeiro teste de paternidade ocorrido em 1985, a análise de DNA vem sendo amplamente utilizada na verificação desse tipo de vínculo (Butler, 2005).

Um fator que contribui para a ampla utilização de perfis de DNA em estudos desse tipo são os avanços que a genética forense tem vivenciado nos últimos anos, em especial com a técnica de PCR, dispondo-se hoje de um conjunto de instrumentos de marcadores muito amplos, altamente polimórficos e padronizados.

Com todo esse aparato, é possível abordar com êxito a maioria das investigações de paternidade. No entanto, uma abordagem estatística é fundamental, visto que nos domínios gerados na análise forense de DNA em estudos de paternidade, há um certo grau de incerteza, pois apesar de cada ser humano possuir um perfil genético único, em estudos forenses é praticamente impossível se analisar todo o genótipo dos indivíduos envolvidos, havendo, pois, a necessidade de se utilizar inferências probabilísticas na análise dos dados. Além disso, em muitos casos é necessário realizar estudos de paternidade sem ter, por exemplo, o perfil genético do suposto pai, conforme será discutido mais adiante.

A metodologia usada pelos laboratórios de genética forense na produção de perfis genéticos para estudo de paternidade é idêntica às análises de material genético deixado em locais de crime. Todavia, a interpretação dos resultados em estudos de paternidade é mais complexa, haja vista que no caso da identificação de criminosos é feita apenas a comparação entre os perfis encontrados na cena do crime e os perfis dos suspeitos (Butler, 2005). Além disso, em muitos casos é necessário realizar estudos de paternidade sem ter, por exemplo, o perfil genético do suposto pai, tendo, pois, que trabalhar sob a incerteza. Contudo, com a posse dos perfis genéticos de familiares desse indivíduo, é possível realizar a análise e, a depender da quantidade e dos indivíduos envolvidos nesse estudo, ter um resultado proveitoso.

4.5.2 Estudo de Paternidade

A prova biológica requer estudos genéticos e a Genética de Populações da população de referência, sendo necessário haver a comprovação do equilíbrio de Hardy-Weinberg no que tange à independência do *locus*. O estudo compreende o cálculo dos seguintes valores:

Índice de Paternidade (IP) Corresponde à razão entre a probabilidade do suposto pai ser o pai biológico da criança dadas as evidências pela probabilidade do pai biológico da criança ser outro indivíduo da população dadas as mesmas evidências. Sendo E o conjunto de evidências, H o evento que representa a hipótese do suposto pai ser o pai biológico da criança e H^c

o evento que representa a hipótese do pai biológico da criança ser outro indivíduo da população, tem-se:

$$IP = \frac{Pr(H \mid E)}{Pr(H^c \mid E)} \tag{4.1}$$

Pela equação (2.4) da página 19, a equação acima pode ser reescrita da seguinte forma:

$$IP = \frac{Pr(H \mid E)}{Pr(H^c \mid E)} = \frac{Pr(E \mid H)}{Pr(E \mid H^c)} \cdot \frac{Pr(H)}{Pr(H^c)} = \mathcal{L}(E \mid H)O(H) \tag{4.2}$$

Dessa forma, o IP, que é o produto entre a 'razão de verossimilhança' ($Pr(e \mid H)/Pr(e \mid H^c)$) e as 'chances relativas da priori' ($Pr(H)/Pr(H^c)$), pode ser chamado de 'chances a posteriori' da hipótese H tendo sido observado o conjunto de evidências E. É importante mencionar que o conjunto de evidências E corresponde exatamente aos genótipos dos indivíduos envolvidos no estudo e que os eventos H e H^c são complementares, ou seja, $Pr(H) = 1 - Pr(H^c)$.

Outra denominação atribuída ao IP é 'razão de verossimilhança', pois é assumido que $Pr(H) = 0.5$ e, por conseguinte, as 'chances relativas da priori' é igual a 1. Portanto, a equação (4.2) pode ser reescrita como segue:

$$IP = \frac{Pr(H \mid E)}{Pr(H^c \mid E)} = \frac{Pr(E \mid H)}{Pr(E \mid H^c)} \tag{4.3}$$

É importante ressaltar que o IP é calculado para cada *locus*.

Índice de Paternidade Acumulado (IPC) Corresponde ao produtório dos IPs de cada *locus* conforme mostrado na equação abaixo, onde n é o número dos *loci* analisados.

$$IPC = \prod_{i=1}^{n} IP_i \tag{4.4}$$

Probabilidade de Exclusão (PE) Corresponde à probabilidade de um indivíduo que fora falsamente acusado ser excluído como pai biológico da criança conforme mostrado na equação abaixo, onde p_{A_1} e p_{A_2} são as freqüências dos alelos A_1 e A_2 da criança.

$$PE = [1 - (p_{A_1} + p_{A_2})]^2 \tag{4.5}$$

Probabilidade de Exclusão Acumulada (PEC) Corresponde à probabilidade de um indivíduo que fora falsamente acusado ser excluído como pai bio-

lógico da criança ao se analisar n *loci* conforme mostrado na equação abaixo.

$$PEC = 1 - \left[\prod_{i=1}^{n}(1 - PE_{locus_i})\right] \cdot 100 \tag{4.6}$$

Probabilidade de Paternidade (PP) É a probabiliade do suposto pai ser o pai biológico da criança, correspondendo à razão entre o IPC e este incrementado de uma unidade conforme mostrado na equação abaixo.

$$PP = \frac{IPC}{IPC + 1} \tag{4.7}$$

Haja vista a prévia apresentação dos conceitos de suma importância ao entendimento das análises de vínculo genético em se tratando de estudos de paternidade, na seção que segue será mostrado como é feito o cálculo do IP no caso padrão (ver genealogia apresentada na Figura 4.6), no qual tem-se os genótipos da criança (G_C), de sua mãe biológica (G_M) e de seu suposto pai (G_{SP}). No capítulo subseqüente, é apresentada a aplicação das redes bayesianas para esse mesmo tipo de estudo.

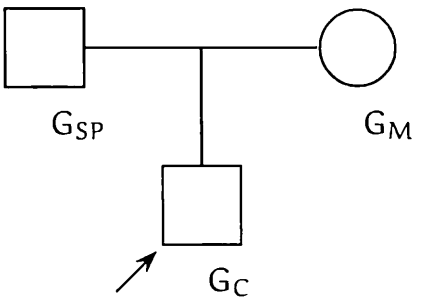

Figura 4.6: Genealogia caso padrão

Cálculo do IP para o Caso Padrão

Assumindo que para um alelo A_i com $i \in \mathbb{N}^*_+$, tem-se $Pr(A_i) = p_i$, são mostrados na Tabela 4.2 os valores do IP dado o conjunto de evidências, ou seja, os genótipos da criança (G_C), de sua mãe biológica (G_M) e de seu suposto pai (G_{SP}).

Analisando o caso em que a criança, sua mãe biológica e seu suposto pai são todos heterozigotos (A_1A_2), tem-se pela Tabela 4.2 que o IP é dado por $1/(p_1 + p_2)$. Abaixo é mostrado o cálculo desse IP, cujo procedimento pode ser aplicado para se calcular o IP dado qualquer conjunto de evidências (genótipos da criança, mãe e suposto pai).

No caso em questão, tem-se que analisar as evidências e calcular as probabilidades $Pr(H \mid E)$ e $Pr(H^c \mid E)$ para, só então, calcular o IP.

G_C	G_M	G_{SP}	IP
A_1A_1	A_1A_1, A_1A_2	A_1A_1	$\frac{1}{p_1}$
		A_1A_2, A_1A_k	$\frac{1}{2p_1}$
A_1A_2	A_1A_1, A_1A_k	A_2A_2	$\frac{1}{p_2}$
		A_1A_2, A_2A_k, A_2A_l	$\frac{1}{2p_2}$
A_1A_2	A_1A_2	A_1A_1, A_2A_2, A_1A_2	$\frac{1}{(p_1+p_2)}$
		A_1A_k, A_2A_k	$\frac{1}{2(p_1+p_2)}$

Tabela 4.2: Valor do IP no caso padrão

1. **Evidências:**

   - $G_C = A_1A_2$
   - $G_M = A_1A_2$
   - $G_{SP} = A_1A_2$

2. **A probabilidade do suposto pai ser o pai biológico da criança dadas as evidências** ($Pr(H \mid E)$)**:**
   Desse cruzamento há 3 valores possíveis para o genótipo da criança: A_1A_1 com probabilidade 1/4, A_1A_2 com probabilidade 1/2 e A_2A_2 com probabilidade 1/4. O diagrama da Figura 4.7 mostra como foram obtidas essas probabilidades.

   Logo, $Pr(H \mid E) = 1/2$.

3. **A pobabilidade do pai biológico da criança ser outro indivíduo da população dadas as evidências** ($Pr(H^c \mid E)$)**:**
   Essa probabilidade é a soma das probabilidades da mãe passar cada um dos alelos para a criança, sendo o outro passado por um indivíduo qualquer da população diferente do suposto pai, ou seja,

$$\begin{aligned} Pr(H^c \mid E) &= Pr(\text{da criança herdar } A_1 \text{ da mãe e } A_2 \text{ de um indivíduo qualquer}) + \\ &\quad Pr(\text{da criança herdar } A_2 \text{ da mãe e } A_1 \text{ de um indivíduo qualquer}) \\ &= 0.5p_{A_2} + 0.5p_{A_1} \\ &= \frac{p_{A_1} + p_{A_2}}{2}. \end{aligned}$$

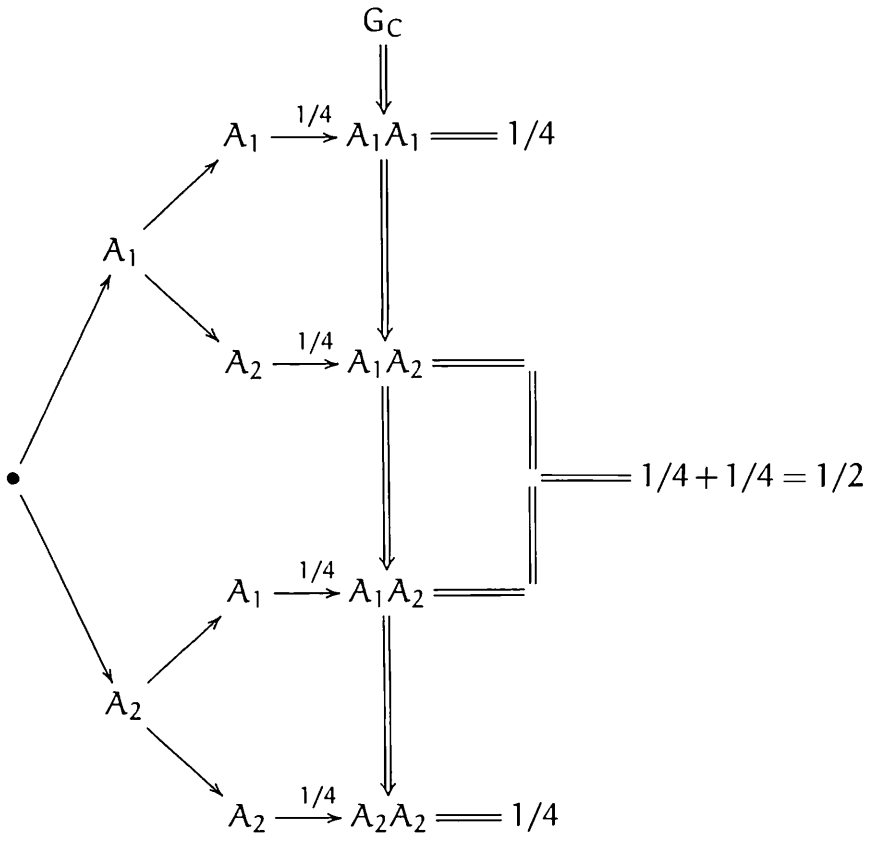

Figura 4.7: Probabilidades dos genótipos da criança

Com essas informações é possível calcular o IP a partir da equação (4.1).

$$\begin{aligned} IP &= \frac{Pr(H \mid E)}{Pr(H^c \mid E)} \\ &= \frac{\frac{1}{2}}{\left(\frac{p_{A_1} + p_{A_2}}{2}\right)} \\ &= \left(\frac{1}{2}\right) \times \left(\frac{2}{p_{A_1} + p_{A_2}}\right) \\ &= \frac{1}{(p_{A_1} + p_{A_2})} \end{aligned}$$

Observação: Em muitos casos de estudo de paternidade, não se possui o perfil genético do suposto pai, sendo necessário, pois, obter o perfil genético desse indivíduo com base nos genótipos de seus familiares para, só então, efetuar o cálculo do IP aplicando os procedimentos acima apresentados. No entanto, conforme será visto no próximo capítulo, ao se utilizar as redes bayesianas para modelar o domínio em questão, uma vez construído o modelo, é preciso apenas inserir as evidências e inferir nesse modelo probabilístico, o qual se encarregará de encontrar o perfil genético do suposto pai, bem como calcular a probabilidade desse indivíduo ser o pai

biológico da criança e a probabilidade do pai biológico ser outro indivíduo da população.

A genealogia da Figura 4.8 representa um estudo de paternidade em que não se tem o perfil genético do suposto pai da criança (G_{SP}), todavia tem-se os perfis genéticos dos pais desse indivíduo (G_{PSP} e G_{MSP}), da criança (G_C) e da mãe desta criança (G_M).

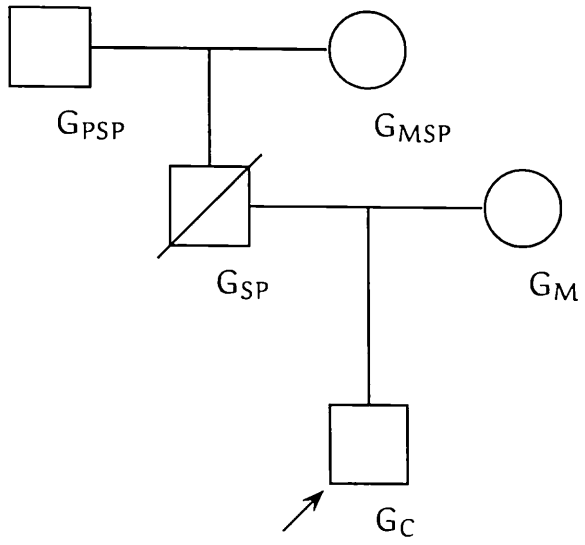

Figura 4.8: Genealogia caso complexo

4.6 Comentários

No presente capítulo, foram abordados alguns conceitos de Genética que serão úteis no entendimento da análise de vínculo genético no que tange ao estudo de paternidade, tema do próximo capítulo. Dentre os conceitos aqui discutidos, estão: genoma humano, reação em cadeia de polimerase, equilíbrio de Hardy-Weinberg, dentre outros.

Capítulo 5

Aplicacão das Redes Bayesianas na Análise Forense de DNA

Haja vista todo o embasamento teórico dado nos capítulos anteriores no que tange às redes bayesianas e à análise forense de DNA; no presente capítulo, será construído um modelo usando esse formalismo matemático na análise de vínculo genético em estudos de paternidade.

É importante ressaltar que o sistema THÊMIS (ver Capítulo 6) utiliza esse modelo para realizar inferências a partir das evidências que se dispõe para o estudo, seja ele o caso padrão (ver Figura 4.6 na página 49), em que se tem evidências (genótipos) da criança, de sua mãe e de seu suposto pai, ou um tipo de caso complexo (ver Figura 4.8 na página 52), em que se tem apenas os genótipos da criança, de sua mãe e dos pais de seu suposto pai.

5.1 Descrição do Problema

Conforme visto no Capítulo 4, há na espécie humana 46 cromossomos, formando 23 pares: 22 pares autossômicos e 1 par envolvido na determinação do sexo.

Nos cromossomos, há regiões denominadas ***locus*** onde é possível encontrar um gene ou uma seqüência de nucleotídeos não-codificadores. Em cromossomos homólogos[1], os genes localizados no mesmo *locus* são chamados genes alelos ou **alelos** e são responsáveis pela mesma característica genética.

Indivíduos heterozigotos para uma dada característica possuem alelos diferentes no *locus* em questão, ao passo que indíviduos homozigotos possuem alelos iguais. A configuração desses alelos corresponde ao **genótipo** do *locus*, sendo o genótipo de um indivíduo o conjunto dos genótipos de seus *loci*.

[1] Cromossomos que se alinham durante a meiose.

A reprodução humana, que é sexuada, está intimamente ligada a dois processos: meiose e fecundação.

Por meiose, o número diplóide de cromossomos (2n) é reduzido à metade (n – haplóide), sendo restabelecido o número 2n típico da espécie pela fecundação. Dessa forma, *em cada par de cromossomos homólogos encontrado em um indivíduo, o genótipo de cada locus é constituído por um alelo proveniente do pai e o outro herdado da mãe.*

5.2 Construção do Modelo

Haja vista a definição de rede bayesiana, bem como a descrição do problema, é possível construir um modelo utilizando esse formalismo para representar o conhecimento acerca do domínio em questão.

Em se tratando do caso padrão, ao se analisar os marcadores STRs no sangue coletado dos indivíduos envolvidos (criança, mãe e suposto pai), o qual fora amplificado via PCR, tem-se para cada **marcador**:

- dois alelos do suposto pai da criança, um herdado do pai dele (modelado pela variável aleatória pppg) e o outro da mãe (modelado pela variável aleatória ppmg), sendo o genótipo desse *locus* modelado pela variável aleatória pgen;
- dois alelos da mãe biológica da criança, um herdado do pai dela (modelado pela variável aleatória mpg) e o outro da mãe (modelado pela variável aleatória mmg), sendo o genótipo desse *locus* modelado pela variável aleatória mgen;
- dois alelos da criança, um herdado do pai (modelado pela variável aleatória cpg) e o outro da mãe (modelado pela variável aleatória cmg), sendo o genótipo desse *locus* modelado pela variável aleatória cgen.

É importante ressaltar que para cada população, há freqüências associadas aos alelos dado um marcador. Por exemplo, a freqüência do alelo 6 para o marcador TPOX pode ser 0,016 no Estado de Alagoas e 0,020 no Estado de Sergipe. Essas freqüências alélicas correspondem à probabilidade de ocorrência do alelo para um certo marcador dada uma população. Portanto, a soma das freqüências dos alelos para cada marcador numa dada população deve totalizar um.

Com base nas informações acima, observa-se que para cada marcador:

- as variáveis aleatórias pppg e ppmg dependem respectivamente das variáveis aleatórias que modelam o genótipo do pai e da mãe do suposto pai da criança. Como em estudos de paternidade caso padrão tem-se o genótipo do suposto pai, essas últimas variáveis são irrelevantes ao cálculo. Sendo assim, pppg e ppmg são regidas por uma lei *a priori*;

- a variável aleatória pgen depende das variáveis aleatórias pppg e ppmg, sendo, pois, regida por uma lei *a posteriori*;

- as variáveis aleatórias mpg e mmg dependem respectivamente das variáveis aleatórias que modelam o genótipo do avô e da avó materna da criança. Como em estudos de paternidade caso padrão tem-se o genótipo da mãe, essas últimas variáveis são irrelevantes ao cálculo. Sendo assim, mpg e mmg são regidas por uma lei *a priori*;

- a variável aleatória mgen depende das variáveis aleatórias mpg e mmg, sendo, pois, regida por uma lei *a posteriori*;

- a variável aleatória cpg depende das variáveis aleatórias pppg e ppmg, ao passo que a variável aleatória cmg depende das variáveis aleatórias mpg e mmg. Dessa forma, cpg e cmg são regidas por leis *a posteriori*;

- a variável aleatória cgen depende das variáveis aleatórias cpg e cmg, sendo, pois, regida por uma lei *a posteriori*.

Como não se sabe a *priori* se o suposto pai é ou não o pai biológico da criança, faz-se necessária a inserção de uma outra variável aleatória (v.a. pb) para modelar a hipótese do suposto pai ser o pai biológico da criança (H). No geral, existem outras evidências que não genéticas e que não são consideradas nesse estudo. Portanto, inicia-se essa variável com 50% ($\Pr(H) = 0{,}5$). Essa variável deve ser sensoriada na rede.

Com isso, a variável aleatória cpg dependerá não mais apenas das variáveis aleatórias pppg e ppmg, mas também da variável aleatória pb.

A rede bayesiana apresentada na Figura 5.1 corresponde ao modelo para o domínio em questão. A aplicação de um modelo similar pode ser vista em Pena (2006).

5.3 Inferência no Modelo

Sendo $\Pr(\mathtt{pb} \cap \mathtt{pppg} \cap \cdots \cap \mathtt{cgen})$ a função que caracteriza a distribuição conjunta das variáveis aleatórias do modelo, utilizando o teorema do produto das probabilidades e as informações fornecidas pela topologia da rede tem-se:

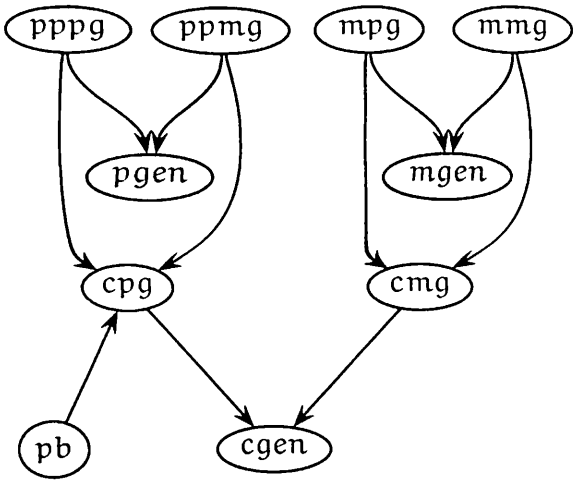

Figura 5.1: Rede bayesiana para caso padrão

$$\begin{aligned}\Pr(pb \cap pppg \cap \cdots \cap cgen) = {} & \Pr(pppg) \times \Pr(ppmg) \times \Pr(pb) \times \Pr(pgen \mid pppg \cap ppmg) \times \\ & \Pr(cpg \mid pppg \cap ppmg \cap pb) \times \Pr(mpg) \times \Pr(mmg) \times \\ & \Pr(mgen \mid mpg \cap mmg) \times \Pr(cmg \mid mpg \cap mmg) \times \\ & \Pr(cgen \mid cpg \cap cmg).\end{aligned}$$

Assumindo que E descreve o conjunto de evidências[2], a equação acima pode ser reescrita da seguinte forma:

$$\begin{aligned}\Pr(pb \cap E) = {} & \Pr(pppg) \times \Pr(ppmg) \times \Pr(pb) \times \Pr(pgen \mid pppg \cap ppmg) \times \\ & \Pr(cpg \mid pppg \cap ppmg \cap pb) \times \Pr(mpg) \times \Pr(mmg) \times \\ & \Pr(mgen \mid mpg \cap mmg) \times \Pr(cmg \mid mpg \cap mmg) \times \\ & \Pr(cgen \mid cpg \cap cmg).\end{aligned}$$

Pela Definição 5 da página 17, pode-se escrever as seguintes equações:

$$\Pr(pb = sim \mid E) = \frac{\Pr(pb = sim \cap E)}{\Pr(E)} \tag{5.1}$$

$$\Pr(E \mid pb = sim) = \frac{\Pr(pb = sim \cap E)}{\Pr(pb = sim)} \Rightarrow \Pr(pb = sim \cap E) = \Pr(E \mid pb = sim) \cdot \Pr(pb = sim) \tag{5.2}$$

$$\Pr(pb = não \mid E) = \frac{\Pr(pb = não \cap E)}{\Pr(E)} \tag{5.3}$$

[2]Valores que as variáveis aleatórias do modelo, com exceção de pb, assumem no estudo em questão.

$$\Pr(E \mid pb = \text{não}) = \frac{\Pr(pb = \text{não} \cap E)}{\Pr(pb = \text{não})} \Rightarrow \Pr(pb = \text{não} \cap E) = \Pr(E \mid pb = \text{não}) \cdot \Pr(pb = \text{não}) \tag{5.4}$$

Substituindo a equação (5.2) na (5.1), obtem-se:

$$\Pr(pb = \text{sim} \mid E) = \frac{\Pr(E \mid pb = \text{sim}) \cdot \Pr(pb = \text{sim})}{\Pr(E)} \tag{5.5}$$

Da mesma forma, substituindo a equação (5.4) na (5.3), obtem-se:

$$\Pr(pb = \text{não} \mid E) = \frac{\Pr(E \mid pb = \text{não}) \cdot \Pr(pb = \text{não})}{\Pr(E)} \tag{5.6}$$

Dividindo a equação (5.5) pela (5.6), tem-se:

$$\frac{\Pr(pb = \text{sim} \mid E)}{\Pr(pb = \text{não} \mid E)} = \frac{\Pr(E \mid pb = \text{sim}) \cdot \Pr(pb = \text{sim})}{\Pr(E \mid pb = \text{não}) \cdot \Pr(pb = \text{não})} \tag{5.7}$$

Sendo H o evento que representa a hipótese do suposto pai ser o pai biológico da criança ($pb = \text{sim}$) e H^c o evento que representa a hipótese do pai biológico da criança ser outro indivíduo da população ($pb = \text{não}$), pode-se reescrever a equação (5.7) como segue:

$$\frac{\Pr(H \mid E)}{\Pr(H^c \mid E)} = \frac{\Pr(E \mid H)}{\Pr(E \mid H^c)} \cdot \frac{\Pr(H)}{\Pr(H^c)} \tag{5.8}$$

A resolução da equação (5.8) dado um conjunto de evidências fornece o valor do **IP** (ver seção 4.5.2 na página 47).

Como $\Pr(H) = 0{,}5$, $\Pr(H)/\Pr(H^c) = 1$, logo

$$\frac{\Pr(H \mid E)}{\Pr(H^c \mid E)} = \frac{\Pr(E \mid H)}{\Pr(E \mid H^c)} \tag{5.9}$$

Tendo em vista a construção completa do modelo, é possível, dado um conjunto de evidências, obter por meio de inferências os resultados requeridos pela genética forense no que tange ao cálculo do IP.

5.4 Aplicação Prática do Modelo

Supondo que esteja sendo feito o cálculo do IP para o marcador genérico M_1 cujas freqüências alélicas para a população em estudo são mostradas na Tabela 5.1, a seguir serão construídas as tabelas de probabilidade *a priori* e *a posteriori* das variáveis aleatórias do modelo para este marcador. É impor-

tante mencionar que os procedimentos aqui aplicados podem ser utilizados na construção das tabelas de probabilidade de quaisquer marcadores.

Marcador: M_1	
Alelo	Freqüência
a	0,122
b	0,222
c	0,656
Soma	1,000

Tabela 5.1: Freqüências alélicas do marcador M_1

5.4.1 Construção das Tabelas de Probabilidade

As tabelas de probabilidades *a priori* das variáveis aleatórias `pppg`, `ppmg`, `mpg` e `mmg` (ver Tabelas 5.2, 5.3, 5.4 e 5.5) correspondem exatamente à tabela de freqüências alélicas (ver Tabela 5.1).

w	Pr(pppg = w)
a	0,122
b	0,222
c	0,656

Tabela 5.2: Tabela de probabilidade *a priori* de `pppg`

w	Pr(ppmg = w)
a	0,122
b	0,222
c	0,656

Tabela 5.3: Tabela de probabilidade *a priori* de `ppmg`

w	Pr(mpg = w)
a	0,122
b	0,222
c	0,656

Tabela 5.4: Tabela de probabilidade *a priori* de `mpg`

A tabela de probabilidade *a priori* da variável aleatória `pb` é mostrada na Tabela 5.6.

A tabela de probabilidade *a posteriori* da variável aleatória `cpg` é mostrada na Tabela 5.7.

Observe que sendo verdadeira a hipótese "o pai biológico da criança é um indivíduo da população diferente do suposto pai" (`pb` = `não`), as probabilidades

w	$\Pr(\texttt{mmg} = w)$
a	0,122
b	0,222
c	0,656

Tabela 5.5: Tabela de probabilidade *a priori* de mmg

w	$\Pr(\texttt{pb} = w)$
sim	0,5
não	0,5

Tabela 5.6: Tabela de probabilidade *a priori* de pb

de cpg correspondem exatamente às freqüências do alelo em questão, independente do genótipo do suposto pai. Todavia, partindo do pressuposto que "o suposto pai é o pai biológico da criança" (pb = sim), as probabilidades dependerão diretamente da configuração dos alelos desse indivíduo.

			$\Pr(\texttt{cpg} \mid \cdot)$		
pppg	ppmg	pb	a	b	c
a	a	sim	1,000	0,000	0,000
a	a	não	0,122	0,222	0,656
a	b	sim	0,500	0,500	0,000
a	b	não	0,122	0,222	0,656
a	c	sim	0,500	0,000	0,500
a	c	não	0,122	0,222	0,656
b	a	sim	0,500	0,500	0,000
b	a	não	0,122	0,222	0,656
b	b	sim	0,000	1,000	0,000
b	b	não	0,122	0,222	0,656
b	c	sim	0,000	0,500	0,500
b	c	não	0,122	0,222	0,656
c	a	sim	0,500	0,000	0,500
c	a	não	0,122	0,222	0,656
c	b	sim	0,000	0,500	0,500
c	b	não	0,122	0,222	0,656
c	c	sim	0,000	0,000	1,000
c	c	não	0,122	0,222	0,656

Tabela 5.7: Tabela de probabilidade *a posteriori* de cpg

A tabela de probabilidade *a posteriori* da variável aleatória cmg é mostrada na Tabela 5.8. As probabilidades de cmg dependem diretamente da configuração dos alelos da mulher, haja vista essa ser realmente a mãe biológica da criança.

Em se tratando dos genótipos, estes dependem diretamente da configuração dos alelos dos indivíduos. Dessa forma, as tabelas de probabilidade *a*

		Pr(cmg | ·)		
mpg	mmg	a	b	c
a	a	1,000	0,000	0,000
a	b	0,500	0,500	0,000
a	c	0,500	0,000	0,500
b	a	0,500	0,500	0,000
b	b	0,000	1,000	0,000
b	c	0,000	0,500	0,500
c	a	0,500	0,000	0,500
c	b	0,000	0,500	0,500
c	c	0,000	0,000	1,000

Tabela 5.8: Tabela de probabilidade *a posteriori* de cmg

posteriori de pgen, mgen e cgen podem ser construídas como mostrado nas Tabelas 5.9, 5.10 e 5.11 respectivamente.

		Pr(pgen | ·)					
pppg	ppmg	a−a	a−b	a−c	b−b	b−c	c−c
a	a	1,000	0,000	0,000	0,000	0,000	0,000
a	b	0,000	1,000	0,000	0,000	0,000	0,000
a	c	0,000	0,000	1,000	0,000	0,000	0,000
b	a	0,000	1,000	0,000	0,000	0,000	0,000
b	b	0,000	0,000	0,000	1,000	0,000	0,000
b	c	0,000	0,000	0,000	0,000	1,000	0,000
c	a	0,000	0,000	1,000	0,000	0,000	0,000
c	b	0,000	0,000	0,000	0,000	1,000	0,000
c	c	0,000	0,000	0,000	0,000	0,000	1,000

Tabela 5.9: Tabela de probabilidade *a posteriori* de pgen

		Pr(mgen | ·)					
mpg	mmg	a−a	a−b	a−c	b−b	b−c	c−c
a	a	1,000	0,000	0,000	0,000	0,000	0,000
a	b	0,000	1,000	0,000	0,000	0,000	0,000
a	c	0,000	0,000	1,000	0,000	0,000	0,000
b	a	0,000	1,000	0,000	0,000	0,000	0,000
b	b	0,000	0,000	0,000	1,000	0,000	0,000
b	c	0,000	0,000	0,000	0,000	1,000	0,000
c	a	0,000	0,000	1,000	0,000	0,000	0,000
c	b	0,000	0,000	0,000	0,000	1,000	0,000
c	c	0,000	0,000	0,000	0,000	0,000	1,000

Tabela 5.10: Tabela de probabilidade *a posteriori* de mgen

Pelas informações fornecidas pela topologia da rede da Figura 5.1, tem-se para $w \in \{a, b, c\}$:

		Pr(cgen | ·)					
cpg	cmg	a−a	a−b	a−c	b−b	b−c	c−c
a	a	1,000	0,000	0,000	0,000	0,000	0,000
a	b	0,000	1,000	0,000	0,000	0,000	0,000
a	c	0,000	0,000	1,000	0,000	0,000	0,000
b	a	0,000	1,000	0,000	0,000	0,000	0,000
b	b	0,000	0,000	0,000	1,000	0,000	0,000
b	c	0,000	0,000	0,000	0,000	1,000	0,000
c	a	0,000	0,000	1,000	0,000	0,000	0,000
c	b	0,000	0,000	0,000	0,000	1,000	0,000
c	c	0,000	0,000	0,000	0,000	0,000	1,000

Tabela 5.11: Tabela de probabilidade *a posteriori* de cgen

1. Equação para o cálculo da $\Pr(\mathrm{cpg} = w)$

$$\Pr(\mathrm{cpg} = w) = \sum_{(i=a,b,c)} \sum_{(j=a,b,c)} \sum_{(k=sim,não)} [\, \Pr(\mathrm{pppg} = i) \times \Pr(\mathrm{ppmg} = j) \times \Pr(\mathrm{pb} = k) \times \Pr(\mathrm{pgen} = i-j \mid \mathrm{pppg} = i \cap \mathrm{ppmg} = j) \times \Pr(\mathrm{cpg} = w \mid \mathrm{pppg} = i \cap \mathrm{ppmg} = j \cap \mathrm{pb} = k) \,]$$

2. Equação para o cálculo da $\Pr(\mathrm{cmg} = w)$

$$\Pr(\mathrm{cmg} = w) = \sum_{(i=a,b,c)} \sum_{(j=a,b,c)} [\, \Pr(\mathrm{mpg} = i) \times \Pr(\mathrm{mmg} = j) \times \Pr(\mathrm{mgen} = i-j \mid \mathrm{mpg} = i \cap \mathrm{mmg} = j) \times \Pr(\mathrm{cmg} = w \mid \mathrm{mpg} = i \cap \mathrm{mmg} = j) \,]$$

Utilizando as equações acima, é possível construir as tabelas de probabilidade *a priori* da variável aleatória cpg (calculando as probabilidades de cpg = a, cpg = b e cpg = c) e da variável aleatória cmg (calculando as probabilidades de cmg = a, cmg = b e cmg = c). Sendo importante destacar que o genótipo i−j é igual ao genótipo j−i.

No Apêndice B, são mostrados em detalhes os cálculos das probabilidades necessárias à construção das tabelas de probabilidade *a priori* de cpg e cmg (ver Tabelas 5.12 e 5.13).

É observado que ao se construir as tabelas de probabilidade *a priori* das variáveis cpg e cmg, obtêm-se tabelas idênticas à Tabela 5.1, visto que uma vez removidos os condicionamentos, os valores assumidos por essas variáveis correspondem exatamente às freqüências alélicas da população em que o

estudo está sendo realizado.

w	$\Pr(cpg = w)$
a	0,122
b	0,222
c	0,656

Tabela 5.12: Tabela de probabilidade *a priori* de cpg

w	$\Pr(cmg = w)$
a	0,122
b	0,222
c	0,656

Tabela 5.13: Tabela de probabilidade *a priori* de cmg

5.4.2 Cálculo do IP

Supondo que os genótipos da criança, de sua mãe e seu suposto pai sejam respectivamente $G_C = a - b$, $G_M = b - b$ e $G_{SP} = a - a$, tem-se $E = \{pppg = a, ppmg = a, pgen = a - a, mpg = b, mmg = b, mgen = b - b, cpg = a, cmg = b, cgen = a - b\}$.

Dessa forma, o IP pode ser calculado como segue:

1. **A probabilidade do suposto pai ser o pai biológico da criança dadas as evidências ($\Pr(H \mid E)$):**

$$
\begin{aligned}
\Pr(H \cap E) &= \Pr(pppg = a) \times \Pr(ppmg = a) \times \Pr(pb = sim) \times \\
&\quad \Pr(pgen = a - a \mid pppg = a \cap ppmg = a) \times \\
&\quad \Pr(cpg = a \mid pppg = a \cap ppmg = a \cap pb = sim) \times \Pr(mpg = b) \times \Pr(mmg = b) \times \\
&\quad \Pr(mgen = b - b \mid mpg = b \cap mmg = b) \times \Pr(cmg = b \mid mpg = b \cap mmg = b) \times \\
&\quad \Pr(cgen = a - b \mid cpg = a \cap cmg = b) \\
&= 0,122 \times 0,122 \times 0,5 \times 1 \times 1 \times 0,222 \times 0,222 \times 1 \times 1 \times 1 \\
&= 0,000366771
\end{aligned}
$$

2. **A probabilidade do pai biológico da criança ser outro indivíduo da**

população dadas as evidências ($Pr(H^c \mid E)$)**:**

$$\begin{aligned}
Pr(H^c \cap E) &= Pr(pppg = a) \times Pr(ppmg = a) \times Pr(pb = \text{não}) \times \\
&\quad Pr(pgen = a - a \mid pppg = a \cap ppmg = a) \times \\
&\quad Pr(cpg = a \mid pppg = a \cap ppmg = a \cap pb = \text{não}) \times Pr(mpg = b) \times Pr(mmg = b) \times \\
&\quad Pr(mgen = b - b \mid mpg = b \cap mmg = b) \times Pr(cmg = b \mid mpg = b \cap mmg = b) \times \\
&\quad Pr(cgen = a - b \mid cpg = a \cap cmg = b) \\
&= 0,122 \times 0,122 \times 0,5 \times 1 \times 0,122 \times 0,222 \times 0,222 \times 1 \times 1 \times 1 \\
&= 0,000044746
\end{aligned}$$

Normalizando, tem-se $Pr(H \mid E) = 0,891265731$ e $Pr(H^c \mid E) = 0,108734268$. Logo,

$$IP = \frac{0,891265731}{0,108734268} \simeq 8,1967.$$

Após propagar as evidências no modelo, os valores de $Pr(H \mid E)$ e $Pr(H^c \mid E)$ ficarão armazenados na variável aleatória pb.

Pela Tabela 4.2 da página 50, tem-se que para esse conjunto de evidências, o IP é dado por $1/p_{A_1}$, onde p_{A_1} é a freqüência do alelo A_1 possivelmente herdado do suposto pai. Dessa forma, o cálculo pode ser feito como segue:

$$IP = \frac{1}{p_{A_1}} = \frac{1}{p_a} = \frac{1}{0,122} \simeq 8,1967.$$

5.5 Casos Complexos de Paternidade

Em muitos casos de estudo de paternidade, não é possível obter o DNA do suposto pai, sendo necessário trabalhar com os genótipos de seus familiares, o que torna os cálculos mais complexos. Todavia ao se utilizar as redes bayesianas como meio de representação do domínio em questão, o cálculo se torna simples, haja vista que uma vez o modelo construído, para o cálculo do IP (muitas vezes chamado de razão de verossimilhança) é necessário apenas a inserção das evidências, deixando a própria rede se encarregar de inferir o resultado desejado.

Para construir uma rede bayesiana para um caso complexo, é preciso apenas fazer uma extensão da rede da Figura 5.1 (página 56), adicionando novas variáveis aleatórias para representar as novas evidências necessárias ao estudo. É importante ressaltar que com um mesmo modelo é possível realizar vários estudos de paternidade, desde que as evidências requeridas pelos estudos estejam representadas no modelo.

A rede da Figura 5.2 corresponde a uma extensão do modelo construído

anteriormente na qual podem ser realizados, dentre outros estudos de paternidade, o caso padrão (em que se tem o genótipo da criança, de sua mãe e de seu suposto pai) e um caso complexo em que se tem o genótipo da criança, de sua mãe e dos seus possíveis avós paternos.

O sistema THÊMIS utiliza o modelo apresentado na Figura 5.2.

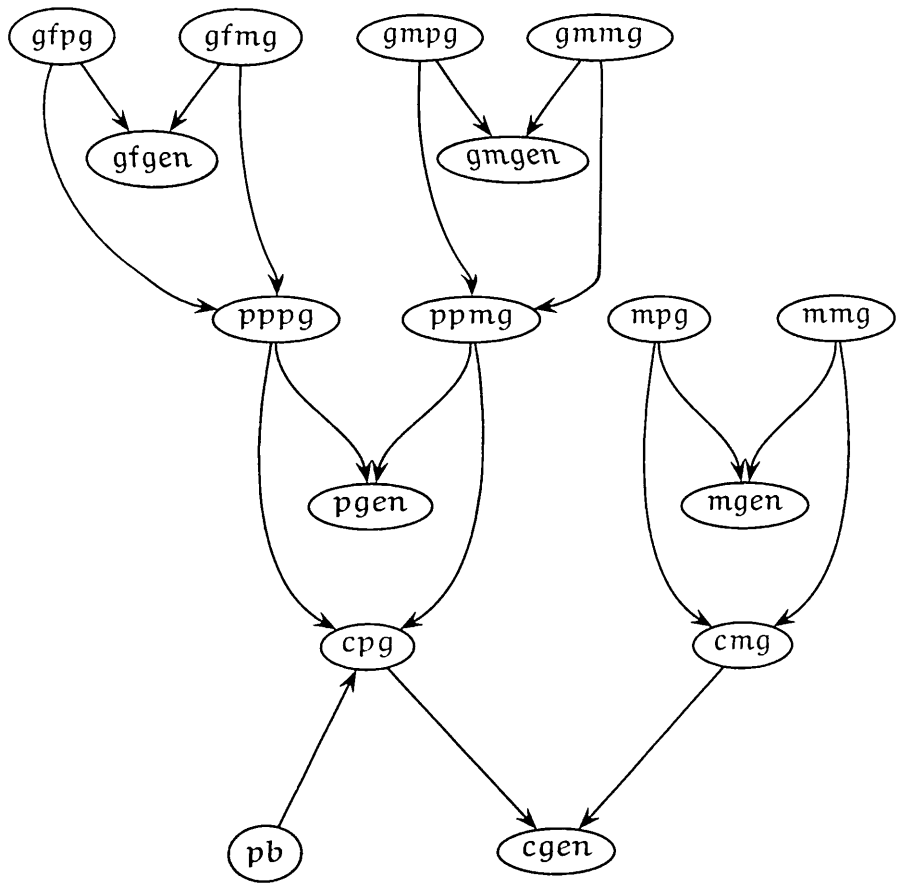

Figura 5.2: Rede bayesiana para caso complexo

5.6 Comentários

No presente capítulo, foram apresentados em detalhes o domínio em que se encontram os estudos de paternidade, a construção de um modelo utilizando redes bayesianas como meio de representação desse domínio, a inferência nesse modelo, bem como uma aplicação prática do mesmo.

Capítulo 6

O Sistema THÊMIS

O presente capítulo visa oferecer uma visão geral do sistema THÊMIS, documentando a sua especificação como um conjunto de modelos de sistema —representações gráficas que descrevem o problema a ser resolvido ou o sistema a ser desenvolvido.

Um modelo de sistema é uma abstração do sistema em estudo, sendo uma importante ponte entre o processo de análise e de projeto (Somerville, 2003).

O capítulo está disposto nas seguintes seções:

Descrição do Sistema Breve descrição do sistema THÊMIS utilizando linguagem natural.

Tecnologias Usadas Apresentação das tecnologias usadas na implementação do software.

Especificação de Requisitos Definição das funções e restrições do sistema.

Projeto Modelagem do sistema com base em seus requisitos.

Validação Apresentação de dois dentre os vários estudos de paternidade reais cedidos pelo Laboratório de DNA Forense da Universidade Federal de Alagoas usados para validar o sistema.

Em toda a modelagem foram utilizadas notações definidas na *Unified Modeling Language* (UML) que está emergindo como uma linguagem padrão de modelagem, particularmente para a modelagem orientada a objetos.

A Programação Orientada a Objetos (POO) ou ainda em inglês *Object-Oriented Programming (OOP)* é um paradigma de análise, projeto e programação de sistemas de software baseado na composição e interação entre diversas unidades de software chamadas de objetos.

Detalhes sobre UML e POO podem ser obtidos em Booch et al. (2005) e Bruegge & Dutoit (2003).

6.1 Descrição do Sistema

O sistema THÊMIS é um software para análise de vínculo genético em estudos de paternidade. Esse software utiliza o último modelo apresentado no capítulo anterior (ver Figura 5.2 na página 64) para realizar inferências a partir das evidências que se dispõe para o estudo, seja ele o caso padrão, em que se tem evidências (genótipos) da criança, de sua mãe e de seu suposto pai (ver Figura 4.6 na página 49), ou um tipo de caso complexo, em que se tem apenas os genótipos da criança, de sua mãe e dos pais de seu suposto pai (ver Figura 4.8 na página 52).

A realização de um estudo de paternidade pode ser dividida em duas etapas principais: a tipagem do DNA das pessoas envolvidas no estudo, por meio da qual se obtêm os perfis genéticos desses indivíduos, e a análise estatística desses perfis em busca de verificar a paternidade da criança.

No Laboratório de DNA Forense da Universidade Federal de Alagoas, há equipamentos que permitem a obtenção dos perfis genéticos de pessoas, dentre eles máquina para execução do PCR e o seqüenciador. A análise probabilística desses perfis genéticos, que engloba o cálculo dos IPs e de outros índices relevantes a esse tipo de estudo, é feita com o auxílio de planilhas eletrônicas, nas quais são construídas funções para esse fim.

Quando se trata de casos complexos, ou seja, casos em que não se pode fazer a tipagem do DNA do suposto pai, é utilizado o software `familias` (ver Egeland et al., 2000) para obter os valores do IP. Estes valores são então inseridos manualmente na planilha eletrônica para que sejam calculados outros índices relevantes ao estudo, os quais não dependem diretamente de uma análise estatística acurada, como é o caso do Índice de Paternidade Combinado (IPC).

No `familias`, que também utiliza o ferramental das redes bayesianas no cálculo do IP, é necessário o usuário construir manualmente a rede que representa o estudo em questão, exigindo, pois, desse indivíduo um certo grau de conhecimento sobre as redes bayesianas. Além disso, o `familias` não possui uma base de dados para o armazenamento dos perfis genéticos, os quais, devido a isso, não poderão ser utilizados em estudos futuros.

Diferentemente do `familias`, no THÊMIS as informações sobre os estudos são armazenadas numa base de dados e não é exigido do usuário a construção da rede que representa a genealogia em questão. Dessa forma, os dados inseridos no sistema podem ser utilizados em análises futuras e o usuário não necessita ter conhecimento algum sobre redes bayesianas, ou seja, o uso desse formalismo fica transparente ao usuário.

Por utilizar as redes bayesianas na análise estatística dos perfis genéticos, o sistema THÊMIS, que atualmente contempla dois tipos de estudo de paternidade com o uso de uma única topologia de rede, pode dar suporte a outros tipos de estudo de paternidade com uma simples extensão da topologia da rede bayesiana atual conforme discutido no Capítulo 5.

Dentre os softwares analisados que utilizam redes bayesianas como abordagem para representação do conhecimento (ver Apêndice A), o escolhido como motor de inferência do sistema na análise dos dados biológicos no que tange à verificação de vínculo genético em estudos de paternidade foi o `UnBBayes`, visto que este software se encontra disponível sob licença *GNU General Public License (GPL)*, possui *Application Programming Interface (API)*, bem como interface gráfica bastante simples e intuitiva, eficiência no processo de inferência e confiabilidade.

6.2 Tecnologias e Ferramentas Usadas

Na implementação do software, as seguintes tecnologias e ferramentas foram utilizadas:

Java Linguagem de programação orientada a objetos desenvolvida pela empresa *Sun Microsystems*. Com o uso dessa linguagem, ao se compilar o código fonte da aplicação, é gerado um conjunto de *bytecodes*, os quais para serem executados necessitam de uma máquina virtual, denominada *Java Virtual Machine (JVM)*. Dessa forma, os *bytecodes* correspondem a um estágio intermédio entre o código-fonte e o sistema final, tendo como principal vantagem a portabilidade, haja vista que os *bytecodes* irão produzir o mesmo resultado em qualquer arquitetura de hardware, independente da plataforma de software. Essa característica foi primordial na seleção dessa linguagem, além do fato dela ser gratuita. Para maiores detalhes sobre essa linguagem acesse `http://java.sun.com/`.

Hibernate *Framework* Java para o mapeamento objeto-relacional, no qual as entidades do banco de dados são representadas por meio de classes e os registros de cada entidade correspondem às instâncias dessas classes. Com o uso do *Hibernate*, não é necessário o desenvolvedor se preocupar com os comandos em linguagem SQL, um vez que este *framework* gera as chamadas SQL, mantendo o programa portável para quaisquer bancos de dados SQL, exigindo, para isso, pequenas modificações nos arquivos de configuração. Essa portabilidade juntamente com a facilidade de uso, a

praticidade e a gratuidade foram de fundamental importância na seleção dessa tecnologia. Para maiores detalhes sobre o *Hibernate* acesse `http://www.hibernate.org/`.

Eclipse Ambiente Integrado de Desenvolvimento (IDE[1]) para a implementação de softwares. Esta IDE possui grande quantidade e variedade de *plugins*, objetivando atender às diferentes necessidades dos desenvolvedores, além de ser gratuita e estar disponível para várias plataformas de software, características estas decisivas no momento da seleção dessa ferramenta. Maiores detalhes sobre o *Eclipse* podem ser encontrados em `http://www.eclipse.org/`.

PostgreSQL Sistema de banco de dados objeto-relacional versátil, seguro, gratuito e bem documentado, possuindo recursos comuns a bancos de dados de grande porte como, por exemplo, o *Oracle* (ver `http://www.oracle.com/`). A segurança e a gratuidade foram as características mais relevantes na seleção dessa ferramenta, além do fato dela estar disponível para várias plataformas de software. Maiores detalhes do *PostgreSQL* em `http://www.postgresql.org/`.

Jude Ferramenta para modelagem de sistemas de software utilizando UML, suportando a construção de diversos diagramas dessa linguagem. Dentre as características desse software que corroboraram a sua seleção, estão a disponibilização de versões gratuitas e o fato dele ser multiplataforma. Maiores detalhes da ferramenta em `http://jude.change-vision.com/`.

brModelo Ferramenta gratuita que provê a construção de modelos conceitual, lógico e físico de bancos de dados relacionais, sendo possível a geração de um modelo a partir de outro: físico a partir de lógico e lógico a partir de conceitual. Maiores detalhes dessa ferramenta em `http://www.sis4.com/brModelo/`.

6.3 Especificação de Requisitos

A especificação de requisitos é uma das fases mais importantes no processo de desenvolvimento de um sistema de software, visto que os problemas que os engenheiros de software se defrontam são, muitas vezes, bastante complexos, propiciando dificuldades em estabelecer com exatidão os serviços que o sistema deve prover.

[1]Sigla para *Integrated Development Environment* – aplicativo que provê ferramentas para apoio ao desenvolvimento de software, visando agilidade no processo.

As descrições das funções e das restrições são os requisitos para o sistema; e o processo de descobrir, analisar, documentar e verificar essas funções e retrições é chamado de Engenharia de Requisitos (Somerville, 2003).

6.3.1 Documento de Requisitos de Software

Abaixo são listados os requisitos do usuário do sistema, bem como os serviços que o software deve oferecer.

Requisitos do Usuário do Sistema

- Inserir Freqüências Alélicas.
- Inserir Processo.
- Inserir Perfis Genéticos.
- Executar Cálculo de Paternidade.
- Verificar Resultado do Cálculo.
- Excluir Processo.

Funções do Sistema

- Oferecer segurança no armazenamento das informações sobre os processos, desde os dados pessoais dos indíviduos envolvidos ao resultado do estudo.
- Oferecer confiabilidade nos resultados dos estudos, utilizando o formalismo das redes bayesianas de modo transparente ao usuário.
- Disponibilizar relatórios de todos os estudos realizados.

6.3.2 Casos de Uso

O diagrama de casos de uso é o diagrama mais geral e informal da UML, sendo utilizado normalmente nas fases de levantamento e análises de requisitos do sistema. Apresenta uma linguagem simples e de fácil compreensão para que os usuários possam ter uma idéia geral de como o sistema irá se comportar. Detalhes sobre esse tipo de diagrama UML podem ser obtidos em Booch et al. (2005).

A Figura 6.1 é o diagrama de casos de uso do sistema THÊMIS.

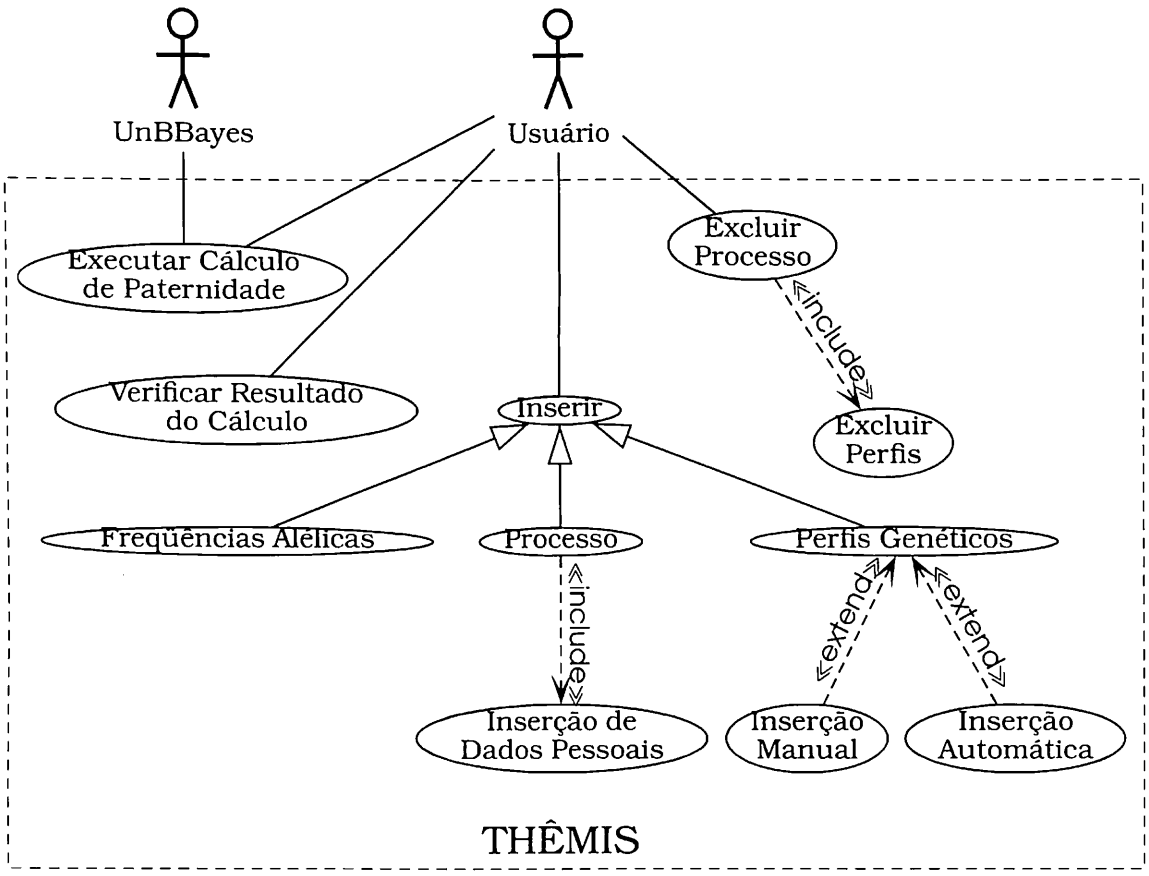

Figura 6.1: Diagrama de Casos de Uso

Descrição dos Casos de Uso

1. <**UC0001**> **Inserir Freqüências Alélicas**

   Descrição: Inserção das tabelas de freqüências alélicas das populações em que os estudos serão realizados.

   Pré-condições: As tabelas de freqüências alélicas estarem num dado formato. Atualmente o software trabalha com essas tabelas em formato csv[2].

   Pós-condições: Permissão do cadastro de processos vinculados às populações cujas tabelas de freqüências alélicas foram inseridas.

2. <**UC0002**> **Inserir Processo**

   Descrição: Inserção de um processo de estudo de paternidade e dos dados pessoais dos indivíduos envolvidos.

   Pré-condições: Ter sido inserida ao menos uma tabela de freqüências alélicas.

[2]Formato de arquivo em que os dados estão separados por vírgula ou ponto-e-vírgula.

Pós-condições: Permissão do cadastro dos perfis genéticos dos indivíduos envolvidos no processo.

3. <**UC0003**> **Inserir Perfis Genéticos**

   Descrição: Inserção dos perfis genéticos dos indivíduos envolvidos num dado processo manualmente ou via arquivo. Atualmente o software trabalha com esses arquivos em formato csv.

   Pré-condições: Ter sido criado o processo.

   Pós-condições: Permissão da execução dos cálculos de paternidade para o processo.

4. <**UC0004**> **Executar Cálculo de Paternidade**

   Descrição: Execução dos cálculos de paternidade para um dado processo utilizando o UnBBayes.

   Pré-condições: Terem sido inseridos os perfis genéticos dos indivíduos vinculados ao processo.

   Pós-condições: Permissão da geração do relatório contendo o resultado do estudo de paternidade para o processo.

5. <**UC0005**> **Verificar Resultado do Cálculo**

   Descrição: Verificação via relatório do resultado do estudo de paternidade para um dado processo.

   Pré-condições: Ter sido executado o cálculo de paternidade para o processo.

   Pós-condições: Exibição do resultado do estudo.

6. <**UC0006**> **Excluir Processo**

   Descrição: Exclusão de um dado processo.

   Pré-condições: Ter sido criado um processo.

   Pós-condições: Indisponibilidade de relatórios vinculados ao processo excluído.

6.4 Projeto

Está etapa compreende todo o planejamento do software, visando atender aos requisitos que lhe foram atribuídos.

6.4.1 Modelagem da Base de Dados

A modelagem da base de dados é uma das principais etapas no processo de desenvolvimento de um sistema de software, podendo, em casos de falhas, comprometer todo o sistema.

As Figuras 6.2 e 6.3 apresentam, respectivamente, os modelos conceitual e lógico do sistema, correspondendo, pois, à modelagem da base de dados do mesmo.

O modelo conceitual está na forma de um modelo entidade relacionamento, ao passo que o lógico está sendo representado por um conjunto de tabelas passíveis de implementação num banco de dados relacional como, por exemplo, o PostgreSQL.

Detalhes sobre modelo conceitual e lógico de banco de dados e tipos de banco de dados (relacional, objeto-relacional, dentre outros) podem ser obtidos em Elmasri & Navathe (2003).

Glossário

Aqui é definido o dicionário de dados do sistema.

[marker] Entidade onde é armazenada a lista de marcadores.

[allele] Entidade onde é armazenada a lista de alelos.

[location] Entidade onde são armazenadas as localizações para os estudos de paternidade.

[study_type] Entidade onde são armazenados os tipos de estudos de paternidade. O software contempla hoje o caso padrão e um caso complexo.

[person_classification] Entidade onde são armazenados os tipos de pessoas que podem estar vinculadas a um processo (criança, mãe da criança, suposto pai da criança, suposto avô paterno da criança e suposta avó paterna da criança).

[marker_frequency] Entidade onde são armazenadas as freqüências contidas nas tabelas de freqüências alélicas, estando essas freqüências vinculadas a um alelo, um marcador e uma localização.

[process] Entidade onde são armazenadas as informações sobre os processos: data de criação, ipcombinado, pec, pp, localização e tipo de estudo de paternidade.

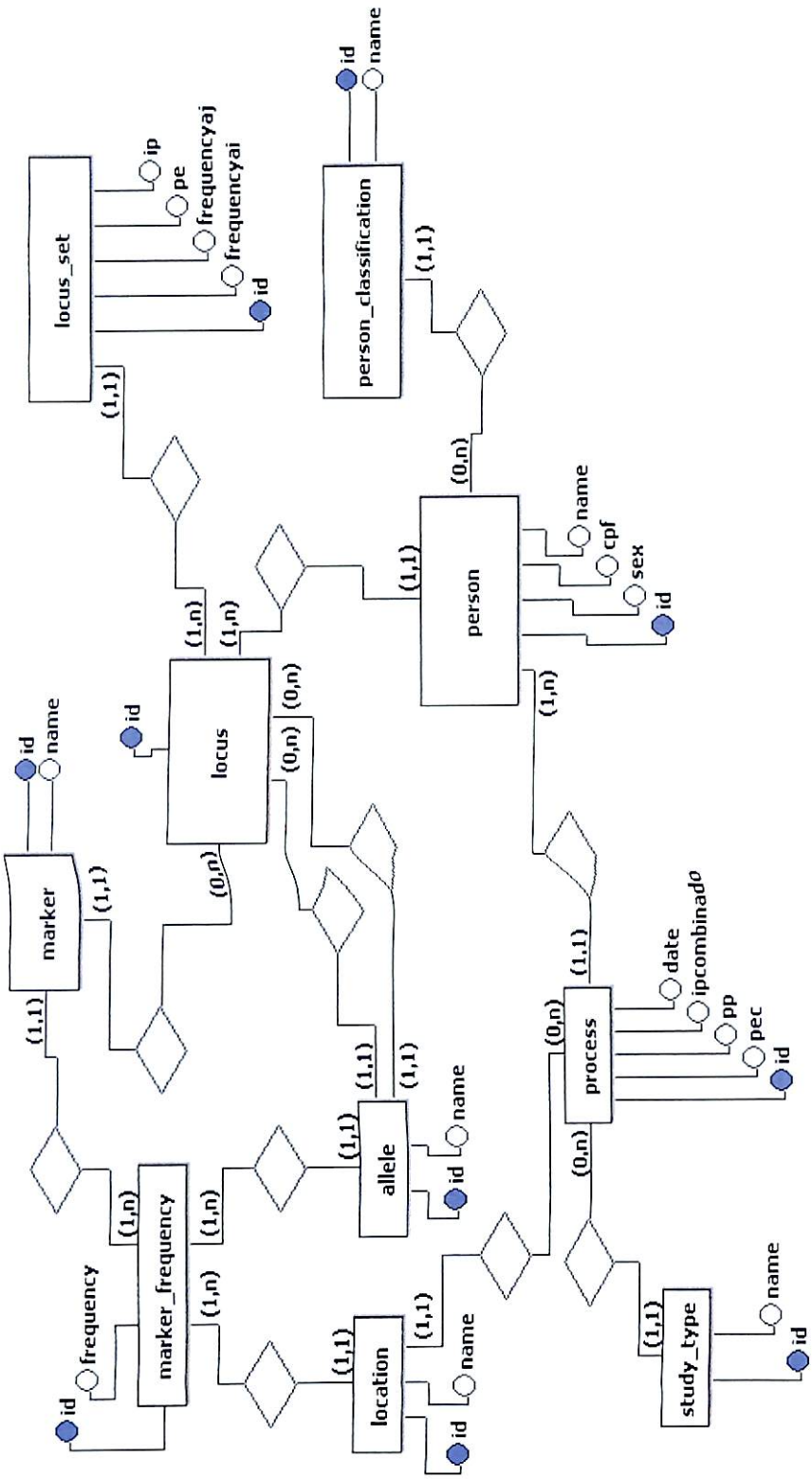

Figura 6.2: Modelo Conceitual

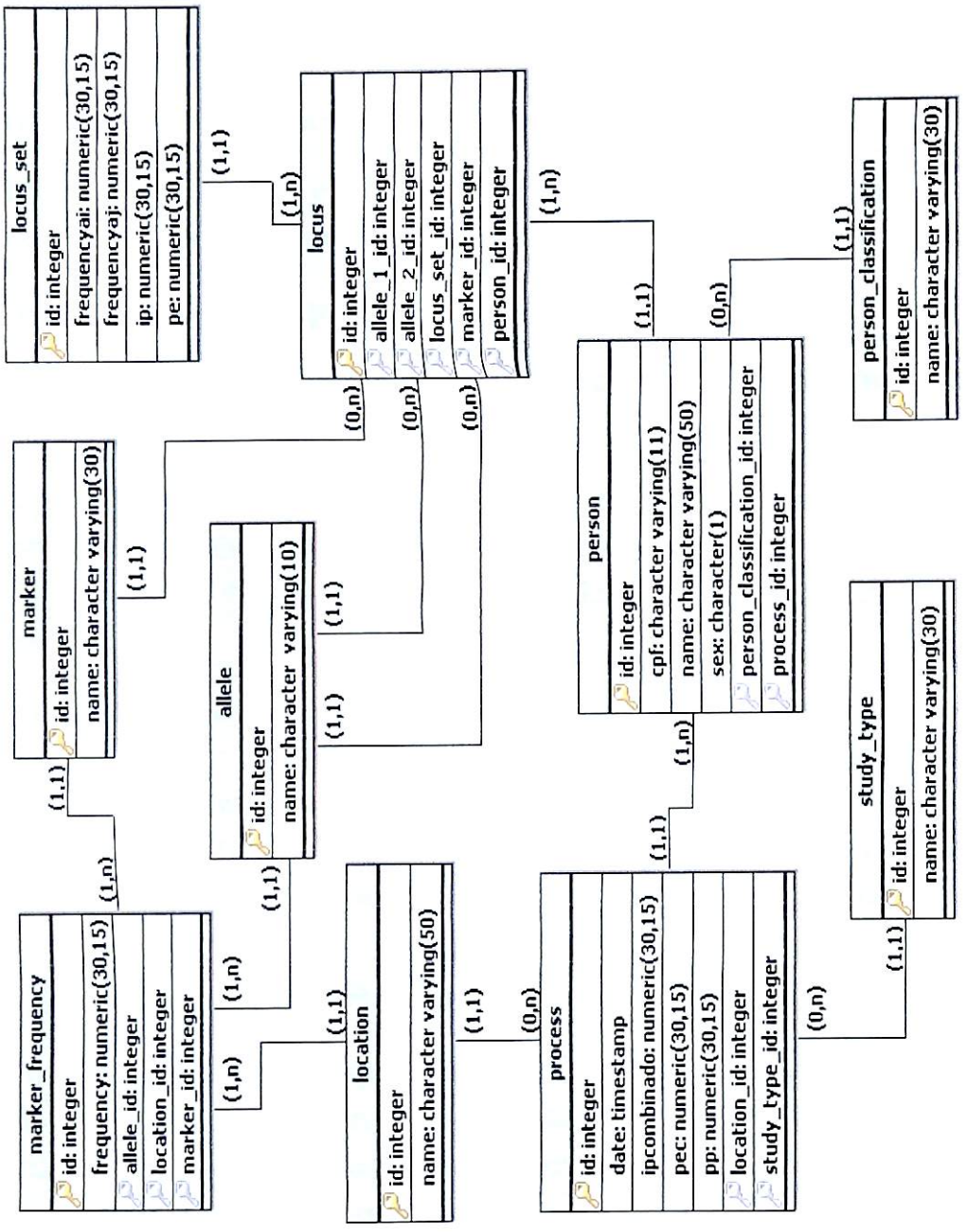

Figura 6.3: Modelo Lógico

[person] Entidade onde são armazenadas as informações sobre as pessoas: CPF, nome, sexo, o processo ao qual a pessoa esta vinculado e classificação da pessoa mediante o processo.

[locus] Entidade onde são armazenados os genótipos (conjuntos de loci) das pessoas.

[locus_set] Entidade usada para armazenar as freqüências dos alelos das crianças e os resultado do IP e PE provenientes dos cálculos.

6.4.2 Arquitetura

Em termos de arquitetura o sistema possui as seguintes camadas:

View Composta pelos componentes de interface do sistema.

Controller Composta pelos componentes responsáveis por mediar a interação entre a *View* e as demais camadas do sistema.

Model Composta pelos componentes de negócio do sistema.

Persistence Composta pelos componentes responsáveis pelo armazenamento dos dados do sistema.

Util Composta pelos componentes utilitários do sistema, responsáveis pelo carregamento das tabelas de freqüências alélicas e dos perfis genéticos via arquivos csv, assim como do cálculo de paternidade utilizando o `UnBBayes`.

A arquitetura do sistema foi modelada dessa forma com o intuito de diminuir o acoplamento no sistema, a fim de que os módulos pudessem ser desenvolvidos de maneira independente.

A Figura 6.4 mostra essa arquitetura. Detalhes sobre os componentes desse tipo de diagrama UML podem ser obtidos em Booch et al. (2005).

6.4.3 Diagrama de Atividades

O diagrama de atividades, empregado para fazer a modelagem de aspectos dinâmicos do sistema, é essencialmente um gráfico de fluxo que mostra o fluxo de controle de uma atividade para outra, explicitando concorrência entre atividades e ramificações de controle.

A Figura 6.5 é o diagrama de atividades do sistema THÊMIS. Detalhes sobre os componentes desse tipo de diagrama UML podem ser obtidos em Booch et al. (2005).

6.4.4 Diagrama de Classes

O projeto orientado a objetos é uma estratégia de projeto em que os projetistas de sistema pensam em termos de 'coisas', em vez de operações ou funções. O sistema em funcionamento é constituído de objetos que interagem entre si (Somerville, 2003).

Um diagrama de classes é uma representação da estrutura e das relações das classes que servem de modelo para objetos. É um diagrama muito útil

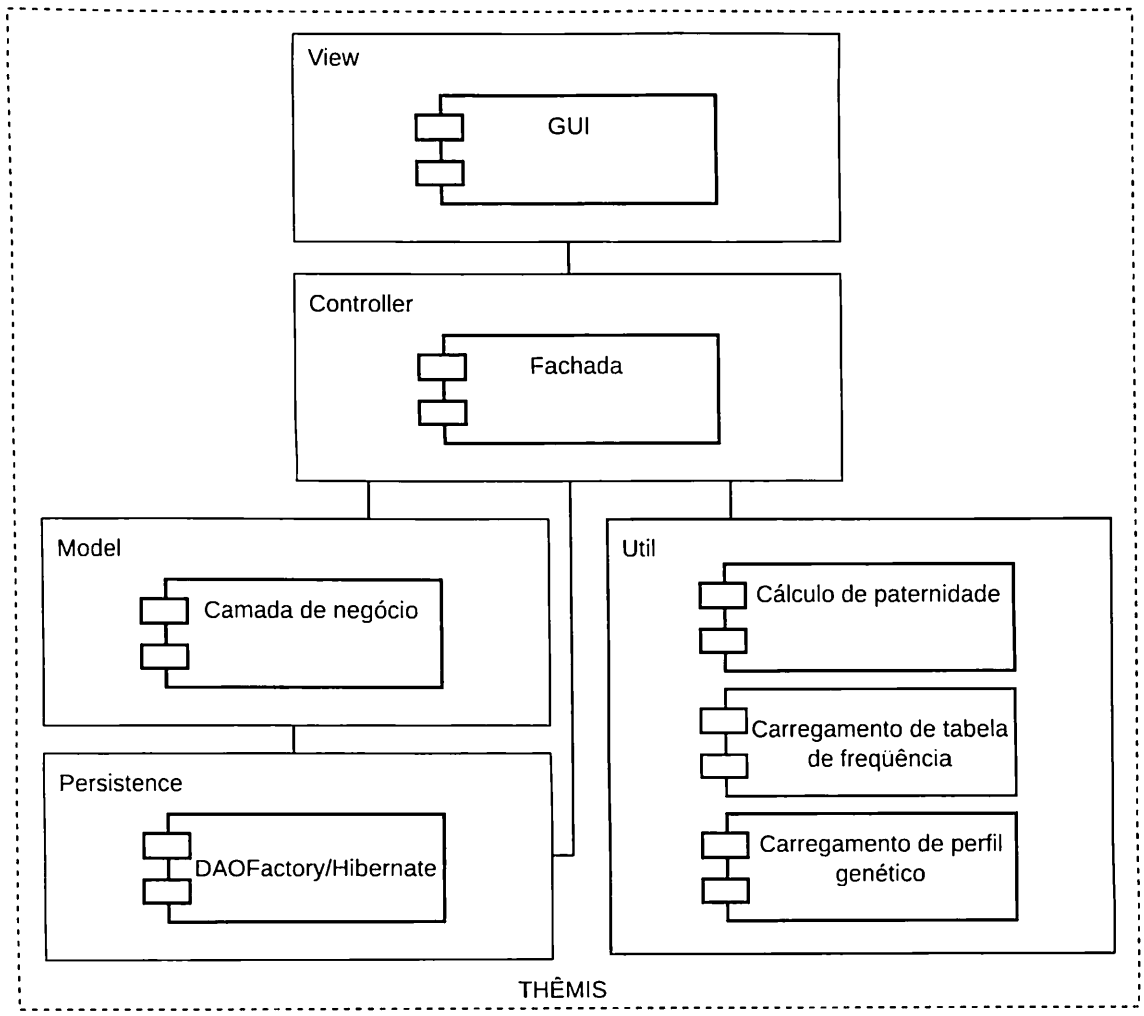

Figura 6.4: Arquitetura do Sistema: camadas, subsistemas e pacotes

para o sistema, define todas as classes necessárias ao sistema e é a base para a construção dos diagramas de comunicação, seqüência e estados (Guedes, 2008).

As classes do sistema estão agrupadas em pacotes, tendo como critério para esse agrupamento os serviços providos por elas.

No desenvolvimento do sistema, foi utilizada uma extensão da abordagem ***MVC*** com o objetivo de facilitar as atividades de desenvolvimento de software, utilizando orientação a objetos e dividindo o sistema em cinco camadas principais: ***M****odel*, ***V****iew*, ***C****ontroller*, *Persistence* e *Util*.

Na *View* (apresentação), encontram-se as classes que provêem as interfaces do sistema por meio das quais o usuário irá interagir com o software.

No *Controller* (controle), encontram-se as classes que definirão o modo como as interfaces do usuário reagem às entradas do mesmo, ou seja, o controle é o responsável por coordenar (controlar) a interação entre a camada de apresentação (*View*) e as demais camadas do sistema: *Model*, *Persistence* e *Util*.

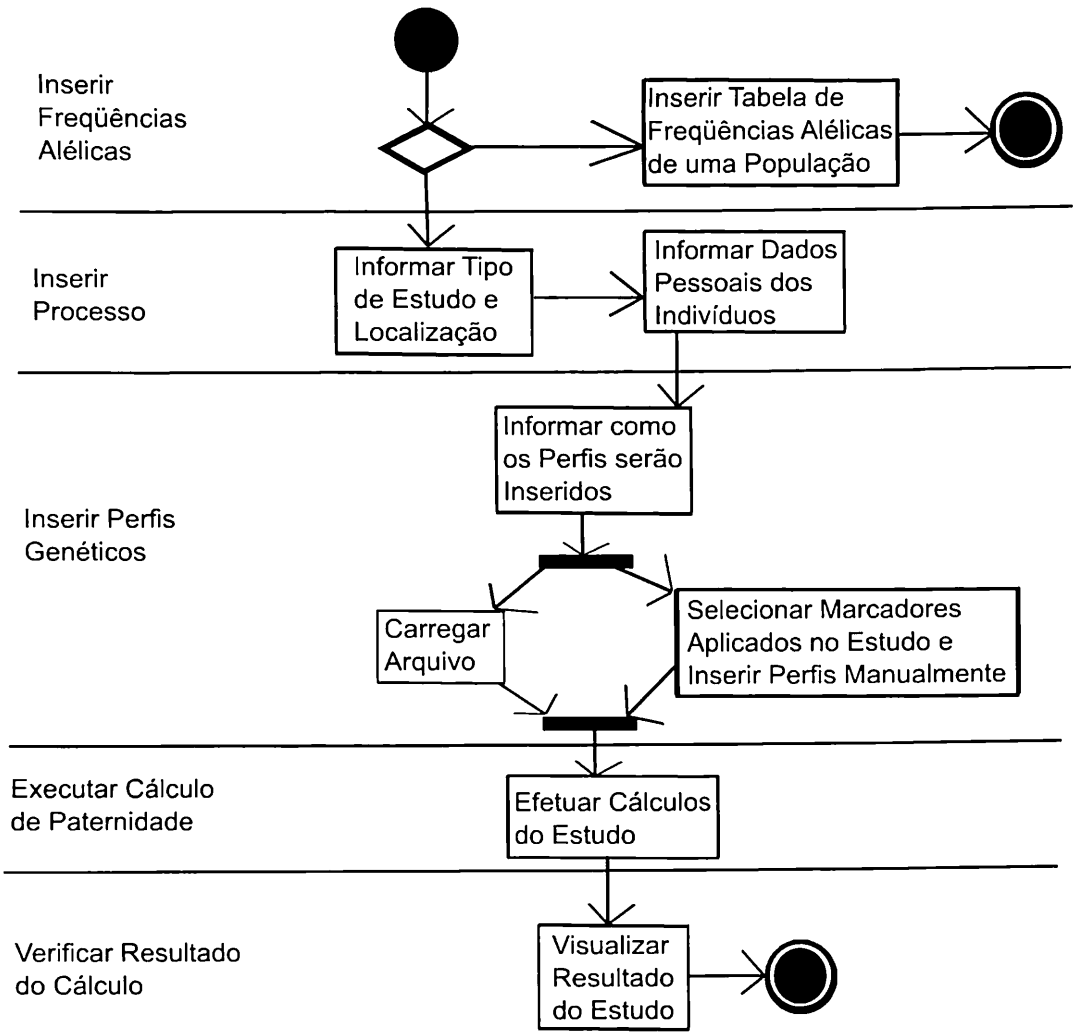

Figura 6.5: Diagrama de Atividades

No *Model* (modelo), encontram-se as classes que compõem o domínio do sistema, sendo, pois, o núcleo da aplicação.

Na camada *Persistence*, estão as classes responsáveis pelo armazenamento dos dados do sistema ao passo que em *Util* estão as classes utilitárias do sistema, responsáveis pelo carregamento das tabelas de freqüências alélicas e dos perfis genéticos via arquivos csv, assim como do cálculo de paternidade utilizando o `UnBBayes`.

Na modelagem do sistema, foram utilizados alguns padrões de projeto (*design patterns*) —descrições de objetos e classes comunicantes que precisam ser personalizadas para resolver um problema geral de projeto num contexto particular (Gamma et al., 2005). A utilização desses padrões teve por objetivo evitar problemas de projeto como, por exemplo, o acoplamento. Dentre os padrões usados, estão: *Facade*, *Factory Method*, *DAOFactory* e *Singleton*.

Facade O padrão *Facade* foi utilizado para simplificar o acesso aos objetos das camadas *Model*, *Persistence* e *Util*, diminuindo a complexidade do

sistema e o acoplamento entre os objetos que o compõem. Isso foi possível com a adição de novos objetos (chamados fachadas) que ficaram responsáveis pela comunicação entre os objetos das camadas mencionadas acima (*Model*, *Persistence* e *Util*) e os objetos que compõem a camada *View* do sistema (ver Figura 6.6).

Factory Method O padrão *Factory Method* foi utilizado no pacote *calculus* da camada *Util* (ver Figura 6.7). Ao ser chamado o método startCalculus(...) a partir de uma instância da classe PaternityStudy, não é possível prever a priori qual tipo de cálculo será utilizado e, conseqüentemente, não é possível saber qual classe deve ser instanciada para se executar o cálculo correto. Isso ocorre devido ao fato de haver mais de um tipo de cálculo, o que dependerá do estudo de paternidade em questão. O uso de dois tipos de algoritmos para o cálculo, sem que a classe PaternityStudy tenha que conhecer antecipadamente as classes que encapsulam tais algoritmos, sugere a aplicação do padrão *Factory Method* para solucionar esse problema, uma vez que com o emprego desse padrão é possível adiar a instanciação para as subclasses. Com isso, a instanciação em vez de ocorrer na classe PaternityStudy, ocorre na classe PaternityCalculusFactory que implementa a interface PaternityCalculusFactoryIF que, por sua vez, fornece serviço à classe PaternityStudy. A escolha da classe a ser instanciada (PaternityTestCaseOne ou PaternityTestCaseTwo) é feita por meio da análise dos objetos que compõem o atributo locusSetList o que está implementado no método calculusPaternityTest(...) da classe PaternityCalculusFactory. Ao final da execução desse método é retornado um objeto da classe PaternityTest que contém todos as informações resultantes do cálculo. Este objeto é então setado no atributo paternityTest da instância da classe PaternityStudy.

DAOFactory O padrão *DAOFactory* foi utilizado na camada *Persistence* (ver Figura 6.8). A diversidade de fontes de dados pode criar uma dependência da aplicação em relação ao código de acesso aos dados, uma vez que os componentes de negócio conterão código específico das APIs e dispositivos de armazenamento utilizados. Isto introduz um alto acoplamento entre a lógica de negócio e a implementação do acesso às fontes de dados, tornando mais difícil migrar a aplicação de uma fonte de dados para outra. Com o intuito de abstrair a camada de persistência, a fim de tornar possível a utilização de qualquer repositório de dados, foi utilizado o padrão DAOFactory juntamente com o *framework Hibernate*. Dessa forma, foi centralizado o serviço de persistência de objetos num pequeno con-

junto de classes, evitando por exemplo que o código SQL se espalhe pelo código da solução. Para modificar o repositório, basta fazer uma pequena alteração nas configurações do *Hibernate*.

Singleton O padrão *Singleton* foi utilizado na classe HibernateDAOFactory filha da classe DAOFactory, a fim de garantir que esta classe tenha uma única instância, fornecendo um ponto global de acesso para tal instância.

Para uma melhor visualização, o diagrama de classes foi dividido em três diagramas. O diagrama da Figura 6.6 mostra todas as camadas do sistema, todavia apenas as camadas *Model*, *View* e *Controller* estão detalhadas, o da Figura 6.7 apresenta o pacote *calculus* da camada *Util* e o da Figura 6.8 exibe a camada *Persistence* em detalhes.

Detalhes sobre os componentes desse tipo de diagrama UML podem ser obtidos em Booch et al. (2005).

6.5 Validação

Nesta seção, serão apresentados dois dos estudos de paternidade cedidos pelo Laboratório de DNA Forense da Universidade Federal de Alagoas que foram usados para validar o sistema THÊMIS. É importante ressaltar que os nomes e CPFs cadastrados no sistema e mostrados nessa seção não correspondem aos das pessoas que realizaram os estudos, objetivando não as expor.

Para a criação de um processo, é necessária a inserção de ao menos uma tabela de freqüências alélicas (ver seção 6.3.2). Sendo assim, antes de mostrar os dois estudos de paternidade (caso padrão e caso complexo), será mostrada a inserção da tabela de freqüências alélicas do Estado de Alagoas, local ao qual os indíviduos envolvidos nos estudos pertencem. Esta tabela também foi cedida pelo Laboratório de DNA Forense da UFAL.

6.5.1 Inserção da Tabela de Freqüências Alélicas

Para inserir uma tabela de freqüências alélicas basta acionar a opção *Inserir Tabela* no menu *Tabela de Freqüência* (ver Figura 6.9).

A inserção da tabela de freqüências é feita por meio de um arquivo no formato csv, sendo necessário para isso informar a localidade (ver Figura 6.10) e carregar o arquivo csv (ver Figura 6.11). Após o arquivo ser carregado, é possível ver todas as freqüências, além de ser permitido editar os valores mostrados antes de inseri-los no banco de dados (ver Figura 6.12). No Apêndice C (seção C.1) é mostrado o conteúdo desse csv.

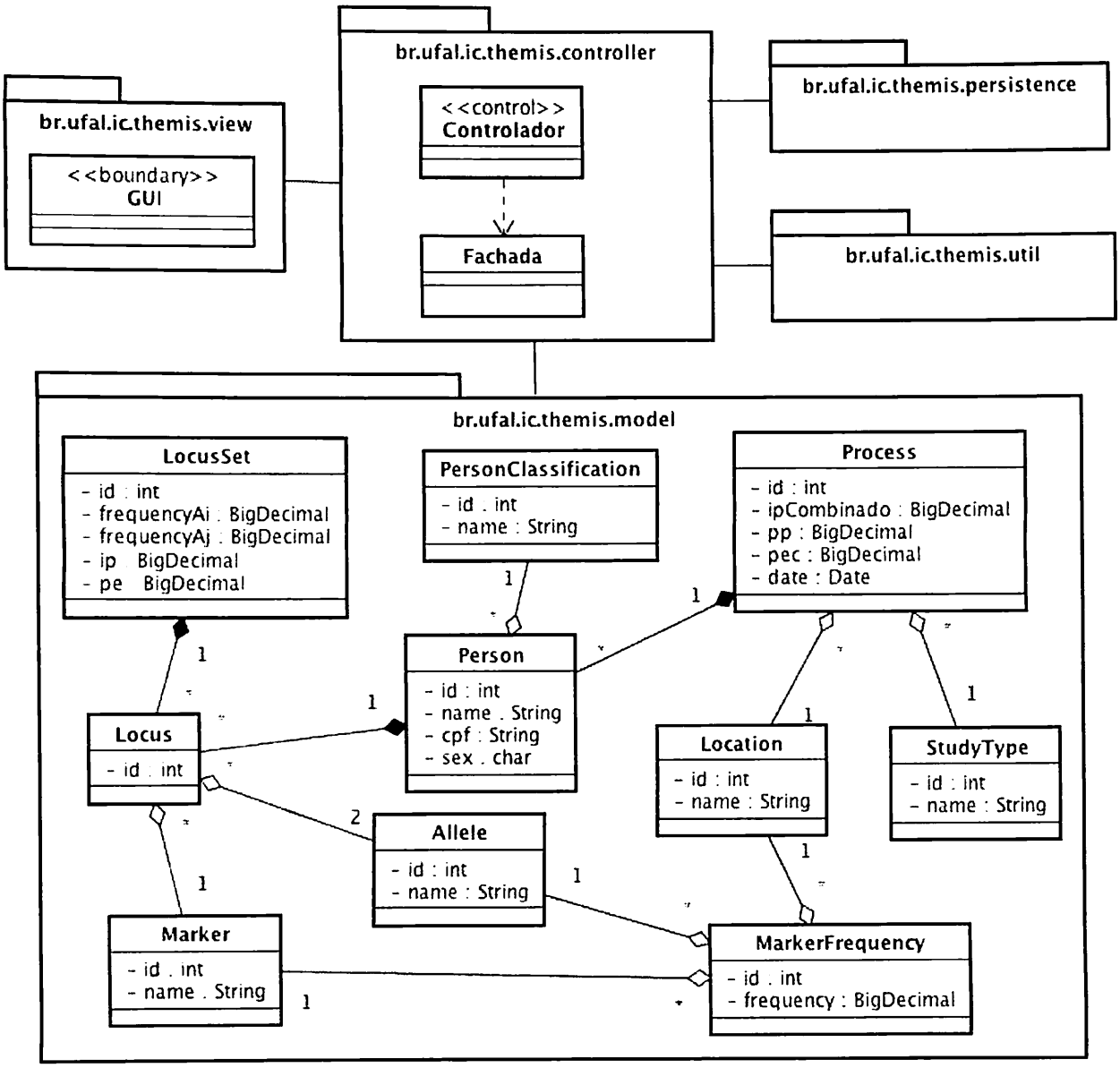

Figura 6.6: Diagrama de Classes (Geral)

Ocorrida a inserção da tabela, nas próximas seções serão mostrados os dois estudos de paternidade.

6.5.2 Estudo Caso Padrão

Conforme mostrado no diagrama de atividades (ver Figura 6.5), primeiro cria-se o processo, depois insere-se os perfis genéticos, em seguida faz-se o cálculo e, por fim, visualiza-se o resultado do estudo. É importante ressaltar que para essas atividades, apesar de serem seqüenciais e dependentes, não existem restrições de tempo entre elas, ou seja, após a execução de uma atividade, a atividade subseqüente pode ser realizada em qualquer momento.

Com base nos dados do estudo contido na Figura 6.13 é mostrada a seguir

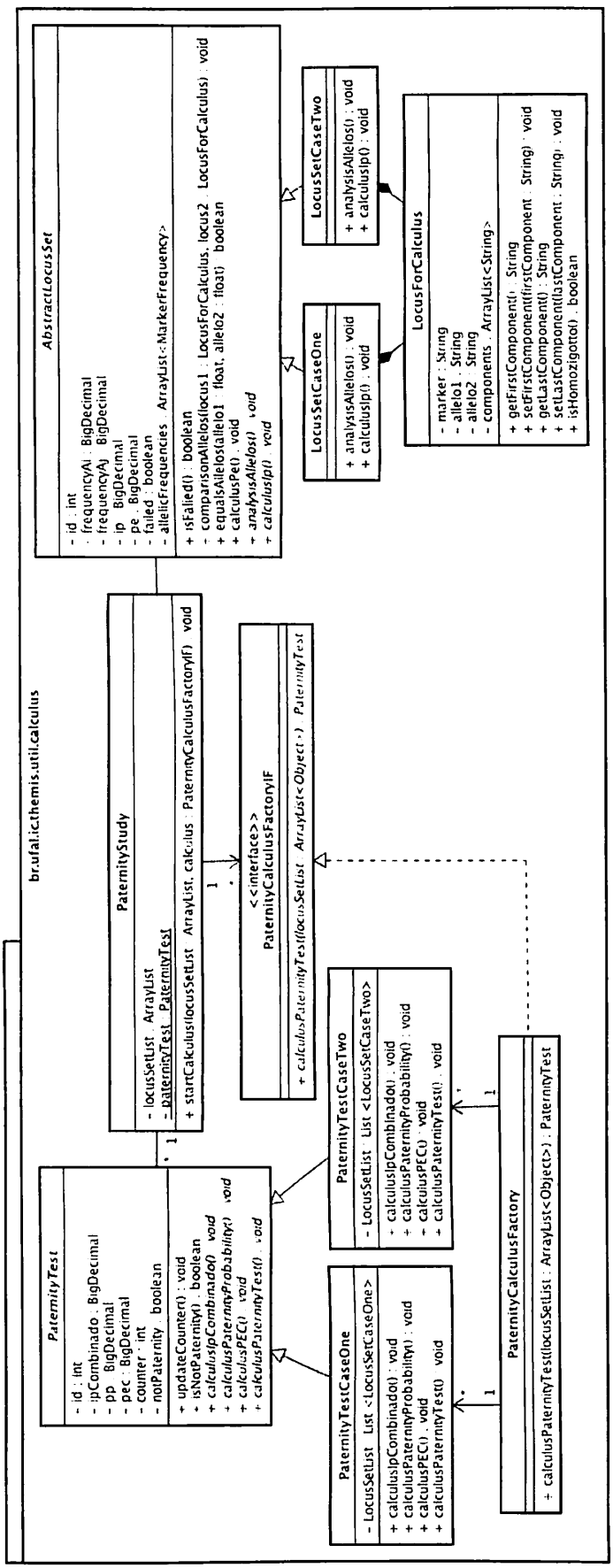

Figura 6.7: Diagrama de Classes (Calculus)

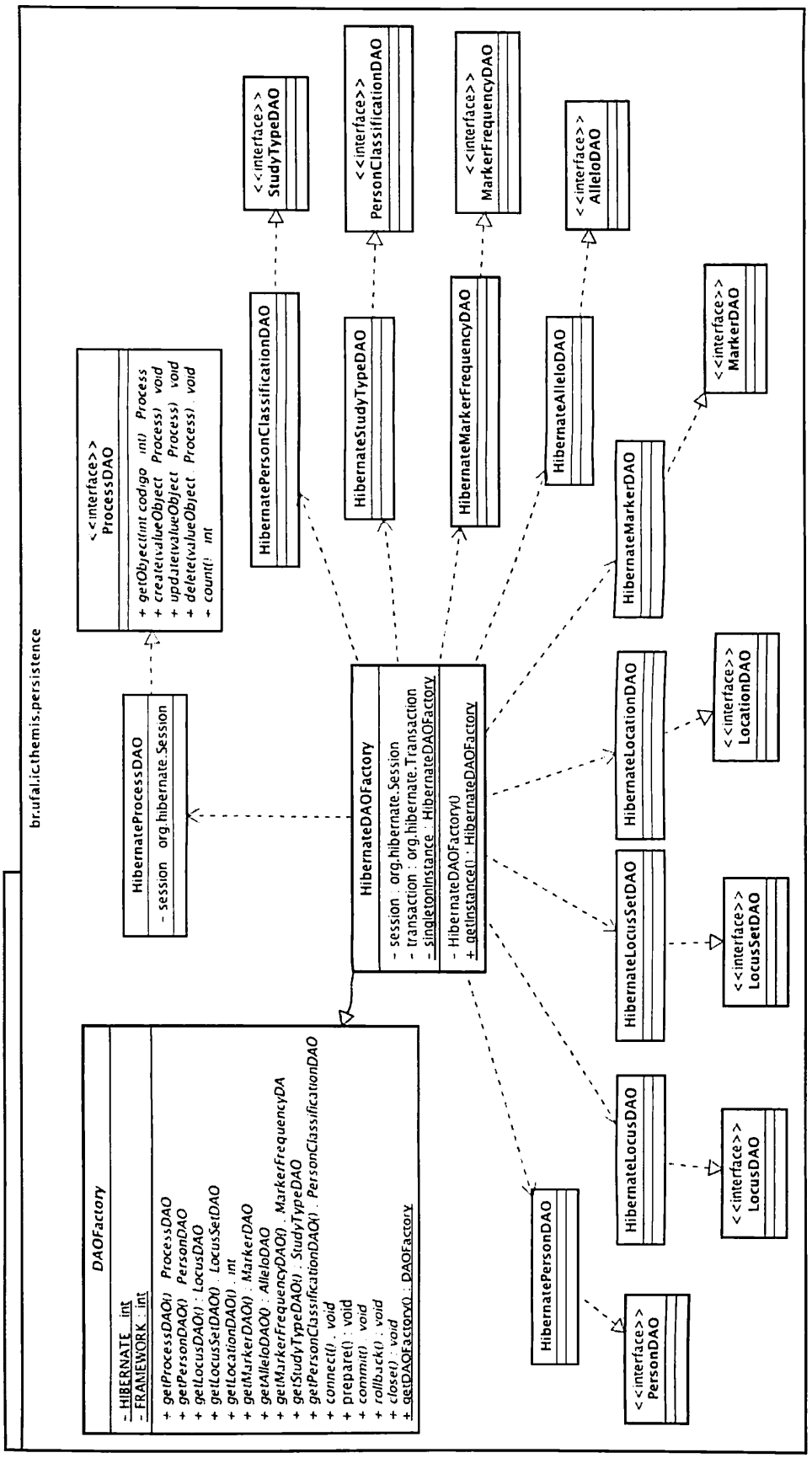

Figura 6.8: Diagrama de Classes (Persistence)

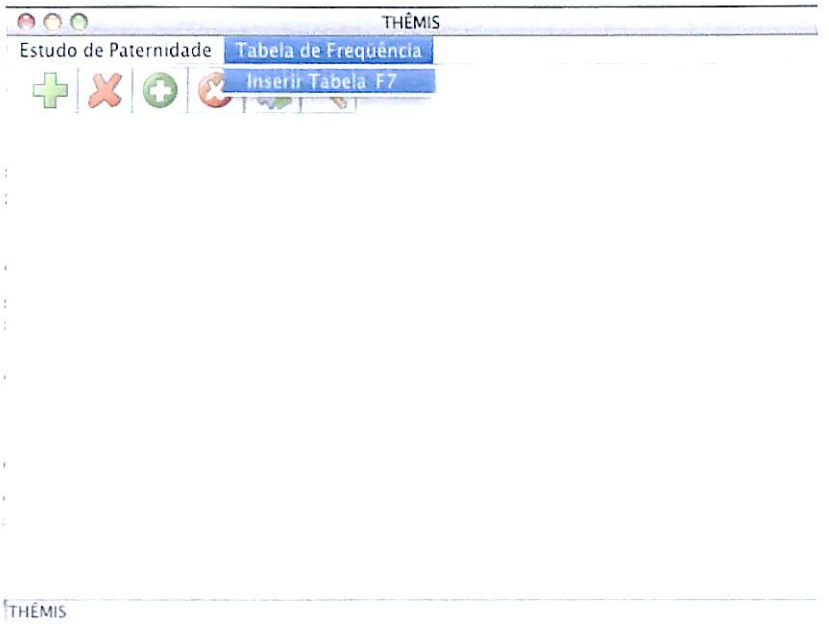

Figura 6.9: Opção Inserir Tabela

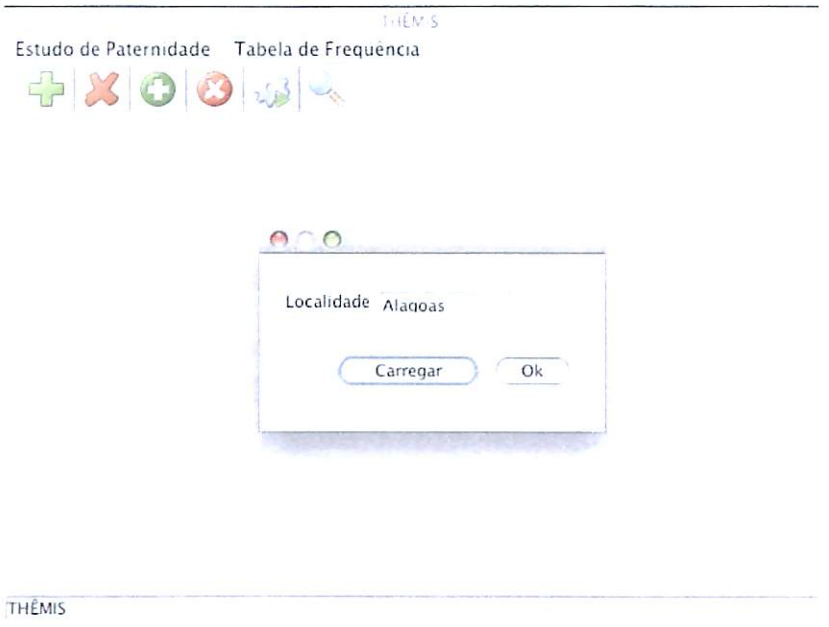

Figura 6.10: Informação da localidade

a execução de cada uma das etapas acima discutidas. É importante mencionar que, nas legendas das figuras, ECP equivale a Estudo Caso Padrão.

Criação do Processo

Para criar um processo basta acionar a opção *Inserir Processo* no menu *Estudo de Paternidade* (ver Figura 6.14).

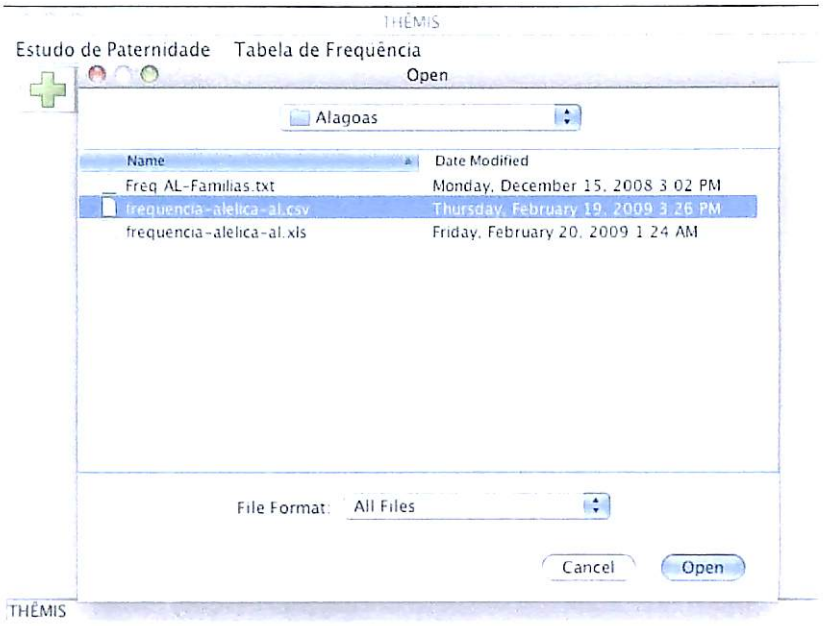

Figura 6.11: Carregamento do csv

A criação do processo consiste em duas etapas: a escolha do tipo de estudo de paternidade e do local ao qual as pessoas vinculadas ao estudo pertencem (ver Figura 6.15) e a inserção dos dados pessoais desses indivíduos (ver Figura 6.16).

Inserção dos Perfis Genéticos

Para inserir os perfis genéticos basta acionar a opção *Inserir Perfis* no menu *Estudo de Paternidade* (ver Figura 6.17), informar na próxima tela o número do processo que deseja (ver Figura 6.18) e na seguinte selecionar o modo de inserção (ver Figura 6.19).

Apenas por questão de praticidade e confiabilidade, a inserção dos perfis genéticos será feita via arquivo csv vindo do seqüenciador. O conteúdo desse arquivo é mostrado no Apêndice C (seção C.2).

Após o carregamento do csv, é possível visualizar os dados (ver Figura 6.20), bem como editá-los, caso seja necessário, antes de inseri-los no banco de dados.

Execução do Cálculo

Para executar o cálculo basta acionar a opção *Iniciar Cálculo* no menu *Estudo de Paternidade* (ver Figura 6.21) e logo após isso informar o número do processo que deseja efetuar o cálculo (ver Figura 6.18).

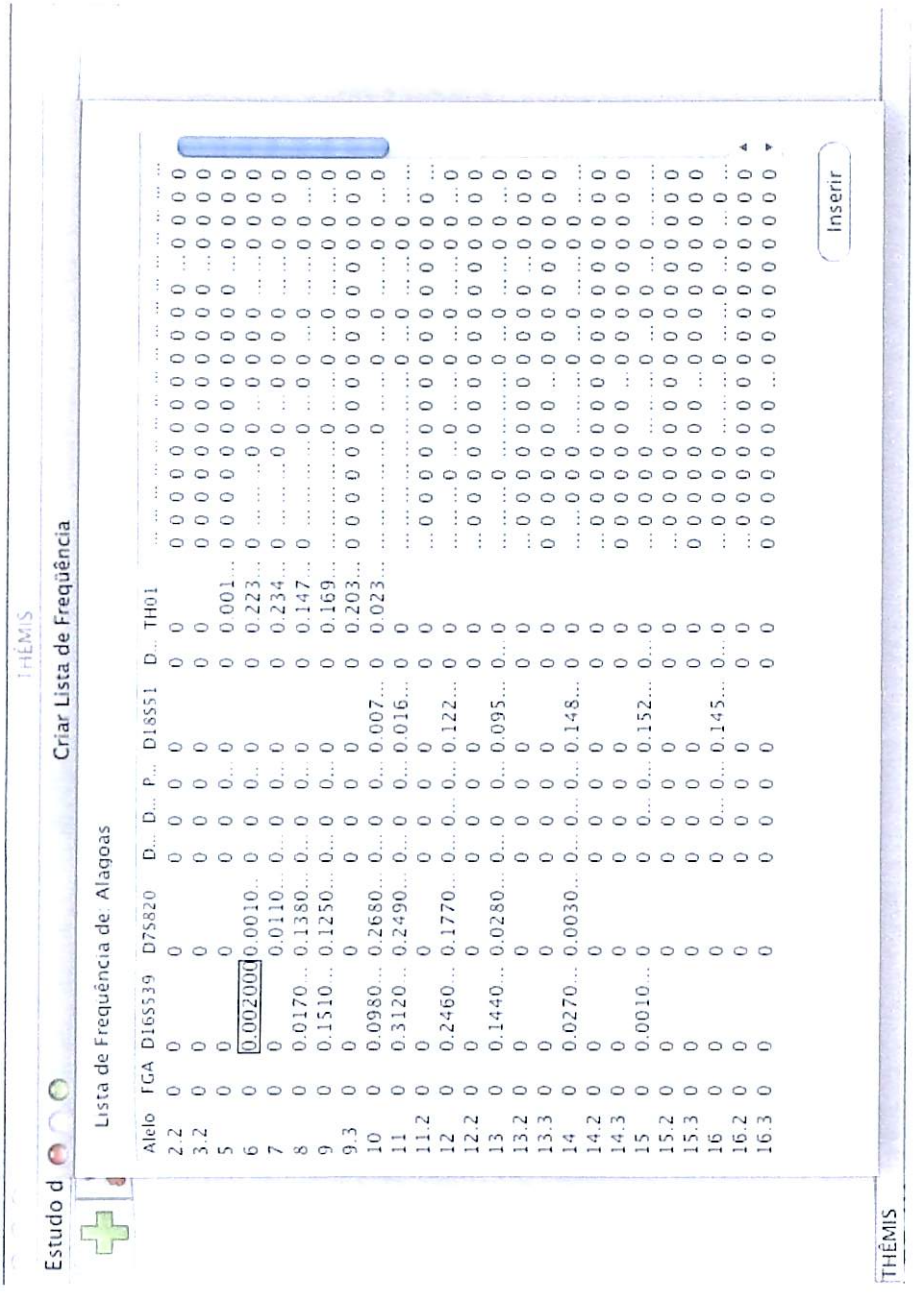

Figura 6.12: Tabela carregada

Verificação do Resultado

Para verificar o resultado do cálculo basta acionar a opção *Exibir Resultado* no menu *Estudo de Paternidade* (ver Figura 6.22) e logo após isso informar o número do processo que deseja ver o resultado do cálculo (ver Figura 6.18). A Figura 6.23 mostra o resultado do cálculo para esse estudo. Observe que os valores são os mesmos contidos na Figura 6.13 na página 86 salvo os arredondamentos contidos no arquivo original com os dados do estudo.

Índice de Paternidade (IP) Combinado 1,241,571,137.494

Probabilidade de Patenidade 99.9999999%

TABELA DE ALELOS

Marcadores	Mãe		Filho(a)		Suposto Pai		IP*	
FGA	21	22	22	23	22	23	4.098	
D16S539	8	10	9	10	9	11	3.311	
D7S820	8	10	8	10	9	10	1.232	
D13S317	11	12	8	11	8	11	5.952	
D3S1358	15	15	15	16	15	16	1.748	
Penta E	5	13	13	19	5	19	45.455	
D18S51	19	21	12	19	12	14	4.098	
D12S391	18	21	18	19	19	23	3.247	
TH01	7	9	9	9	8	9	2.959	
D19S433	13	13	15	15	14	15	1.000	Mutação
CSF1PO	11	13	10	11	10	12	1.773	
TPOX	8	11	8	12	11	12	10.000	
F13B	9	10	9	9	9	9	3.922	
LPL	10	11	10	11	11	11	1.508	
vWA	17	18	18	19	16	19	9.434	
D10S2325	10	12	10	13	7	13	4.032	

*IP = Índice de Paternidade

Nº de Marcadores Analisados 16

Figura 6.13: Dados para estudo caso padrão

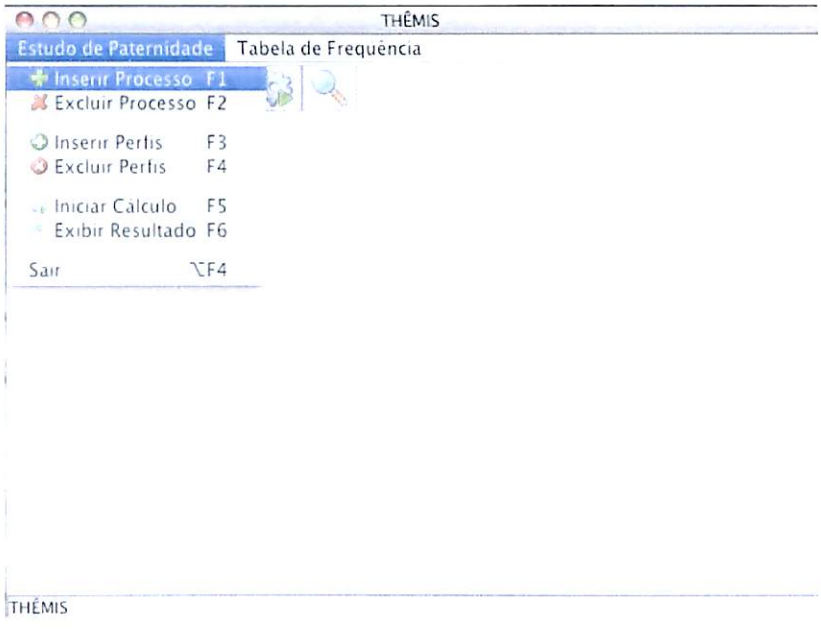

Figura 6.14: Opção Inserir Processo

6.5.3 Estudo Caso Complexo

Com base nos dados do estudo contido na Figura 6.24 (onde PSP é a sigla para suposto avô paterno e MSP, para suposta avó paterna) é mostrada a seguir a execução de cada uma das etapas envolvidas nos processos de estudo de paternidade seguindo as mesmas diretrizes da seção anterior. É importante mencionar que, nas legendas das figuras, ECC equivale a Estudo Caso

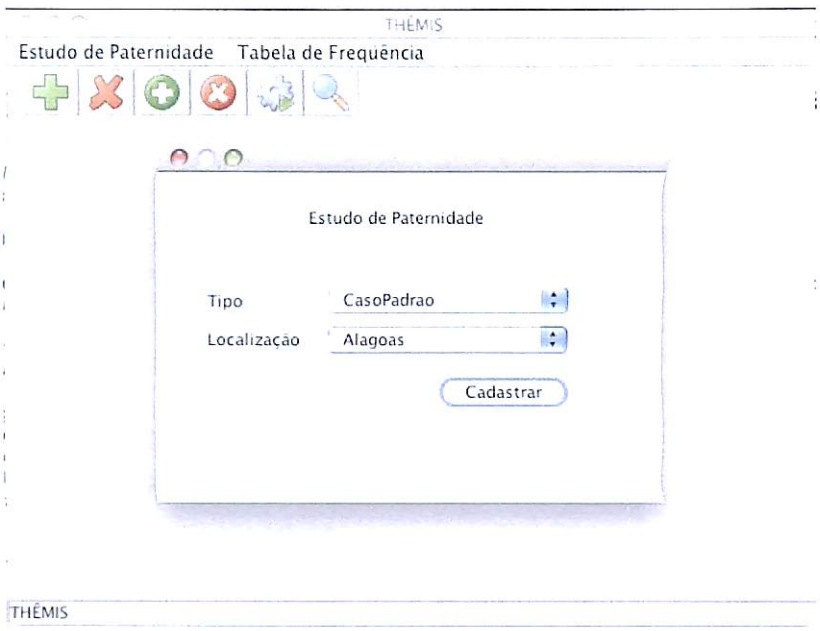

Figura 6.15: Inserção processo - Parte 1 (ECP)

Complexo.

Criação do Processo

Para criar um processo basta acionar a opção *Inserir Processo* no menu *Estudo de Paternidade* (ver Figura 6.14 na página 86).

A criação do processo consiste em duas etapas: a escolha do tipo de estudo de paternidade e do local ao qual as pessoas vinculadas ao estudo pertencem (ver Figura 6.25) e a inserção dos dados pessoais desses indivíduos (ver Figura 6.26).

Inserção dos Perfis Genéticos

Para inserir os perfis genéticos basta acionar a opção *Inserir Perfis* no menu *Estudo de Paternidade* (ver Figura 6.17 ná página 89), informar na próxima tela o número do processo que deseja (ver Figura 6.18 na página 89) e na seguinte selecionar o modo de inserção (ver Figura 6.19 na página 90).

Apenas por questão de praticidade e confiabilidade, a inserção dos perfis genéticos será feita via arquivo csv vindo do seqüenciador. O conteúdo desse arquivo é mostrado no Apêndice C (seção C.3).

Após o carregamento do csv, é possível visualizar os dados (ver Figura 6.27), bem como editá-los, caso seja necessário, antes de inseri-los no banco de dados.

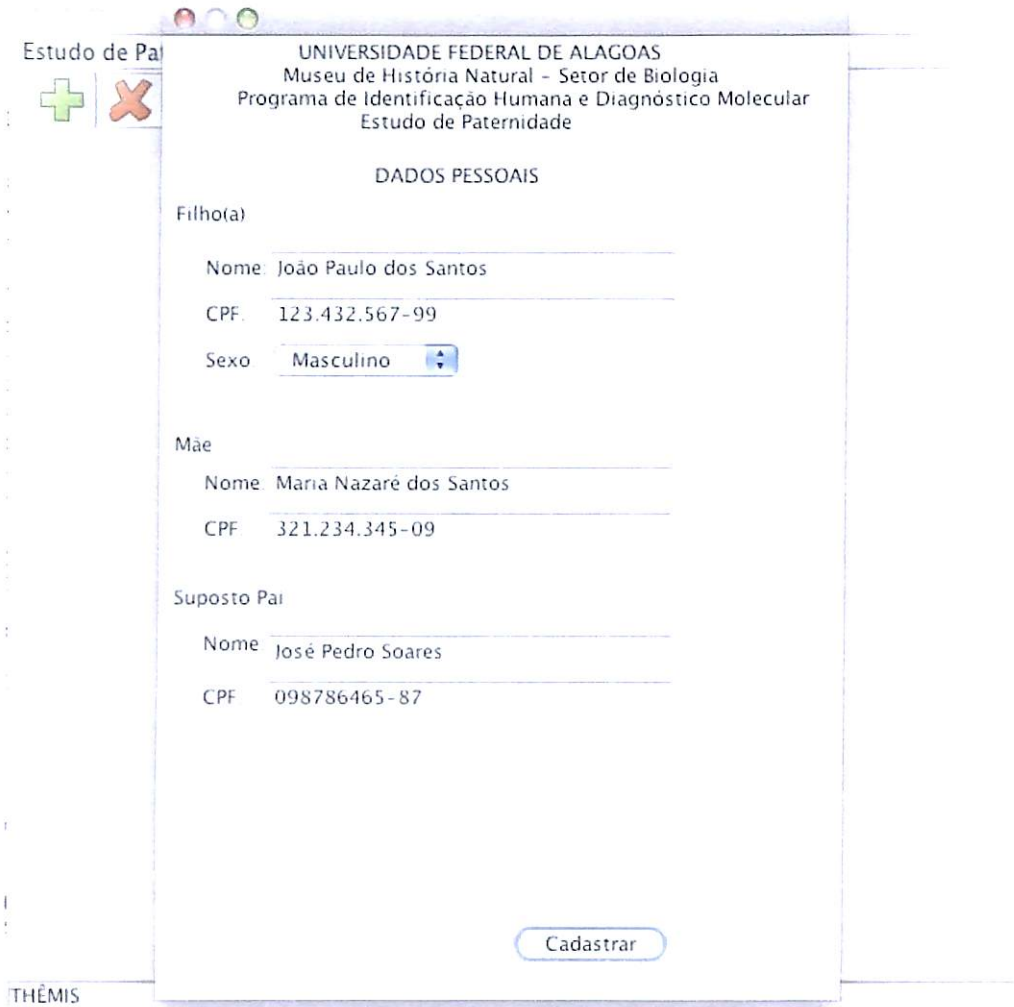

Figura 6.16: Inserção processo - Parte 2 (ECP)

Execução do Cálculo

Para executar o cálculo basta acionar a opção *Iniciar Cálculo* no menu *Estudo de Paternidade* (ver Figura 6.21 na página 92) e logo após isso informar o número do processo que deseja efetuar o cálculo (ver Figura 6.18 na página 89).

Verificação do Resultado

Para verificar o resultado do cálculo basta acionar a opção *Exibir Resultado* no menu *Estudo de Paternidade* (ver Figura 6.22 na página 92) e logo após isso informar o número do processo que deseja ver o resultado do cálculo (ver Figura 6.18 na página 89). A Figura 6.28 mostra o resultado do cálculo para esse estudo. Observe que os valores são os mesmos contidos na Figura 6.24 na página 94 salvo os arredondamentos contidos no arquivo original com os dados do estudo.

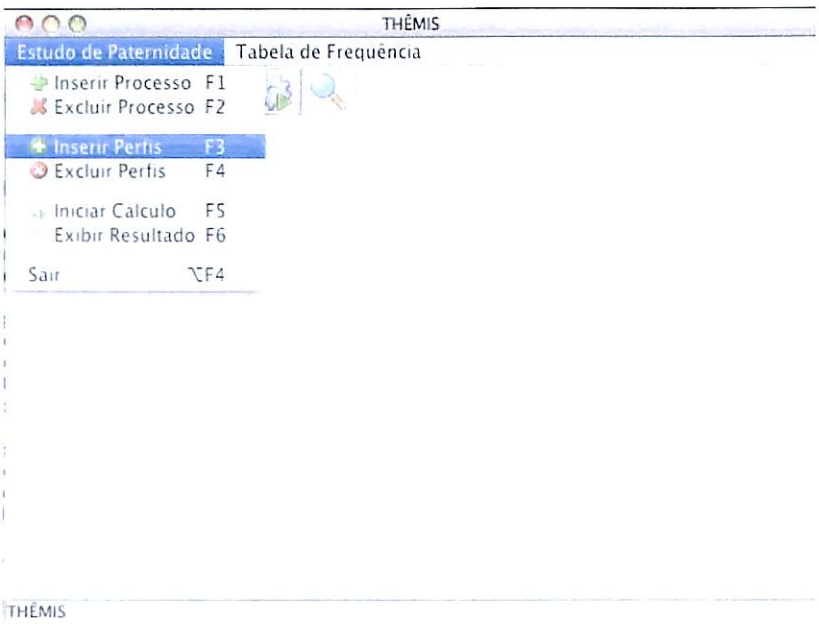

Figura 6.17: Opção Inserir Perfis

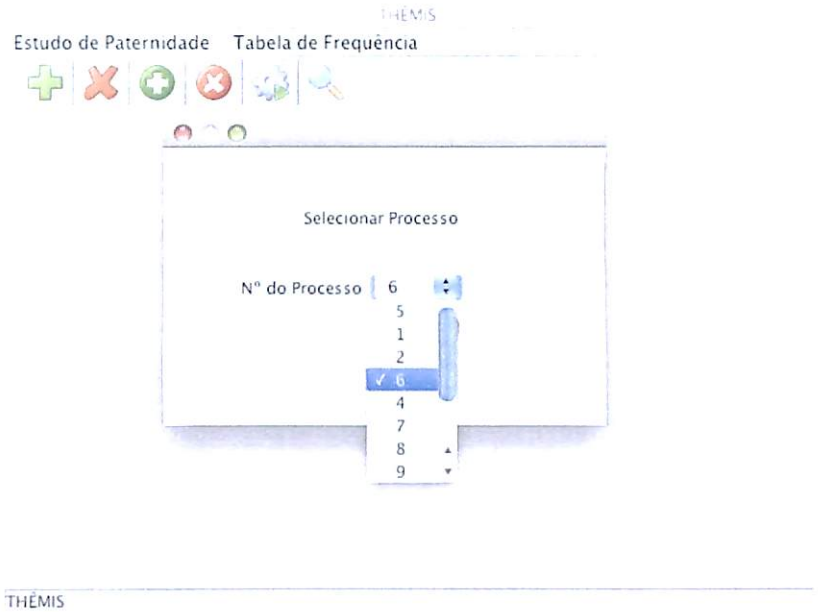

Figura 6.18: Seleção do número do processo

6.6 Comentários

Neste capítulo, foram apresentadas a descrição do sistema THÊMIS utilizando linguagem natural, as tecnologias usadas na implementação desse software, as funções e restrições desse sistema, bem como o seu uso na execução de dois dentre os vários estudos de paternidade reais usados para validá-lo.

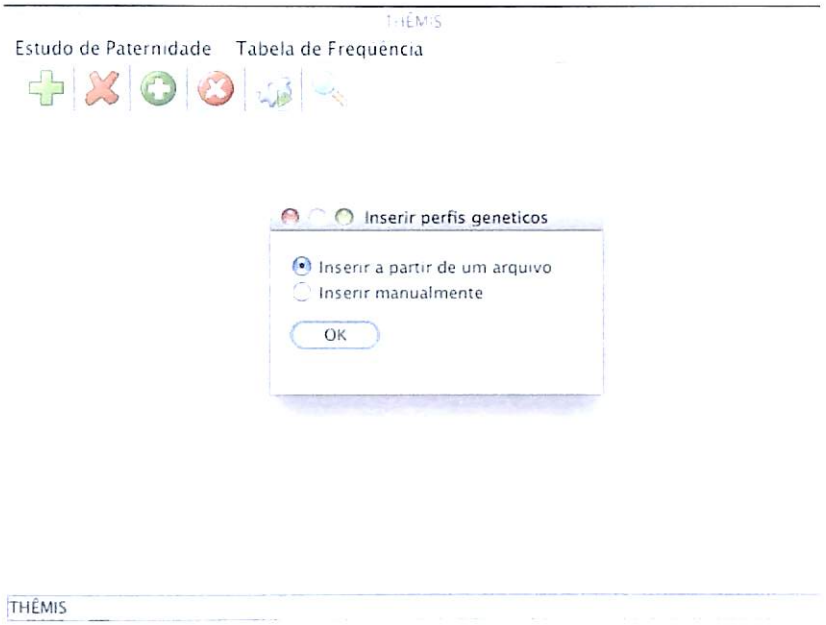

Figura 6.19: Seleção do modo de inserção

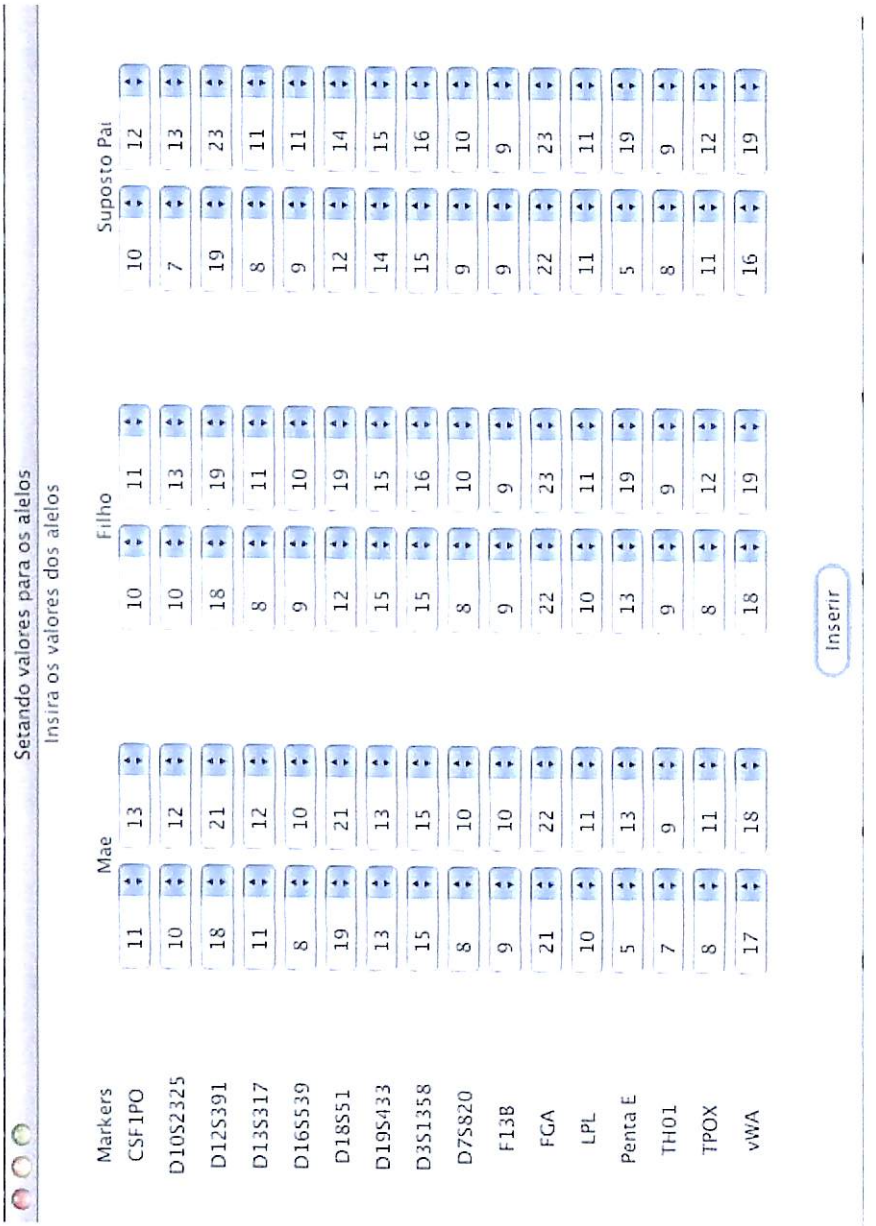

Figura 6.20: Visualização dos perfis inseridos (ECP)

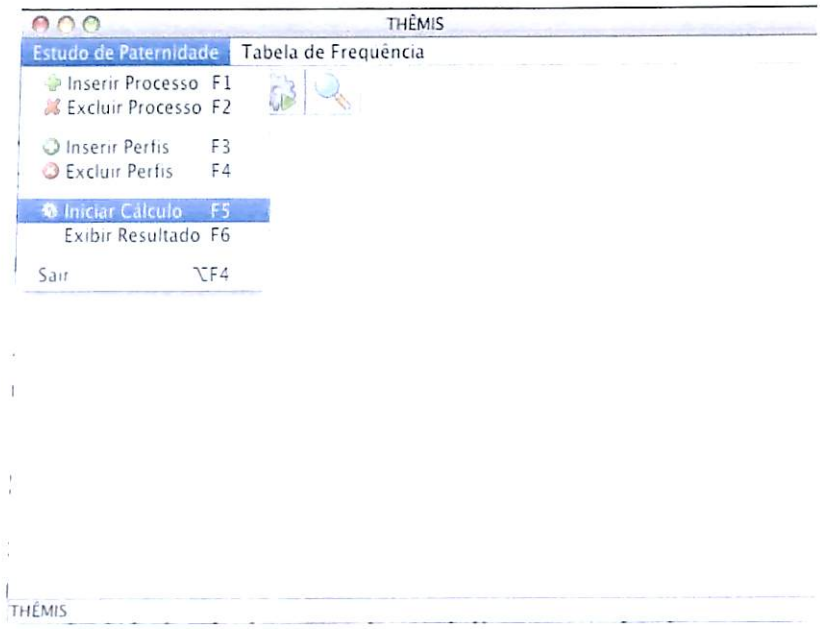

Figura 6.21: Opção Iniciar Cálculo

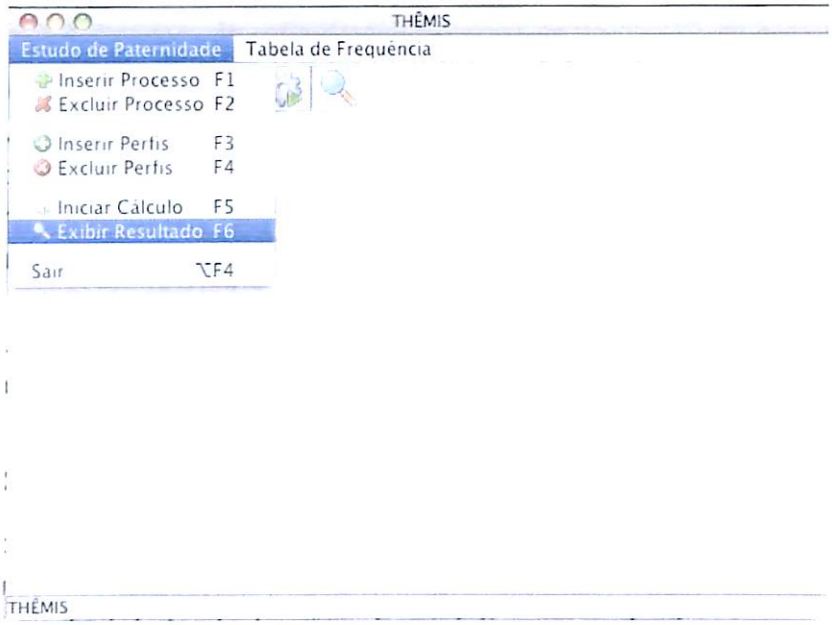

Figura 6.22: Opção Exibir Resultado

Resultado do Estudo de Paternidade

Resultado do Estudo de Paternidade

Markers	Mae		Filho		Suposto Pai		Freq. Ai	Freq. Aj	Ip	Pe
CSF1PO	11	13	10	11	10	12	0.2820000	0.2740000	1.7730495	0.1971360
D10S2325	10	12	10	13	7	13	0.1420000	0.1240000	4.0322579	0.5387560
D12S391	18	21	18	19	19	23	0.2230000	0.1540000	3.2467539	0.3881290
D13S317	11	12	8	11	8	11	0.0840000	0.2900000	5.9523803	0.3918760
D16S539	8	10	9	10	9	11	0.1510000	0.0980000	3.3112584	0.5640010
D18S51	19	21	12	19	12	14	0.1220000	0.0560000	4.0983599	0.6756840
D19S433	13	13	15	15	14	15	0.1434290	0.1434290	1.0000000	0.5085715
D3S1358	15	15	15	16	15	16	0.2980000	0.2860000	1.7482517	0.1730560
D7S820	8	10	8	10	9	10	0.1380000	0.2680000	1.2315271	0.3528360
F13B	9	10	9	9	9	9	0.2550000	0.2550000	3.9215683	0.2401000
FGA	21	22	22	23	22	23	0.1440000	0.1220000	4.0983602	0.5387560
LPL	10	11	10	11	11	11	0.3830000	0.2800000	1.5082958	0.1135690
Penta E	5	13	13	19	5	19	0.1520600	0.0110800	45.126351	0.7003346
TH01	7	9	9	9	8	9	0.1690000	0.1690000	2.9585800	0.4382440
TPOX	8	11	8	12	11	12	0.4730000	0.0500000	9.9999995	0.2275290
vWA	17	18	18	19	16	19	0.1690000	0.0530000	9.4339625	0.6052840

IPC 1232606681.0000

Pp 99.999999920000

Pec 99.991828300000

Fechar

Figura 6.23: Resultado do cálculo (ECP)

TABELA DE ALELOS

Marcadores	Mãe		Filha		PSP		MSP		RV
FGA	22	25	23	25	23	24	20	21	2.04918
D16S539	9	9	9	9	9	12	9	12	3.31126
D7S820	11	12	11	12	9	11	9	11	1.17371
D13S317	10	11	10	11	10	13	8	9	0.71839
D3S1358	15	17	15	17	14	17	15	18	1.04384
Penta E	5	11	10	11	7	10	12	19	4.43341
D18S51	14	16	12	16	12	15	14	15	2.04918
D12S391	19	22	15	19	15	19	17	20	7.57576
TH01	8	9	8	9	7	8	7	8	1.58228
D19S433	12	14	12	16	15.2	16	15	16	10.82720
CSF1PO	8	12	11	12	11	12	12	12	0.91241
TPOX	8	8	8	8	10	10	8	12	0.52854
F13B	9	10	6	10	10	11	6	10	1.52439
LPL	-	-	-	-	-	-	-	-	-
vWA	16	16	14	16	15	17	14	19	4.23729
D10S2325	10	13	9	10	9	14	11	12	2.21239
D21S11	-	-	-	-	-	-	-	-	-
D8S1179	-	-	-	-	-	-	-	-	-
D2S1338	20	22	22	23	17	23	19	20	2.51765
D5S818	13	13	10	13	10	10	10	12	14.40937
SE33	19	19	19	21.2	24.2	25.2	21.2	27.2	25.00000

RV = Razão de Verossimilhança

Marcadores Analisados = 18
Razão de Verossimilhança Total = 44,011,490.342
Probabilidade = 99.999998%

Figura 6.24: Dados para estudo caso complexo

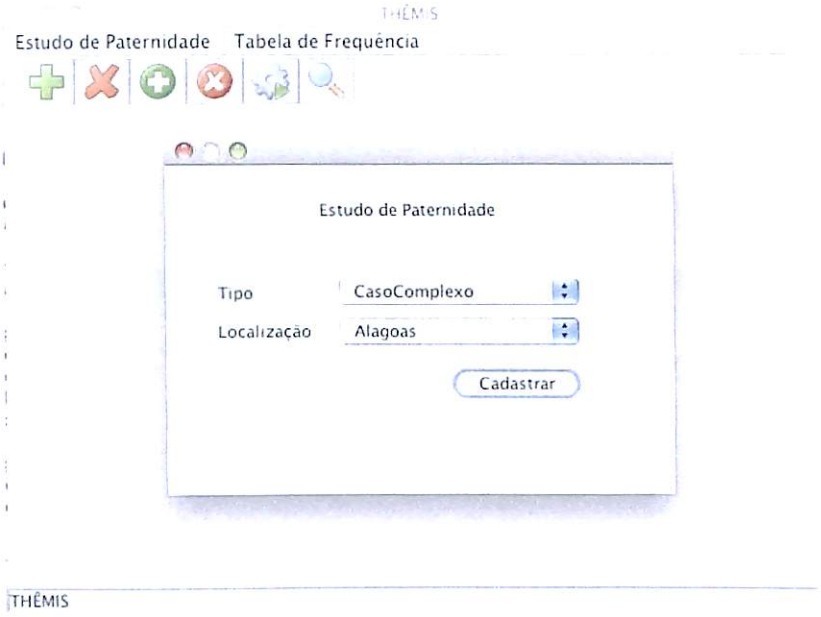

Figura 6.25: Inserção processo - Parte 1 (ECC)

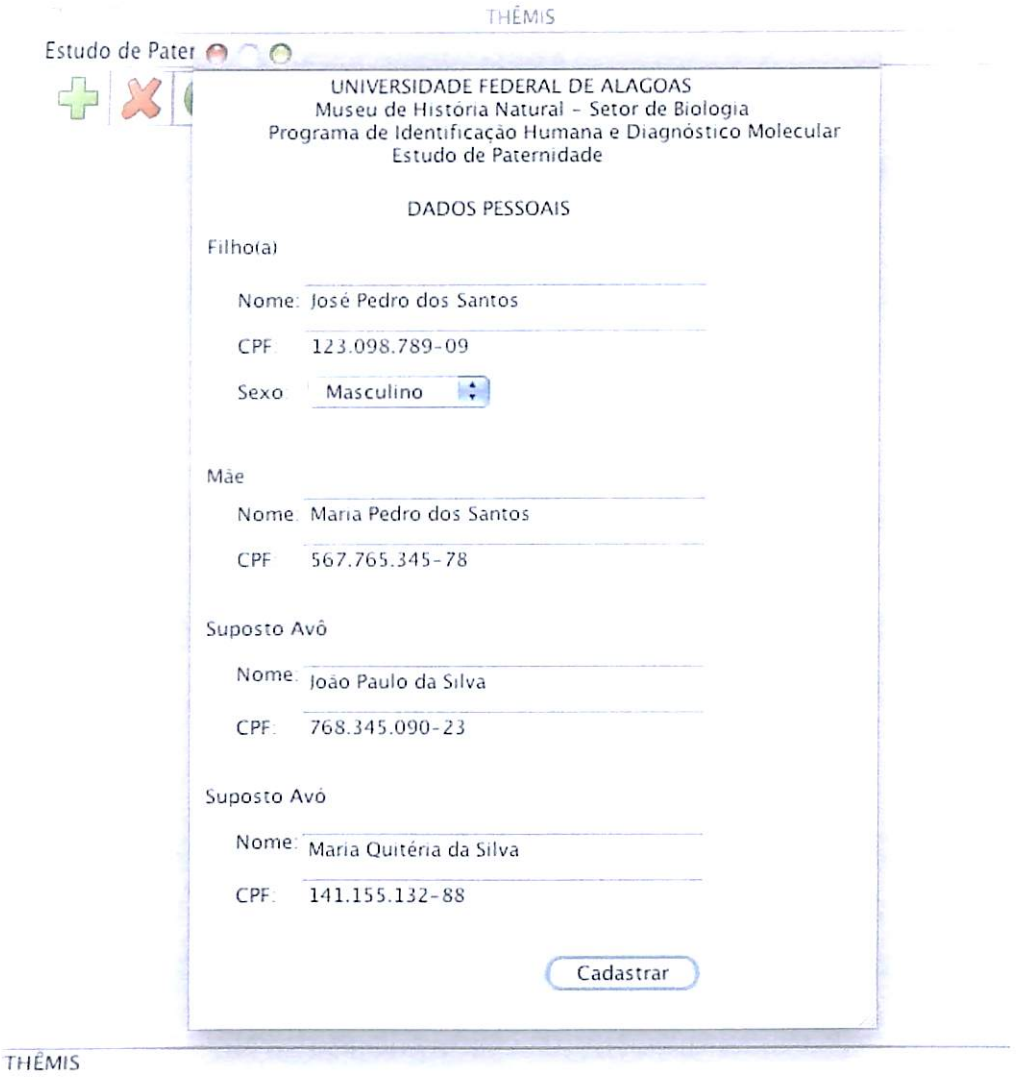

Figura 6.26: Inserção processo - Parte 2 (ECC)

Figura 6.27: Visualização dos perfis inseridos (ECC)

Resultado do Estudo de Paternidade

Resultado do Estudo de Paternidade

Markers	Mae		Filho		Suposto Avô		Suposta Avó		Freq. Ai	Freq. Aj	Pe	Ip
CSF1PO	8	12	11	12	11	12	12	12	0.2740000	0.3320000	0.1552360	0.9124089
D10S2325	10	13	9	10	9	14	11	12	0.1130000	0.1420000	0.5550250	2.2123895
D12S391	19	22	15	19	15	19	17	20	0.0330000	0.1540000	0.6609690	7.5757577
D13S317	10	11	10	11	10	13	8	9	0.0580000	0.2900000	0.4251040	0.7183908
D16S539	9	9	9	9	9	12	9	12	0.1510000	0.1510000	0.4872040	3.3112585
D18S51	14	16	12	16	12	15	14	15	0.1220000	0.1450000	0.5372890	2.0491804
D19S433	12	14	12	16	15.2	16	15	16	0.0640670	0.0439810	0.7955783	11.368543
D2S1338	20	22	22	23	17	23	19	20	0.0952900	0.0993000	0.6486852	2.5176234
D3S1358	15	17	15	17	14	17	15	18	0.2980000	0.1810000	0.2714410	1.0438413
D5S818	13	13	10	13	10	10	10	12	0.0520500	0.1781800	0.5925458	14.409221
D7S820	11	12	11	12	9	11	9	11	0.2490000	0.1770000	0.3294760	1.1737088
F13B	9	10	6	10	10	11	6	10	0.1640000	0.3280000	0.2580640	1.5243902
FGA	22	25	23	25	23	24	20	21	0.1220000	0.1280000	0.5625000	2.0491802
Penta E	5	11	10	11	7	10	12	19	0.0563900	0.0906300	0.7275748	4.4334103
SE33	19	19	19	21.2	24.2	25.2	21.2	27.2	0.0800000	0.0100000	0.8281000	24.999997
TH01	8	9	8	9	7	8	7	8	0.1470000	0.1690000	0.4678560	1.5822783
TPOX	8	8	8	8	10	10	8	12	0.4730000	0.4730000	0.0029160	0.5285412
vWA	16	16	14	16	15	17	14	19	0.0590000	0.2780000	0.4395690	4.2372872

IPC 46211087.

Pp 99.999997

Pec 99.999887

Fechar

Figura 6.28: Resultado do cálculo (ECC)

Capítulo 7

Considerações Finais

A representação do conhecimento por meio de computadores é, na grande maioria dos casos, uma tarefa bastante complexa, uma vez que exige, além do conhecimento acurado do domínio a ser representado, o entendimento de diversos conceitos cuja compreensão é de fundamental importância para a construção de uma boa representação. Assim como as abordagens para representação, os domínios são muito diversos. Devido à infinidade de domínios existentes, os quais possuem muitas características peculiares, deve haver grande cautela na escolha da abordagem a ser utilizada para representá-los.

Nos domínios gerados na análise forense de DNA em estudos de paternidade, há um certo grau de incerteza, haja vista que apesar de cada ser humano possuir um perfil genético único, em estudos forenses é impossível se analisar todo o genótipo dos indivíduos envolvidos, havendo, pois, a necessidade de se utilizar inferências na análise dos dados.

A rede bayesiana é um método interessante para representar domínios incertos com alta complexidade, uma vez que une a teoria dos grafos, que permite descrever graficamente as relações de dependência entre as variáveis, e a teoria das probabilidades, que atribui níveis de crença às variáveis. Esta abordagem estocástica para a incerteza permite por meio de inferências a obtenção de conclusões relevantes que não são evidentes observando-se apenas os dados de entrada.

Por ser uma rede probabilística, a rede de crença bayesiana apresenta as seguintes vantagens em relação a outras abordagens para representação do conhecimento: representação e manipulação da incerteza baseadas em modelos matemáticos, modelagem do conhecimento de forma intuitiva acerca do domínio e permissão à realização de inferência causal, de diagnóstico, intercausal e mista.

Buscando auxiliar os biólogos na análise forense de DNA, bem como dar

continuidade aos estudos e às atividades desenvolvidas no Instituto de Computação da Universidade Federal de Alagoas na área de Bioinformática e no uso das Redes Bayesianas como uma abordagem estocástica para a incerteza, foi modelado, implementado e validado o sistema THÊMIS, o qual utiliza o ferramental das redes bayesianas como meio de representação do conhecimento acerca de estudos de paternidade, obtendo por meio de inferências os resultados requeridos pela genética forense no que tange ao cálculo do IP.

Diferentemente do `familias`, que é utilizado juntamente com planilhas eletrônicas pelo Laboratório de DNA Forense da Universidade Federal de Alagoas, no THÊMIS as informações sobre os estudos são armazenadas numa base de dados e não é exigido do usuário a construção da rede que representa a genealogia em questão. Dessa forma, os dados inseridos no sistema podem ser utilizados em análises futuras e o usuário não necessita ter conhecimento algum sobre redes bayesianas, ou seja, o uso desse formalismo fica transparente ao usuário.

Por utilizar as redes bayesianas na análise estatística dos perfis genéticos, o sistema THÊMIS, que atualmente contempla dois tipos de estudo de paternidade com o uso de uma única topologia de rede, pode dar suporte a outros tipos de estudo de paternidade com uma simples extensão da topologia da rede bayesiana atual.

É importante mencionar que o uso do THÊMIS torna o processo atual executado pelo Laboratório de DNA Forense da Universidade Federal de Alagoas mais rápido e seguro, pois minimiza erros humanos na digitação dos dados ou até mesmo evita-os ao usar arquivos no formato csv para importação destes. Além de utilizar um único aplicativo, o THÊMIS, em substituição ao conjunto {*familias*, planilhas eletrônicas}.

Em se tratando dos resultados obtidos com o sistema THÊMIS a partir dos dados reais cedidos e anteriormente estudados pelo referido laboratório, deve-se ressaltar que foram obtidos os mesmos resultados na análise.

Este trabalho gerou a publicação de dois artigos: um no VIII Encontro Regional de Matemática Aplicada e Computacional (ERMAC 2008), ocorrido de 20 a 22 de novembro de 2008 em Natal–RN (ver Santos Júnior et al., 2008), e outro no V Simpósio Brasileiro de Sistemas de Informação (SBSI 2009), a se realizar entre os dias 20 e 22 de maio do presente ano na cidade de Brasília–DF (ver Santos Júnior et al., 2009). Ambos os artigos podem ser utilizados como base nos estudos que visam à análise de vínculo genético em estudos de paternidade usando a rede de crença bayesiana como abordagem para a representação do conhecimento.

Dentre os trabalhos futuros, pode-se citar:

- a análise da confiabilidade numérica do software;
- a análise do intervalo de confiança do IP;
- a adição de novos tipos de estudo de paternidade no sistema, o qual continuará a utilizar um modelo probabilístico único, diferenciando o tipo de estudo a partir das evidências inseridas;
- a incorporação de mutação em quaisquer tipos de estudo de paternidade em que se faça necessária essa análise;
- a construção de um módulo para a identificação de criminosos por meio do DNA obtido a partir de vestígios deixados no local do crime. Esse módulo proverá serviços similares aos do *Combined DNA Index System–CODIS* (`http://www.fbi.gov/hq/lab/html/codis1.htm`), realizando estudos de coincidência de perfis genéticos e armazenando esses perfis numa base de dados única, desde que a legislação vigente no Brasil venha a permitir a geração de um banco de dados de perfis de criminosos;
- a construção de um módulo para a identificação de indivíduos desaparecidos por meio da comparação dos perfis genéticos desses indivíduos com os de pessoas que tenham parentes desaparecidos.

Apêndice A

Ferramentas para a Computação de Redes Bayesianas

Dentre as ferramentas que provêem suporte à construção de redes bayesianas e à inferência nesse modelo probabilístico, foram analisados o `UnBBayes`, o qual foi utilizado como motor de inferência do sistema THÊMIS, o `Weka` e o `BayesBuilder`. Nas seções seguintes será dada um breve descrição dessas ferramentas.

A.1 UnBBayes

O `UnBBayes` (Universidade de Brasília, n.d.) é um software que permite a edição e a compilação de redes bayesianas (BN), diagramas de influências (ID) ou redes bayesianas múltiplas secionadas (MSBN), a entrada e propagação de evidências, a realização de inferência e a aprendizagem da topologia e/ou parâmetros de uma BN. Para propagação de evidências, esta ferramenta utiliza o método da árvore de junções; ao passo que para a aprendizagem de BN, são utilizados os algoritmos *K2* e *B* (baseados em métodos de busca e pontuação) e *CBL-A* e *CBL-B* (baseados em independência condicional).

Este sistema foi implementado em Java, documentado com Javadoc e JavaHelp, e se encontra disponível sob licença *GNU General Public License (GPL)*. Toda a documentação da API pode ser acessada através da ajuda ao usuário.

Este software foi criticado em Costa (2007) cuja alegação era que o software apresentara resultados inválidos quando utilizado na obtenção de informações sobre a rede bayesiana usada no estudo de caso da referida monografia. Todavia, essa alegação fora equivocada, haja vista que os resultados inválidos foram gerados não por falha da ferramenta, mas sim por erro humano na hora da inserção das probabilidades em uma das tabelas de probabilidade

condicional da rede.

A.2 Weka

O Waikato Environment for Knowledge Analysis - WEKA (University of Waikato, n.d.) é uma ferramenta de sofware utilizada na descoberta de conhecimento em sistemas de banco de dados. Possui uma série de algoritmos de preparação de dados, de aprendizagem de máquina (mineração) e de validação de resultados. Este software, desenvolvido na Universidade de Waikato na Nova Zelândia, foi implementado em Java (sendo, pois, portável) e se encontra disponível sob licença *GNU General Public License (GPL)*. A equipe de desenvolvedores desse software lança periodicamente correções e releases. A maioria dos seus componentes são provenientes de teses e dissertações de grupos de pesquisa desta universidade.

O sistema possui uma interface gráfica amigável (GUI) e seus algoritmos fornecem relatórios com dados analíticos e estatísticos do domínio minerado. Por meio de sua GUI é possível acessar a maioria de seus recursos. Por ser escrito em Java, conforme mencionado anteriormente, há uma boa portabilidade ao software. No entanto, em suas versões atuais o volume de dados a ser manipulado está limitado à dimensão de memória principal (Silva, 2004).

O Weka possui um formato próprio, o *ARFF*, o qual é constituído basicamente por duas partes:

(P1) contém a lista de todos os atributos, onde deve ser definido o tipo do atributo ou o seu domínio;

(P2) contém os registros (instâncias) a serem minerados.

Além da lista de atributos e das instâncias, é possível adicionar comentários.

É importante mencionar que se faz necessário converter os dados para o formato *ARFF* antes de submetê-los a qualquer algoritmo do pacote Weka.

Diversos classificadores para redes bayesianas estão implementados no Weka. NaiveBayes e BayesNet são alguns dos classificadores para esse tipo de rede disponíveis nesse software. Para utilizar um desses classificadores é necessário que sejam satisfeitas as seguintes condições:

(C1) todas as variáveis sejam finitas e discretas;

(C2) nenhuma instância possua valores faltantes.

Caso o conjunto de dados não satisfaça a essas duas condições, é necessário que se utilize filtros para tratar o conjunto de dados antes de submetê-lo. As classes `Discretize` e `ReplaceMissingValues` do pacote `weka.filters.unsupervised.attribute` são utilizadas como filtros para tornar as variáveis discretas e preencher os valores faltantes nas instâncias respectivamente.

A.3 BayesBuilder

O `BayesBuilder` (SNN Nijmegen, n.d.) é uma ferrramenta de software livre que permite a construção de redes bayesianas, bem como a inferência nesse tipo de modelo probabilístico. A versão atual dessa ferramenta contém algoritmos de inferências básicos assim como outros ambientes de desenvolvimento para redes bayesianas. Adicionalmente, foram implementados métodos eficientes de inferência para redes com nós *noisy-OR*, sem os quais a inferência seria, em termos computacionais, intratável (Foundation for Neural Networks–SNN and University Medical Centre Utrecht–UMCU, n.d.).

Este software, desenvolvido na Universidade de Nijmegen por um grupo de pesquisa denominado SNN Nijmegen, tem sua interface gráfica (GUI) escrita em Java, a qual é bastante intuitiva. O kernel que contém o motor de inferência, por sua vez, está escrito em C++, o que torna possível incorporar esse motor a outras aplicações.

Dentre os recursos providos pelo `BayesBuilder`, merecem destaque:

(R1) definição de várias visões para uma mesma rede bayesiana, o que é essencial para a construção de grandes redes;

(R2) uso de variáveis aleatórias gaussianas discretizadas e `portas OR` implementadas de forma eficiente; com isso é possível lidar com grandes volumes de dados;

(R3) importação de redes em diversos formatos, como, por exemplo, Hugin, Netica, Microsoft Bayesian Network e Bayesian Interchange;

(R4) exportação do estado da rede para um banco de dados de casos e importação de banco de dados de casos para a rede;

(R5) suporte à busca de nós.

Apêndice B

Cálculos das Probabilidades de cpg e cmg

Aqui são apresentados em detalhes os cálculos das probabilidades necessárias à construção das tabelas de probabilidade *a priori* (ver Tabelas 5.12 e 5.13 na página 62) das v.a. cpg e cmg apresentadas no Capítulo 5.

B.1 Cálculos das Probabilidades de cpg

B.1.1 $\Pr(\text{cpg} = \text{a})$

$$
\begin{aligned}
\Pr(\text{cpg}=\text{a}) \;=\; & \sum_{(\text{i}=\text{a},\text{b},\text{c})} \sum_{(\text{j}=\text{a},\text{b},\text{c})} \sum_{(\text{k}=\text{sim},\text{não})} [\, \Pr(\text{pppg}=\text{i}) \times \Pr(\text{ppmg}=\text{j}) \times \Pr(\text{pb}=\text{k}) \times \\
& \times \Pr(\text{pgen}=\text{i}-\text{j} \mid \text{pppg}=\text{i} \cap \text{ppmg}=\text{j}) \times \\
& \times \Pr(\text{cpg}=\text{a} \mid \text{pppg}=\text{i} \cap \text{ppmg}=\text{j} \cap \text{pb}=\text{k}) \,] \\
= \; & [\, \Pr(\text{pppg}=\text{a}) \times \Pr(\text{ppmg}=\text{a}) \times \Pr(\text{pb}=\text{sim}) \times \\
& \times \Pr(\text{pgen}=\text{a}-\text{a} \mid \text{pppg}=\text{a} \cap \text{ppmg}=\text{a}) \times \\
& \times \Pr(\text{cpg}=\text{a} \mid \text{pppg}=\text{a} \cap \text{ppmg}=\text{a} \cap \text{pb}=\text{sim}) \,] \; + \\
& [\, \Pr(\text{pppg}=\text{a}) \times \Pr(\text{ppmg}=\text{a}) \times \Pr(\text{pb}=\text{não}) \times \\
& \times \Pr(\text{pgen}=\text{a}-\text{a} \mid \text{pppg}=\text{a} \cap \text{ppmg}=\text{a}) \times \\
& \times \Pr(\text{cpg}=\text{a} \mid \text{pppg}=\text{a} \cap \text{ppmg}=\text{a} \cap \text{pb}=\text{não}) \,] \; + \\
& [\, \Pr(\text{pppg}=\text{a}) \times \Pr(\text{ppmg}=\text{b}) \times \Pr(\text{pb}=\text{sim}) \times \\
& \times \Pr(\text{pgen}=\text{a}-\text{b} \mid \text{pppg}=\text{a} \cap \text{ppmg}=\text{b}) \times \\
& \times \Pr(\text{cpg}=\text{a} \mid \text{pppg}=\text{a} \cap \text{ppmg}=\text{b} \cap \text{pb}=\text{sim}) \,] \; + \\
& [\, \Pr(\text{pppg}=\text{a}) \times \Pr(\text{ppmg}=\text{b}) \times \Pr(\text{pb}=\text{não}) \times \\
& \times \Pr(\text{pgen}=\text{a}-\text{b} \mid \text{pppg}=\text{a} \cap \text{ppmg}=\text{b}) \times \\
& \times \Pr(\text{cpg}=\text{a} \mid \text{pppg}=\text{a} \cap \text{ppmg}=\text{b} \cap \text{pb}=\text{não}) \,] \; +
\end{aligned}
$$

$$
\begin{aligned}
&[\ \Pr(\mathrm{pppg}=\mathrm{a})\times\Pr(\mathrm{ppmg}=\mathrm{c})\times\Pr(\mathrm{pb}=\mathrm{sim})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{a}-\mathrm{c}\mid\mathrm{pppg}=\mathrm{a}\cap\mathrm{ppmg}=\mathrm{c})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{a}\cap\mathrm{ppmg}=\mathrm{c}\cap\mathrm{pb}=\mathrm{sim})\ ]\ +\\
&[\ \Pr(\mathrm{pppg}=\mathrm{a})\times\Pr(\mathrm{ppmg}=\mathrm{c})\times\Pr(\mathrm{pb}=\mathrm{não})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{a}-\mathrm{c}\mid\mathrm{pppg}=\mathrm{a}\cap\mathrm{ppmg}=\mathrm{c})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{a}\cap\mathrm{ppmg}=\mathrm{c}\cap\mathrm{pb}=\mathrm{não})\ ]\ +\\
&[\ \Pr(\mathrm{pppg}=\mathrm{b})\times\Pr(\mathrm{ppmg}=\mathrm{a})\times\Pr(\mathrm{pb}=\mathrm{sim})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{b}-\mathrm{a}\mid\mathrm{pppg}=\mathrm{b}\cap\mathrm{ppmg}=\mathrm{a})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{b}\cap\mathrm{ppmg}=\mathrm{a}\cap\mathrm{pb}=\mathrm{sim})\ ]\ +\\
&[\ \Pr(\mathrm{pppg}=\mathrm{b})\times\Pr(\mathrm{ppmg}=\mathrm{a})\times\Pr(\mathrm{pb}=\mathrm{não})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{b}-\mathrm{a}\mid\mathrm{pppg}=\mathrm{b}\cap\mathrm{ppmg}=\mathrm{a})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{b}\cap\mathrm{ppmg}=\mathrm{a}\cap\mathrm{pb}=\mathrm{não})\ ]\ +\\
&[\ \Pr(\mathrm{pppg}=\mathrm{b})\times\Pr(\mathrm{ppmg}=\mathrm{b})\times\Pr(\mathrm{pb}=\mathrm{sim})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{b}-\mathrm{b}\mid\mathrm{pppg}=\mathrm{b}\cap\mathrm{ppmg}=\mathrm{b})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{b}\cap\mathrm{ppmg}=\mathrm{b}\cap\mathrm{pb}=\mathrm{sim})\ ]\ +\\
&[\ \Pr(\mathrm{pppg}=\mathrm{b})\times\Pr(\mathrm{ppmg}=\mathrm{b})\times\Pr(\mathrm{pb}=\mathrm{não})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{b}-\mathrm{b}\mid\mathrm{pppg}=\mathrm{b}\cap\mathrm{ppmg}=\mathrm{b})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{b}\cap\mathrm{ppmg}=\mathrm{b}\cap\mathrm{pb}=\mathrm{não})\ ]\ +\\
&[\ \Pr(\mathrm{pppg}=\mathrm{b})\times\Pr(\mathrm{ppmg}=\mathrm{c})\times\Pr(\mathrm{pb}=\mathrm{sim})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{b}-\mathrm{c}\mid\mathrm{pppg}=\mathrm{b}\cap\mathrm{ppmg}=\mathrm{c})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{b}\cap\mathrm{ppmg}=\mathrm{c}\cap\mathrm{pb}=\mathrm{sim})\ ]\ +\\
&[\ \Pr(\mathrm{pppg}=\mathrm{b})\times\Pr(\mathrm{ppmg}=\mathrm{c})\times\Pr(\mathrm{pb}=\mathrm{não})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{b}-\mathrm{c}\mid\mathrm{pppg}=\mathrm{b}\cap\mathrm{ppmg}=\mathrm{c})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{b}\cap\mathrm{ppmg}=\mathrm{c}\cap\mathrm{pb}=\mathrm{não})\ ]\ +\\
&[\ \Pr(\mathrm{pppg}=\mathrm{c})\times\Pr(\mathrm{ppmg}=\mathrm{a})\times\Pr(\mathrm{pb}=\mathrm{sim})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{c}-\mathrm{a}\mid\mathrm{pppg}=\mathrm{c}\cap\mathrm{ppmg}=\mathrm{a})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{c}\cap\mathrm{ppmg}=\mathrm{a}\cap\mathrm{pb}=\mathrm{sim})\ ]\ +\\
&[\ \Pr(\mathrm{pppg}=\mathrm{c})\times\Pr(\mathrm{ppmg}=\mathrm{a})\times\Pr(\mathrm{pb}=\mathrm{não})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{c}-\mathrm{a}\mid\mathrm{pppg}=\mathrm{c}\cap\mathrm{ppmg}=\mathrm{a})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{c}\cap\mathrm{ppmg}=\mathrm{a}\cap\mathrm{pb}=\mathrm{não})\ ]\ +\\
&[\ \Pr(\mathrm{pppg}=\mathrm{c})\times\Pr(\mathrm{ppmg}=\mathrm{b})\times\Pr(\mathrm{pb}=\mathrm{sim})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{c}-\mathrm{b}\mid\mathrm{pppg}=\mathrm{c}\cap\mathrm{ppmg}=\mathrm{b})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{c}\cap\mathrm{ppmg}=\mathrm{b}\cap\mathrm{pb}=\mathrm{sim})\ ]\ +\\
&[\ \Pr(\mathrm{pppg}=\mathrm{c})\times\Pr(\mathrm{ppmg}=\mathrm{b})\times\Pr(\mathrm{pb}=\mathrm{não})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{c}-\mathrm{b}\mid\mathrm{pppg}=\mathrm{c}\cap\mathrm{ppmg}=\mathrm{b})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{c}\cap\mathrm{ppmg}=\mathrm{b}\cap\mathrm{pb}=\mathrm{não})\ ]\ +\\
&[\ \Pr(\mathrm{pppg}=\mathrm{c})\times\Pr(\mathrm{ppmg}=\mathrm{c})\times\Pr(\mathrm{pb}=\mathrm{sim})\times\\
&\times\Pr(\mathrm{pgen}=\mathrm{c}-\mathrm{c}\mid\mathrm{pppg}=\mathrm{c}\cap\mathrm{ppmg}=\mathrm{c})\times\\
&\times\Pr(\mathrm{cpg}=\mathrm{a}\mid\mathrm{pppg}=\mathrm{c}\cap\mathrm{ppmg}=\mathrm{c}\cap\mathrm{pb}=\mathrm{sim})\ ]\ +
\end{aligned}
$$

$$
\begin{aligned}
& [\ \Pr(\text{pppg} = \text{c}) \times \Pr(\text{ppmg} = \text{c}) \times \Pr(\text{pb} = \text{não}) \times \\
& \times \Pr(\text{pgen} = \text{c} - \text{c} \mid \text{pppg} = \text{c} \cap \text{ppmg} = \text{c}) \times \\
& \times \Pr(\text{cpg} = \text{a} \mid \text{pppg} = \text{c} \cap \text{ppmg} = \text{c} \cap \text{pb} = \text{não})\ ] \\
= \ & [\ 0,122 \times 0,122 \times 0,5 \times 1 \times 1,0\ ] + [\ 0,122 \times 0,122 \times 0,5 \times 1 \times 0,122\ ] + \\
& [\ 0,122 \times 0,222 \times 0,5 \times 1 \times 0,5\ ] + [\ 0,122 \times 0,222 \times 0,5 \times 1 \times 0,122\ ] + \\
& [\ 0,122 \times 0,656 \times 0,5 \times 1 \times 0,5\ ] + [\ 0,122 \times 0,656 \times 0,5 \times 1 \times 0,122\ ] + \\
& [\ 0,222 \times 0,122 \times 0,5 \times 1 \times 0,5\ ] + [\ 0,222 \times 0,122 \times 0,5 \times 1 \times 0,122\ ] + \\
& [\ 0,222 \times 0,222 \times 0,5 \times 1 \times 0,0\ ] + [\ 0,222 \times 0,222 \times 0,5 \times 1 \times 0,122\ ] + \\
& [\ 0,222 \times 0,656 \times 0,5 \times 1 \times 0,0\ ] + [\ 0,222 \times 0,656 \times 0,5 \times 1 \times 0,122\ ] + \\
& [\ 0,656 \times 0,122 \times 0,5 \times 1 \times 0,5\ ] + [\ 0,656 \times 0,122 \times 0,5 \times 1 \times 0,122\ ] + \\
& [\ 0,656 \times 0,222 \times 0,5 \times 1 \times 0,0\ ] + [\ 0,656 \times 0,222 \times 0,5 \times 1 \times 0,122\ ] + \\
& [\ 0,656 \times 0,656 \times 0,5 \times 1 \times 0,0\ ] + [\ 0,656 \times 0,656 \times 0,5 \times 1 \times 0,122\ ] \\
= \ & 0,0074420000 + 0,0009079240 + 0,0067710000 + \\
& 0,0016521240 + 0,0200080000 + 0,0048819520 + \\
& 0,0067710000 + 0,0016521240 + 0,0000000000 + \\
& 0,0030063240 + 0,0000000000 + 0,0088835520 + \\
& 0,0200080000 + 0,0048819520 + 0,0000000000 + \\
& 0,0088835520 + 0,0000000000 + 0,0262504960 \\
= \ & 0,122
\end{aligned}
$$

B.1.2 $\Pr(\text{cpg} = \text{b})$

$$
\begin{aligned}
\Pr(\text{cpg} = \text{b}) = & \sum_{(\text{i}=\text{a},\text{b},\text{c})} \sum_{(\text{j}=\text{a},\text{b},\text{c})} \sum_{(\text{k}=\text{sim},\text{não})} [\ \Pr(\text{pppg} = \text{i}) \times \Pr(\text{ppmg} = \text{j}) \times \Pr(\text{pb} = \text{k}) \times \\
& \times \Pr(\text{pgen} = \text{i} - \text{j} \mid \text{pppg} = \text{i} \cap \text{ppmg} = \text{j}) \times \\
& \times \Pr(\text{cpg} = \text{b} \mid \text{pppg} = \text{i} \cap \text{ppmg} = \text{j} \cap \text{pb} = \text{k})\ ] \\
= \ & [\ \Pr(\text{pppg} = \text{a}) \times \Pr(\text{ppmg} = \text{a}) \times \Pr(\text{pb} = \text{sim}) \times \\
& \times \Pr(\text{pgen} = \text{a} - \text{a} \mid \text{pppg} = \text{a} \cap \text{ppmg} = \text{a}) \times \\
& \times \Pr(\text{cpg} = \text{b} \mid \text{pppg} = \text{a} \cap \text{ppmg} = \text{a} \cap \text{pb} = \text{sim})\ ] + \\
& [\ \Pr(\text{pppg} = \text{a}) \times \Pr(\text{ppmg} = \text{a}) \times \Pr(\text{pb} = \text{não}) \times \\
& \times \Pr(\text{pgen} = \text{a} - \text{a} \mid \text{pppg} = \text{a} \cap \text{ppmg} = \text{a}) \times \\
& \times \Pr(\text{cpg} = \text{b} \mid \text{pppg} = \text{a} \cap \text{ppmg} = \text{a} \cap \text{pb} = \text{não})\ ] + \\
& [\ \Pr(\text{pppg} = \text{a}) \times \Pr(\text{ppmg} = \text{b}) \times \Pr(\text{pb} = \text{sim}) \times \\
& \times \Pr(\text{pgen} = \text{a} - \text{b} \mid \text{pppg} = \text{a} \cap \text{ppmg} = \text{b}) \times \\
& \times \Pr(\text{cpg} = \text{b} \mid \text{pppg} = \text{a} \cap \text{ppmg} = \text{b} \cap \text{pb} = \text{sim})\ ] + \\
& [\ \Pr(\text{pppg} = \text{a}) \times \Pr(\text{ppmg} = \text{b}) \times \Pr(\text{pb} = \text{não}) \times \\
& \times \Pr(\text{pgen} = \text{a} - \text{b} \mid \text{pppg} = \text{a} \cap \text{ppmg} = \text{b}) \times \\
& \times \Pr(\text{cpg} = \text{b} \mid \text{pppg} = \text{a} \cap \text{ppmg} = \text{b} \cap \text{pb} = \text{não})\ ] +
\end{aligned}
$$

$$
\begin{aligned}
&[\ \Pr(pppg = a) \times \Pr(ppmg = c) \times \Pr(pb = sim) \times \\
&\times \Pr(pgen = a - c \mid pppg = a \cap ppmg = c) \times \\
&\times \Pr(cpg = b \mid pppg = a \cap ppmg = c \cap pb = sim)\ ] + \\
&[\ \Pr(pppg = a) \times \Pr(ppmg = c) \times \Pr(pb = não) \times \\
&\times \Pr(pgen = a - c \mid pppg = a \cap ppmg = c) \times \\
&\times \Pr(cpg = b \mid pppg = a \cap ppmg = c \cap pb = não)\ ] + \\
&[\ \Pr(pppg = b) \times \Pr(ppmg = a) \times \Pr(pb = sim) \times \\
&\times \Pr(pgen = b - a \mid pppg = b \cap ppmg = a) \times \\
&\times \Pr(cpg = b \mid pppg = b \cap ppmg = a \cap pb = sim)\ ] + \\
&[\ \Pr(pppg = b) \times \Pr(ppmg = a) \times \Pr(pb = não) \times \\
&\times \Pr(pgen = b - a \mid pppg = b \cap ppmg = a) \times \\
&\times \Pr(cpg = b \mid pppg = b \cap ppmg = a \cap pb = não)\ ] + \\
&[\ \Pr(pppg = b) \times \Pr(ppmg = b) \times \Pr(pb = sim) \times \\
&\times \Pr(pgen = b - b \mid pppg = b \cap ppmg = b) \times \\
&\times \Pr(cpg = b \mid pppg = b \cap ppmg = b \cap pb = sim)\ ] + \\
&[\ \Pr(pppg = b) \times \Pr(ppmg = b) \times \Pr(pb = não) \times \\
&\times \Pr(pgen = b - b \mid pppg = b \cap ppmg = b) \times \\
&\times \Pr(cpg = b \mid pppg = b \cap ppmg = b \cap pb = não)\ ] + \\
&[\ \Pr(pppg = b) \times \Pr(ppmg = c) \times \Pr(pb = sim) \times \\
&\times \Pr(pgen = b - c \mid pppg = b \cap ppmg = c) \times \\
&\times \Pr(cpg = b \mid pppg = b \cap ppmg = c \cap pb = sim)\ ] + \\
&[\ \Pr(pppg = b) \times \Pr(ppmg = c) \times \Pr(pb = não) \times \\
&\times \Pr(pgen = b - c \mid pppg = b \cap ppmg = c) \times \\
&\times \Pr(cpg = b \mid pppg = b \cap ppmg = c \cap pb = não)\ ] + \\
&[\ \Pr(pppg = c) \times \Pr(ppmg = a) \times \Pr(pb = sim) \times \\
&\times \Pr(pgen = c - a \mid pppg = c \cap ppmg = a) \times \\
&\times \Pr(cpg = b \mid pppg = c \cap ppmg = a \cap pb = sim)\ ] + \\
&[\ \Pr(pppg = c) \times \Pr(ppmg = a) \times \Pr(pb = não) \times \\
&\times \Pr(pgen = c - a \mid pppg = c \cap ppmg = a) \times \\
&\times \Pr(cpg = b \mid pppg = c \cap ppmg = a \cap pb = não)\ ] + \\
&[\ \Pr(pppg = c) \times \Pr(ppmg = b) \times \Pr(pb = sim) \times \\
&\times \Pr(pgen = c - b \mid pppg = c \cap ppmg = b) \times \\
&\times \Pr(cpg = b \mid pppg = c \cap ppmg = b \cap pb = sim)\ ] + \\
&[\ \Pr(pppg = c) \times \Pr(ppmg = b) \times \Pr(pb = não) \times \\
&\times \Pr(pgen = c - b \mid pppg = c \cap ppmg = b) \times \\
&\times \Pr(cpg = b \mid pppg = c \cap ppmg = b \cap pb = não)\ ] + \\
&[\ \Pr(pppg = c) \times \Pr(ppmg = c) \times \Pr(pb = sim) \times \\
&\times \Pr(pgen = c - c \mid pppg = c \cap ppmg = c) \times
\end{aligned}
$$

$$
\begin{aligned}
& \times \Pr(cpg = b \mid pppg = c \cap ppmg = c \cap pb = sim) \,] \,+ \\
& [\, \Pr(pppg = c) \times \Pr(ppmg = c) \times \Pr(pb = não) \times \\
& \times \Pr(pgen = c - c \mid pppg = c \cap ppmg = c) \times \\
& \times \Pr(cpg = b \mid pppg = c \cap ppmg = c \cap pb = não) \,] \\
= \; & [\, 0{,}122 \times 0{,}122 \times 0{,}5 \times 1 \times 0{,}0 \,] + [\, 0{,}122 \times 0{,}122 \times 0{,}5 \times 1 \times 0{,}222 \,] \,+ \\
& [\, 0{,}122 \times 0{,}222 \times 0{,}5 \times 1 \times 0{,}5 \,] + [\, 0{,}122 \times 0{,}222 \times 0{,}5 \times 1 \times 0{,}222 \,] \,+ \\
& [\, 0{,}122 \times 0{,}656 \times 0{,}5 \times 1 \times 0{,}0 \,] + [\, 0{,}122 \times 0{,}656 \times 0{,}5 \times 1 \times 0{,}222 \,] \,+ \\
& [\, 0{,}222 \times 0{,}122 \times 0{,}5 \times 1 \times 0{,}5 \,] + [\, 0{,}222 \times 0{,}122 \times 0{,}5 \times 1 \times 0{,}222 \,] \,+ \\
& [\, 0{,}222 \times 0{,}222 \times 0{,}5 \times 1 \times 1{,}0 \,] + [\, 0{,}222 \times 0{,}222 \times 0{,}5 \times 1 \times 0{,}222 \,] \,+ \\
& [\, 0{,}222 \times 0{,}656 \times 0{,}5 \times 1 \times 0{,}5 \,] + [\, 0{,}222 \times 0{,}656 \times 0{,}5 \times 1 \times 0{,}222 \,] \,+ \\
& [\, 0{,}656 \times 0{,}122 \times 0{,}5 \times 1 \times 0{,}0 \,] + [\, 0{,}656 \times 0{,}122 \times 0{,}5 \times 1 \times 0{,}222 \,] \,+ \\
& [\, 0{,}656 \times 0{,}222 \times 0{,}5 \times 1 \times 0{,}5 \,] + [\, 0{,}656 \times 0{,}222 \times 0{,}5 \times 1 \times 0{,}222 \,] \,+ \\
& [\, 0{,}656 \times 0{,}656 \times 0{,}5 \times 1 \times 0{,}0 \,] + [\, 0{,}656 \times 0{,}656 \times 0{,}5 \times 1 \times 0{,}222 \,] \\
= \; & 0{,}0000000000 + 0{,}0016521240 + 0{,}0067710000 + \\
& 0{,}0030063240 + 0{,}0000000000 + 0{,}0088835520 + \\
& 0{,}0067710000 + 0{,}0030063240 + 0{,}0246420000 + \\
& 0{,}0054705240 + 0{,}0364080000 + 0{,}0161651520 + \\
& 0{,}0000000000 + 0{,}0088835520 + 0{,}0364080000 + \\
& 0{,}0161651520 + 0{,}0000000000 + 0{,}0477672960 \\
= \; & 0{,}222
\end{aligned}
$$

B.1.3 $\Pr(cpg = c)$

$$
\begin{aligned}
\Pr(cpg = c) \; = \; & \sum_{(i=a,b,c)} \sum_{(j=a,b,c)} \sum_{(k=sim,não)} [\, \Pr(pppg = i) \times \Pr(ppmg = j) \times \Pr(pb = k) \times \\
& \times \Pr(pgen = i - j \mid pppg = i \cap ppmg = j) \times \\
& \times \Pr(cpg = c \mid pppg = i \cap ppmg = j \cap pb = k) \,] \\
= \; & [\, \Pr(pppg = a) \times \Pr(ppmg = a) \times \Pr(pb = sim) \times \\
& \times \Pr(pgen = a - a \mid pppg = a \cap ppmg = a) \times \\
& \times \Pr(cpg = c \mid pppg = a \cap ppmg = a \cap pb = sim) \,] \,+ \\
& [\, \Pr(pppg = a) \times \Pr(ppmg = a) \times \Pr(pb = não) \times \\
& \times \Pr(pgen = a - a \mid pppg = a \cap ppmg = a) \times \\
& \times \Pr(cpg = c \mid pppg = a \cap ppmg = a \cap pb = não) \,] \,+ \\
& [\, \Pr(pppg = a) \times \Pr(ppmg = b) \times \Pr(pb = sim) \times \\
& \times \Pr(pgen = a - b \mid pppg = a \cap ppmg = b) \times \\
& \times \Pr(cpg = c \mid pppg = a \cap ppmg = b \cap pb = sim) \,] \,+ \\
& [\, \Pr(pppg = a) \times \Pr(ppmg = b) \times \Pr(pb = não) \times \\
& \times \Pr(pgen = a - b \mid pppg = a \cap ppmg = b) \times
\end{aligned}
$$

$$\times \Pr(cpg = c \mid pppg = a \cap ppmg = b \cap pb = não) \;] \; +$$
$$[\Pr(pppg = a) \times \Pr(ppmg = c) \times \Pr(pb = sim) \times$$
$$\times \Pr(pgen = a - c \mid pppg = a \cap ppmg = c) \times$$
$$\times \Pr(cpg = c \mid pppg = a \cap ppmg = c \cap pb = sim) \;] \; +$$
$$[\Pr(pppg = a) \times \Pr(ppmg = c) \times \Pr(pb = não) \times$$
$$\times \Pr(pgen = a - c \mid pppg = a \cap ppmg = c) \times$$
$$\times \Pr(cpg = c \mid pppg = a \cap ppmg = c \cap pb = não) \;] \; +$$
$$[\Pr(pppg = b) \times \Pr(ppmg = a) \times \Pr(pb = sim) \times$$
$$\times \Pr(pgen = b - a \mid pppg = b \cap ppmg = a) \times$$
$$\times \Pr(cpg = c \mid pppg = b \cap ppmg = a \cap pb = sim) \;] \; +$$
$$[\Pr(pppg = b) \times \Pr(ppmg = a) \times \Pr(pb = não) \times$$
$$\times \Pr(pgen = b - a \mid pppg = b \cap ppmg = a) \times$$
$$\times \Pr(cpg = c \mid pppg = b \cap ppmg = a \cap pb = não) \;] \; +$$
$$[\Pr(pppg = b) \times \Pr(ppmg = b) \times \Pr(pb = sim) \times$$
$$\times \Pr(pgen = b - b \mid pppg = b \cap ppmg = b) \times$$
$$\times \Pr(cpg = c \mid pppg = b \cap ppmg = b \cap pb = sim) \;] \; +$$
$$[\Pr(pppg = b) \times \Pr(ppmg = b) \times \Pr(pb = não) \times$$
$$\times \Pr(pgen = b - b \mid pppg = b \cap ppmg = b) \times$$
$$\times \Pr(cpg = c \mid pppg = b \cap ppmg = b \cap pb = não) \;] \; +$$
$$[\Pr(pppg = b) \times \Pr(ppmg = c) \times \Pr(pb = sim) \times$$
$$\times \Pr(pgen = b - c \mid pppg = b \cap ppmg = c) \times$$
$$\times \Pr(cpg = c \mid pppg = b \cap ppmg = c \cap pb = sim) \;] \; +$$
$$[\Pr(pppg = b) \times \Pr(ppmg = c) \times \Pr(pb = não) \times$$
$$\times \Pr(pgen = b - c \mid pppg = b \cap ppmg = c) \times$$
$$\times \Pr(cpg = c \mid pppg = b \cap ppmg = c \cap pb = não) \;] \; +$$
$$[\Pr(pppg = c) \times \Pr(ppmg = a) \times \Pr(pb = sim) \times$$
$$\times \Pr(pgen = c - a \mid pppg = c \cap ppmg = a) \times$$
$$\times \Pr(cpg = c \mid pppg = c \cap ppmg = a \cap pb = sim) \;] \; +$$
$$[\Pr(pppg = c) \times \Pr(ppmg = a) \times \Pr(pb = não) \times$$
$$\times \Pr(pgen = c - a \mid pppg = c \cap ppmg = a) \times$$
$$\times \Pr(cpg = c \mid pppg = c \cap ppmg = a \cap pb = não) \;] \; +$$
$$[\Pr(pppg = c) \times \Pr(ppmg = b) \times \Pr(pb = sim) \times$$
$$\times \Pr(pgen = c - b \mid pppg = c \cap ppmg = b) \times$$
$$\times \Pr(cpg = c \mid pppg = c \cap ppmg = b \cap pb = sim) \;] \; +$$
$$[\Pr(pppg = c) \times \Pr(ppmg = b) \times \Pr(pb = não) \times$$
$$\times \Pr(pgen = c - b \mid pppg = c \cap ppmg = b) \times$$
$$\times \Pr(cpg = c \mid pppg = c \cap ppmg = b \cap pb = não) \;] \; +$$
$$[\Pr(pppg = c) \times \Pr(ppmg = c) \times \Pr(pb = sim) \times$$

$$
\begin{aligned}
& \times \Pr(pgen = c - c \mid pppg = c \cap ppmg = c) \times \\
& \times \Pr(cpg = c \mid pppg = c \cap ppmg = c \cap pb = sim) \,] \; + \\
& [\, \Pr(pppg = c) \times \Pr(ppmg = c) \times \Pr(pb = não) \times \\
& \times \Pr(pgen = c - c \mid pppg = c \cap ppmg = c) \times \\
& \times \Pr(cpg = c \mid pppg = c \cap ppmg = c \cap pb = não) \,] \\
= \; & [\, 0,122 \times 0,122 \times 0,5 \times 1 \times 0,0 \,] + [\, 0,122 \times 0,122 \times 0,5 \times 1 \times 0,656 \,] \; + \\
& [\, 0,122 \times 0,222 \times 0,5 \times 1 \times 0,0 \,] + [\, 0,122 \times 0,222 \times 0,5 \times 1 \times 0,656 \,] \; + \\
& [\, 0,122 \times 0,656 \times 0,5 \times 1 \times 0,5 \,] + [\, 0,122 \times 0,656 \times 0,5 \times 1 \times 0,656 \,] \; + \\
& [\, 0,222 \times 0,122 \times 0,5 \times 1 \times 0,0 \,] + [\, 0,222 \times 0,122 \times 0,5 \times 1 \times 0,656 \,] \; + \\
& [\, 0,222 \times 0,222 \times 0,5 \times 1 \times 0,0 \,] + [\, 0,222 \times 0,222 \times 0,5 \times 1 \times 0,656 \,] \; + \\
& [\, 0,222 \times 0,656 \times 0,5 \times 1 \times 0,5 \,] + [\, 0,222 \times 0,656 \times 0,5 \times 1 \times 0,656 \,] \; + \\
& [\, 0,656 \times 0,122 \times 0,5 \times 1 \times 0,5 \,] + [\, 0,656 \times 0,122 \times 0,5 \times 1 \times 0,656 \,] \; + \\
& [\, 0,656 \times 0,222 \times 0,5 \times 1 \times 0,5 \,] + [\, 0,656 \times 0,222 \times 0,5 \times 1 \times 0,656 \,] \; + \\
& [\, 0,656 \times 0,656 \times 0,5 \times 1 \times 1,0 \,] + [\, 0,656 \times 0,656 \times 0,5 \times 1 \times 0,656 \,] \\
= \; & 0,0000000000 + 0,0048819520 + 0,0000000000 + \\
& 0,0088835520 + 0,0200080000 + 0,0262504960 + \\
& 0,0000000000 + 0,0088835520 + 0,0000000000 + \\
& 0,0161651520 + 0,0364080000 + 0,0477672960 + \\
& 0,0200080000 + 0,0262504960 + 0,0364080000 + \\
& 0,0477672960 + 0,2151680000 + 0,1411502080 \\
= \; & 0,656
\end{aligned}
$$

B.2 Cálculos das Probabilidades de cmg

B.2.1 $\Pr(cmg = a)$

$$
\begin{aligned}
\Pr(cmg = a) \; = \; & \sum_{(i=a,b,c)} \sum_{(j=a,b,c)} [\, \Pr(mpg = i) \times \Pr(mmg = j) \times \\
& \times \Pr(mgen = i - j \mid mpg = i \cap mmg = j) \times \\
& \times \Pr(cmg = a \mid mpg = i \cap mmg = j \,] \\
= \; & [\, \Pr(mpg = a) \times \Pr(mmg = a) \times \\
& \times \Pr(mgen = a - a \mid mpg = a \cap mmg = a) \times \\
& \times \Pr(cmg = a \mid mpg = a \cap mmg = a) \,] \; + \\
& [\, \Pr(mpg = a) \times \Pr(mmg = b) \times \\
& \times \Pr(mgen = a - b \mid mpg = a \cap mmg = b) \times \\
& \times \Pr(cmg = a \mid mpg = a \cap mmg = b) \,] \; + \\
& [\, \Pr(mpg = a) \times \Pr(mmg = c) \times \\
& \times \Pr(mgen = a - c \mid mpg = a \cap mmg = c) \times
\end{aligned}
$$

$$
\begin{aligned}
&\times \Pr(cmg = a \mid mpg = a \cap mmg = c) \,] \,+ \\
&[\, \Pr(mpg = b) \times \Pr(mmg = a) \times \\
&\times \Pr(mgen = b - a \mid mpg = b \cap mmg = a) \times \\
&\times \Pr(cmg = a \mid mpg = b \cap mmg = a) \,] \,+ \\
&[\, \Pr(mpg = b) \times \Pr(mmg = b) \times \\
&\times \Pr(mgen = b - b \mid mpg = b \cap mmg = b) \times \\
&\times \Pr(cmg = a \mid mpg = b \cap mmg = b) \,] \,+ \\
&[\, \Pr(mpg = b) \times \Pr(mmg = c) \times \\
&\times \Pr(mgen = b - c \mid mpg = b \cap mmg = c) \times \\
&\times \Pr(cmg = a \mid mpg = b \cap mmg = c) \,] \,+ \\
&[\, \Pr(mpg = c) \times \Pr(mmg = a) \times \\
&\times \Pr(mgen = c - a \mid mpg = c \cap mmg = a) \times \\
&\times \Pr(cmg = a \mid mpg = c \cap mmg = a) \,] \,+ \\
&[\, \Pr(mpg = c) \times \Pr(mmg = b) \times \\
&\times \Pr(mgen = c - b \mid mpg = c \cap mmg = b) \times \\
&\times \Pr(cmg = a \mid mpg = c \cap mmg = b) \,] \,+ \\
&[\, \Pr(mpg = c) \times \Pr(mmg = c) \times \\
&\times \Pr(mgen = c - c \mid mpg = c \cap mmg = c) \times \\
&\times \Pr(cmg = a \mid mpg = c \cap mmg = c) \,] \\
= \;& [\, 0,122 \times 0,122 \times 1 \times 1,0 \,] + [\, 0,122 \times 0,222 \times 1 \times 0,5 \,] \,+ \\
&[\, 0,122 \times 0,656 \times 1 \times 0,5 \,] + [\, 0,222 \times 0,122 \times 1 \times 0,5 \,] \,+ \\
&[\, 0,222 \times 0,222 \times 1 \times 0,0 \,] + [\, 0,222 \times 0,656 \times 1 \times 0,0 \,] \,+ \\
&[\, 0,656 \times 0,122 \times 1 \times 0,5 \,] + [\, 0,656 \times 0,222 \times 1 \times 0,0 \,] \,+ \\
&[\, 0,656 \times 0,656 \times 1 \times 0,0 \,] \\
= \;& 0,0148840000 + 0,0135420000 + 0,0400160000 + \\
&0,0135420000 + 0,0 + 0,0 + 0,0400160000 + 0,0 + 0,0 \\
= \;& 0,122
\end{aligned}
$$

B.2.2 $\Pr(cmg = b)$

$$
\begin{aligned}
\Pr(cmg = b) \;=\; & \sum_{(i=a,b,c)} \sum_{(j=a,b,c)} [\, \Pr(mpg = i) \times \Pr(mmg = j) \times \\
& \times \Pr(mgen = i - j \mid mpg = i \cap mmg = j) \times \\
& \times \Pr(cmg = b \mid mpg = i \cap mmg = j \,] \\
=\; & [\, \Pr(mpg = a) \times \Pr(mmg = a) \times \\
& \times \Pr(mgen = a - a \mid mpg = a \cap mmg = a) \times \\
& \times \Pr(cmg = b \mid mpg = a \cap mmg = a) \,] \,+
\end{aligned}
$$

$$
\begin{aligned}
&[\ Pr(mpg = a) \times Pr(mmg = b) \times \\
&\times Pr(mgen = a - b \mid mpg = a \cap mmg = b) \times \\
&\times Pr(cmg = b \mid mpg = a \cap mmg = b)\ ] + \\
&[\ Pr(mpg = a) \times Pr(mmg = c) \times \\
&\times Pr(mgen = a - c \mid mpg = a \cap mmg = c) \times \\
&\times Pr(cmg = b \mid mpg = a \cap mmg = c)\ ] + \\
&[\ Pr(mpg = b) \times Pr(mmg = a) \times \\
&\times Pr(mgen = b - a \mid mpg = b \cap mmg = a) \times \\
&\times Pr(cmg = b \mid mpg = b \cap mmg = a)\ ] + \\
&[\ Pr(mpg = b) \times Pr(mmg = b) \times \\
&\times Pr(mgen = b - b \mid mpg = b \cap mmg = b) \times \\
&\times Pr(cmg = b \mid mpg = b \cap mmg = b)\ ] + \\
&[\ Pr(mpg = b) \times Pr(mmg = c) \times \\
&\times Pr(mgen = b - c \mid mpg = b \cap mmg = c) \times \\
&\times Pr(cmg = b \mid mpg = b \cap mmg = c)\ ] + \\
&[\ Pr(mpg = c) \times Pr(mmg = a) \times \\
&\times Pr(mgen = c - a \mid mpg = c \cap mmg = a) \times \\
&\times Pr(cmg = b \mid mpg = c \cap mmg = a)\ ] + \\
&[\ Pr(mpg = c) \times Pr(mmg = b) \times \\
&\times Pr(mgen = c - b \mid mpg = c \cap mmg = b) \times \\
&\times Pr(cmg = b \mid mpg = c \cap mmg = b)\ ] + \\
&[\ Pr(mpg = c) \times Pr(mmg = c) \times \\
&\times Pr(mgen = c - c \mid mpg = c \cap mmg = c) \times \\
&\times Pr(cmg = b \mid mpg = c \cap mmg = c)\ ] \\
= \ &[\ 0,122 \times 0,122 \times 1 \times 0,0\ ] + [\ 0,122 \times 0,222 \times 1 \times 0,5\ ] + \\
&[\ 0,122 \times 0,656 \times 1 \times 0,0\ ] + [\ 0,222 \times 0,122 \times 1 \times 0,5\ ] + \\
&[\ 0,222 \times 0,222 \times 1 \times 1,0\ ] + [\ 0,222 \times 0,656 \times 1 \times 0,5\ ] + \\
&[\ 0,656 \times 0,122 \times 1 \times 0,0\ ] + [\ 0,656 \times 0,222 \times 1 \times 0,5\ ] + \\
&[\ 0,656 \times 0,656 \times 1 \times 0,0\ ] \\
= \ &0,0 + 0,0135420000 + 0,0 + 0,0135420000 + 0,0492840000 + \\
&0,0728160000 + 0,0 + 0,0728160000 + 0,0 \\
= \ &0,222
\end{aligned}
$$

B.2.3 $Pr(cmg = c)$

$$
\begin{aligned}
Pr(cmg = c) \ = \ &\sum_{(i=a,b,c)} \sum_{(j=a,b,c)} [\ Pr(mpg = i) \times Pr(mmg = j) \times \\
&\times Pr(mgen = i - j \mid mpg = i \cap mmg = j) \times
\end{aligned}
$$

$$
\begin{aligned}
& \times \Pr(cmg = c \mid mpg = i \cap mmg = j \,] \\
= \; & [\, \Pr(mpg = a) \times \Pr(mmg = a) \times \\
& \times \Pr(mgen = a - a \mid mpg = a \cap mmg = a) \times \\
& \times \Pr(cmg = c \mid mpg = a \cap mmg = a) \,] \; + \\
& [\, \Pr(mpg = a) \times \Pr(mmg = b) \times \\
& \times \Pr(mgen = a - b \mid mpg = a \cap mmg = b) \times \\
& \times \Pr(cmg = c \mid mpg = a \cap mmg = b) \,] \; + \\
& [\, \Pr(mpg = a) \times \Pr(mmg = c) \times \\
& \times \Pr(mgen = a - c \mid mpg = a \cap mmg = c) \times \\
& \times \Pr(cmg = c \mid mpg = a \cap mmg = c) \,] \; + \\
& [\, \Pr(mpg = b) \times \Pr(mmg = a) \times \\
& \times \Pr(mgen = b - a \mid mpg = b \cap mmg = a) \times \\
& \times \Pr(cmg = c \mid mpg = b \cap mmg = a) \,] \; + \\
& [\, \Pr(mpg = b) \times \Pr(mmg = b) \times \\
& \times \Pr(mgen = b - b \mid mpg = b \cap mmg = b) \times \\
& \times \Pr(cmg = c \mid mpg = b \cap mmg = b) \,] \; + \\
& [\, \Pr(mpg = b) \times \Pr(mmg = c) \times \\
& \times \Pr(mgen = b - c \mid mpg = b \cap mmg = c) \times \\
& \times \Pr(cmg = c \mid mpg = b \cap mmg = c) \,] \; + \\
& [\, \Pr(mpg = c) \times \Pr(mmg = a) \times \\
& \times \Pr(mgen = c - a \mid mpg = c \cap mmg = a) \times \\
& \times \Pr(cmg = c \mid mpg = c \cap mmg = a) \,] \; + \\
& [\, \Pr(mpg = c) \times \Pr(mmg = b) \times \\
& \times \Pr(mgen = c - b \mid mpg = c \cap mmg = b) \times \\
& \times \Pr(cmg = c \mid mpg = c \cap mmg = b) \,] \; + \\
& [\, \Pr(mpg = c) \times \Pr(mmg = c) \times \\
& \times \Pr(mgen = c - c \mid mpg = c \cap mmg = c) \times \\
& \times \Pr(cmg = c \mid mpg = c \cap mmg = c) \,] \\
= \; & [\, 0,122 \times 0,122 \times 1 \times 0,0 \,] + [\, 0,122 \times 0,222 \times 1 \times 0,0 \,] \; + \\
& [\, 0,122 \times 0,656 \times 1 \times 0,5 \,] + [\, 0,222 \times 0,122 \times 1 \times 0,0 \,] \; + \\
& [\, 0,222 \times 0,222 \times 1 \times 0,0 \,] + [\, 0,222 \times 0,656 \times 1 \times 0,5 \,] \; + \\
& [\, 0,656 \times 0,122 \times 1 \times 0,5 \,] + [\, 0,656 \times 0,222 \times 1 \times 0,5 \,] \; + \\
& [\, 0,656 \times 0,656 \times 1 \times 1,0 \,] \\
= \; & 0,0 + 0,0 + 0,0400160000 + 0,0 + 0,0 + 0,0728160000 \\
& 0,0400160000 + 0,0728160000 + 0,4303360000 \\
= \; & 0,656
\end{aligned}
$$

Apêndice C

Aquivos em Formato CSV

Aqui são apresentados os arquivos csv usados no Capítulo 6 (seção 6.5).

C.1 CSV da Tabela de Freqüências Alélicas de Alagoas

```
Alelo;FGA;D16S539;D7S820;D13S317;D3S1358;Penta E;D18S51;D12S391;TH01;D19S433
;CSF1PO;TPOX;F13B;LPL;vWA;D10S2325;D1S1656;D21S11;D8S1179;D2S1338;D5S818;Penta D
;D14S299;D6S477;D8S1179;SE33
2.2;;;;;;;;;;;;;;;;;;;;;;0.011200;;;;
3.2;;;;;;;;;;;;;;;;;;;;;;0.011200;;;;
5;;;;;;0.068480;;;0.001000;;;;;;;;;;;;0.011200;;;;
6;;0.002000;0.001000;;;0.005040;;;0.223000;;0.001000;0.016000;0.164000;;;0.003000
;;;;;0.002000;0.011200;;;;
7;;;0.011000;0.003000;;0.106750;;;0.234000;;0.011000;0.004000;0.066000;;;0.136000
;;;;;0.015020;0.011200;;;;
8;;0.017000;0.138000;0.084000;;0.056390;;;0.147000;;0.012000;0.473000;0.164000
;0.005000;;0.055000;;;0.005010;;0.015020;0.022410;;;0.005800;
9;;0.151000;0.125000;0.088000;;0.023160;;;0.169000;0.001914;0.025000;0.104000
;0.255000;0.054000;;0.113000;0.007000;;0.003010;;0.026030;0.173670;;;0.011600;
9.3;;;;;;;;;0.203000;;;;;;;;;;;;;;;;
10;;0.098000;0.268000;0.058000;;0.056390;0.007000;;0.023000;0.005733;0.282000
;0.053000;0.328000;0.383000;;0.142000;0.032000;;0.052100;;0.052050;0.173670;;
;0.063900;11;;0.312000;0.249000;0.290000;;0.090630;0.016000;;;0.025819;0.274000
;0.296000;0.023000;0.280000;0.004000;0.116000;0.099000;;0.065130;;0.328330
;0.164330;;;0.087200;0.005000
11.2;;;;;;;;;0.047619;;;;;;;;;;;;;;;0.005000
12;;0.246000;0.177000;0.308000;0.006000;0.164150;0.122000;;;0.064067;0.332000
;0.050000;;0.229000;;0.176000;0.063000;;0.130260;;0.370370;0.146590;;;0.122100;
```

```
12.2;;;;;;;;;;0.011476;;;;;;;;;;;;;;;;
13;;0.144000;0.028000;0.131000;;0.152060;0.095000;0.002000;;0.223752;0.055000
;0.004000;;0.049000;0.001000;0.124000;0.173000;;0.290580;;0.178180;0.167130;;
;0.244200;
13.2;;;;;;;;;;0.033467;;;;;;;;;;;;0.011200;;;;
13.3;;;;;;;;;;;;;;;;;;0.011000;;;;;;;;
14;;0.027000;0.003000;0.038000;0.093000;0.085600;0.148000;;;0.273476;0.008000;;;
;0.059000;0.082000;0.134000;;0.256510;;0.013010;0.051350;;;0.279100;0.015000
14.2;;;;;;;;;;0.049724;;;;;;;;;;;;;;;;
14.3;;;;;;;;;;;;;;;;;;0.049000;;;;;;;;
15;;0.001000;;;0.298000;0.051360;0.152000;0.033000;;0.143429;;;;;0.164000
;0.047000;0.136000;;0.158320;;;0.011200;;0.005800;0.162800;0.030000
15.2;;;;;;;;;;0.047810;;;;;;;;;;;;;;;;
15.3;;;;;;;;;;;;;;;;;;0.053000;;;;;;;;
16;;;;;0.286000;0.048340;0.145000;0.039000;;0.043981;;;;;0.278000;0.003000
;0.077000;;0.031060;0.041120;;0.011200;;0.151200;;0.040000
16.2;;;;;;;;;;0.027733;;;;;;;;;;;;;;;;
16.3;;;;;;;;;;;;;;;;;;0.099000;;;;;;;;
17;;0.001000;;;0.181000;0.035250;0.141000;0.091000;;;;;;;0.256000;0.003000
;0.007000;;0.008020;0.226680;;0.011200;;0.157000;;0.140000
17.3;;;;;;;;;0.013000;;;;;;;;;0.049000;;;;;;;;
18;0.022000;0.001000;;;0.124000;0.023160;0.082000;0.223000;;;;;;;0.169000;
;0.004000;;;0.058170;;;;0.174400;;0.155000
18.3;;;;;;;;;0.013000;;;;;;;;;0.007000;;;;;;;;
19;0.089000;;;;0.012000;0.011080;0.056000;0.154000;;;;;;;0.053000;;;;
;0.118360;;;;0.325600;0.023300;0.080000
19.3;;;;;;;;;0.006000;;;;;;;;;;;;;;;;;;
20;0.089000;;;;;0.010070;0.013000;0.138000;;;;;;;0.013000;;;;;0.136410;;;
;0.168600;;0.060000
20.2;;;;;;;;;;;;;;;;;;;;;;;;;;;;0.005000
21;0.178000;;;;;0.008060;0.010000;0.086000;;;;;;;0.002000;;;;;0.062190;;;
;0.011600;;0.045000
21.2;;;;;;;;;;;;;;;;;;;;;;;;;;;;0.010000
22;0.144000;;;;;0.003020;0.010000;0.106000;;;;;;;0.001000;;;;;0.095290;;;;;
;0.005000
22.2;;;;;;;;;;;;;;;;;;;;;;;;;;;;0.025000
23;0.122000;;;;;0.001010;;0.059000;;;;;;;;;;;;0.099300;;;;0.005800;;
23.2;;;;;;;;;;;;;;;;;;;;;;;;;;;;0.025000
24;0.178000;;;;;;0.003000;0.024000;;;;;;;;;;0.003010;;0.081240;;;;;;
24.2;;;;;;;;;;;;;;;;;;;;;;;;;;;;0.030000
25;0.128000;;;;;;;0.011000;;;;;;;;;;0.003010;;0.060180;;;;;;
```

```
25.2;;;;;;;;;;;;;;;;;;;;;;;;;;;;0.010000
26;0.028000;;;;;;;0.002000;;;;;;;;;;;;0.019060;;;;;;
26.2;;;;;;;;;;;;;;;;;;;;;;;;;;;;0.070000
27;0.017000;;;;;;;;;;;;;;;;;;;0.032100;;0.002010;;;;;;
27.2;;;;;;;;;;;;;;;;;;;;;;;;;;;;0.060000
28;0.005000;;;;;;;;;;;;;;;;;;;0.161480;;;;;;;;
28.2;;;;;;;;;;;;;;;;;;;;;;;;;;;;0.045000
29;;;;;;;;;;;;;;;;;;0.200600;;;;;;;;
29.2;;;;;;;;;;;;;;;;;;;;;;;;;;;;0.080000
30;;;;;;;;;;;;;;;;;;0.210630;;;;;0.069790;;;
30.2;;;;;;;;;;;;;;;;;;0.019060;;;;;;;;0.025000
31;;;;;;;;;;;;;;;;;;0.081240;;;;;0.191880;;;
31.2;;;;;;;;;;;;;;;;;;0.110330;;;;;;;;0.010000
32;;;;;;;;;;;;;;;;;;0.013040;;;;;0.279070;;;
32.2;;;;;;;;;;;;;;;;;;0.110330;;;;;;;;0.010000
33;;;;;;;;;;;;;;;;;;0.010030;;;;;0.215080;;;0.005000
33.2;;;;;;;;;;;;;;;;;;0.023070;;;;;;;;0.005000
34;;;;;;;;;;;;;;;;;;0.006020;;;;;0.087190;;;0.005000
35;;;;;;;;;;;;;;;;;;0.013040;;;;;0.058090;;;
35.2;;;;;;;;;;;;;;;;;;0.003010;;;;;;;;
36;;;;;;;;;;;;;;;;;;;;;;0.075590;;;
37;;;;;;;;;;;;;;;;;;;;;;0.023300;;;
```

C.2 CSV do Estudo de Paternidade Caso Padrão

```
Sample Name, Marker, Allele 1, Allele 2, Allele 3, Allele 4, Allele 5, Allele 6
, Allele 7, Allele 8, Allele 9, Allele 10, UD2,
224-A-M, Amel, X, X, , , , , , , , , 1,
224-A-C, Amel, X, X, , , , , , , , , 2,
224-A-SP, Amel, X, Y, , , , , , , , , 3,
224-A-M, CSF1PO, 11, 13, , , , , , , , , 1,
224-A-C, CSF1PO, 10, 11, , , , , , , , , 2,
224-A-SP, CSF1PO, 10, 12, , , , , , , , , 3,
224-A-M, D10S2325, 10, 12, , , , , , , , , 1,
224-A-C, D10S2325, 10, 13, , , , , , , , , 2,
224-A-SP, D10S2325, 7, 13, , , , , , , , , 3,
224-A-M, D12S391, 18, 21, , , , , , , , , 1,
224-A-C, D12S391, 18, 19, , , , , , , , , 2,
224-A-SP, D12S391, 19, 23, , , , , , , , , 3,
224-A-M, D13S317, 11, 12, , , , , , , , , 1,
224-A-C, D13S317, 8, 11, , , , , , , , , 2,
```

```
224-A-SP, D13S317, 8, 11, , , , , , , , , 3,
224-A-M, D16S539, 8, 10, , , , , , , , , 1,
224-A-C, D16S539, 9, 10, , , , , , , , , 2,
224-A-SP, D16S539, 9, 11, , , , , , , , , 3,
224-A-M, D18S51, 19, 21, , , , , , , , , 1,
224-A-C, D18S51, 12, 19, , , , , , , , , 2,
224-A-SP, D18S51, 12, 14, , , , , , , , , 3,
224-A-M, D19S433, 13, 13, , , , , , , , , 1,
224-A-C, D19S433, 15, 15, , , , , , , , , 2,
224-A-SP, D19S433, 14, 15, , , , , , , , , 3,
224-A-M, D3S1358, 15, 15, , , , , , , , , 1,
224-A-C, D3S1358, 15, 16, , , , , , , , , 2,
224-A-SP, D3S1358, 15, 16, , , , , , , , , 3,
224-A-M, D7S820, 8, 10, , , , , , , , , 1,
224-A-C, D7S820, 8, 10, , , , , , , , , 2,
224-A-SP, D7S820, 9, 10, , , , , , , , , 3,
224-A-M, F13B, 9, 10, , , , , , , , , 1,
224-A-C, F13B, 9, 9, , , , , , , , , 2,
224-A-SP, F13B, 9, 9, , , , , , , , , 3,
224-A-M, FGA, 21, 22, , , , , , , , , 1,
224-A-C, FGA, 22, 23, , , , , , , , , 2,
224-A-SP, FGA, 22, 23, , , , , , , , , 3,
224-A-M, LPL, 10, 11, , , , , , , , , 1,
224-A-C, LPL, 10, 11, , , , , , , , , 2,
224-A-SP, LPL, 11, 11, , , , , , , , , 3,
224-A-M, Penta E, 5, 13, , , , , , , , , 1,
224-A-C, Penta E, 13, 19, , , , , , , , , 2,
224-A-SP, Penta E, 5, 19, , , , , , , , , 3,
224-A-M, TH01, 7, 9, , , , , , , , , 1,
224-A-C, TH01, 9, 9, , , , , , , , , 2,
224-A-SP, TH01, 8, 9, , , , , , , , , 3,
224-A-M, TPOX, 8, 11, , , , , , , , , 1,
224-A-C, TPOX, 8, 12, , , , , , , , , 2,
224-A-SP, TPOX, 11, 12, , , , , , , , , 3,
224-A-M, vWA, 17, 18, , , , , , , , , 1,
224-A-C, vWA, 18, 19, , , , , , , , , 2,
224-A-SP, vWA, 16, 19, , , , , , , , , 3,
```

C.3 CSV do Estudo de Paternidade Caso Complexo

```
Sample Name, Marker, Allele 1, Allele 2, Allele 3, Allele 4, Allele 5, Allele 6
, Allele 7, Allele 8, Allele 9, Allele 10, UD2,
224-A-M, Amel, X, X, , , , , , , , , 1,
224-A-C, Amel, X, X, , , , , , , , , 2,
224-A-PSP, Amel, X, Y, , , , , , , , , 3,
224-A-MSP, Amel, X, X, , , , , , , , , 4,
224-A-M, CSF1PO, 8, 12, , , , , , , , , 1,
224-A-C, CSF1PO, 11, 12, , , , , , , , , 2,
224-A-PSP, CSF1PO, 11, 12, , , , , , , , , 3,
224-A-MSP, CSF1PO, 12, 12, , , , , , , , , 4,
224-A-M, D10S2325, 10, 13, , , , , , , , , 1,
224-A-C, D10S2325, 9, 10, , , , , , , , , 2,
224-A-PSP, D10S2325, 9, 14, , , , , , , , , 3,
224-A-MSP, D10S2325, 11, 12, , , , , , , , , 4,
224-A-M, D12S391, 19, 22, , , , , , , , , 1,
224-A-C, D12S391, 15, 19, , , , , , , , , 2,
224-A-PSP, D12S391, 15, 19, , , , , , , , , 3,
224-A-MSP, D12S391, 17, 20, , , , , , , , , 4,
224-A-M, D13S317, 10, 11, , , , , , , , , 1,
224-A-C, D13S317, 10, 11, , , , , , , , , 2,
224-A-PSP, D13S317, 10, 13, , , , , , , , , 3,
224-A-MSP, D13S317, 8, 9, , , , , , , , , 4,
224-A-M, D16S539, 9, 9, , , , , , , , , 1,
224-A-C, D16S539, 9, 9, , , , , , , , , 2,
224-A-PSP, D16S539, 9, 12, , , , , , , , , 3,
224-A-MSP, D16S539, 9, 12, , , , , , , , , 4,
224-A-M, D18S51, 14, 16, , , , , , , , , 1,
224-A-C, D18S51, 12, 16, , , , , , , , , 2,
224-A-PSP, D18S51, 12, 15, , , , , , , , , 3,
224-A-MSP, D18S51, 14, 15, , , , , , , , , 4,
224-A-M, D19S433, 12, 14, , , , , , , , , 1,
224-A-C, D19S433, 12, 16, , , , , , , , , 2,
224-A-PSP, D19S433, 15,2, 16, , , , , , , , , 3,
224-A-MSP, D19S433, 15, 16, , , , , , , , , 4,
224-A-M, D2S1338, 20, 22, , , , , , , , , 1,
224-A-C, D2S1338, 22, 23, , , , , , , , , 2,
224-A-PSP, D2S1338, 17, 23, , , , , , , , , 3,
224-A-MSP, D2S1338, 19, 20, , , , , , , , , 4,
```

```
224-A-M, D3S1358, 15, 17, , , , , , , , , 1,
224-A-C, D3S1358, 15, 17, , , , , , , , , 2,
224-A-PSP, D3S1358, 14, 17, , , , , , , , , 3,
224-A-MSP, D3S1358, 15, 18, , , , , , , , , 4,
224-A-M, D5S818, 13, 13, , , , , , , , , 1,
224-A-C, D5S818, 10, 13, , , , , , , , , 2,
224-A-PSP, D5S818, 10, 10, , , , , , , , , 3,
224-A-MSP, D5S818, 10, 12, , , , , , , , , 4,
224-A-M, D7S820, 11, 12, , , , , , , , , 1,
224-A-C, D7S820, 11, 12, , , , , , , , , 2,
224-A-PSP, D7S820, 9, 11, , , , , , , , , 3,
224-A-MSP, D7S820, 9, 11, , , , , , , , , 4,
224-A-M, F13B, 9, 10, , , , , , , , , 1,
224-A-C, F13B, 6, 10, , , , , , , , , 2,
224-A-PSP, F13B, 10, 11, , , , , , , , , 3,
224-A-MSP, F13B, 6, 10, , , , , , , , , 4,
224-A-M, FGA, 22, 25, , , , , , , , , 1,
224-A-C, FGA, 23, 25, , , , , , , , , 2,
224-A-PSP, FGA, 23, 24, , , , , , , , , 3,
224-A-MSP, FGA, 20, 21, , , , , , , , , 4,
224-A-M, Penta E, 5, 11, , , , , , , , , 1,
224-A-C, Penta E, 10, 11, , , , , , , , , 2,
224-A-PSP, Penta E, 7, 10, , , , , , , , , 3,
224-A-MSP, Penta E, 12, 19, , , , , , , , , 4,
224-A-M, SE33, 19, 19, , , , , , , , , 1,
224-A-C, SE33, 19, 21,2, , , , , , , , , 2,
224-A-PSP, SE33, 24,2, 25,2, , , , , , , , , 3,
224-A-MSP, SE33, 21,2, 27,2, , , , , , , , , 4,
224-A-M, TH01, 8, 9, , , , , , , , , 1,
224-A-C, TH01, 8, 9, , , , , , , , , 2,
224-A-PSP, TH01, 7, 8, , , , , , , , , 3,
224-A-MSP, TH01, 7, 8, , , , , , , , , 4,
224-A-M, TPOX, 8, 8, , , , , , , , , 1,
224-A-C, TPOX, 8, 8, , , , , , , , , 2,
224-A-PSP, TPOX, 10, 10, , , , , , , , , 3,
224-A-MSP, TPOX, 8, 12, , , , , , , , , 4,
224-A-M, vWA, 16, 16, , , , , , , , , 1,
224-A-C, vWA, 14, 16, , , , , , , , , 2,
224-A-PSP, vWA, 15, 17, , , , , , , , , 3,
224-A-MSP, vWA, 14, 19, , , , , , , , , 4,
```

Referências Bibliográficas

Adriaenssens, V., Goethals, P. L. M., Charles, J. & De Pauw, N. (2004), 'Application of Bayesian Belief Networks for the prediction of macroinvertebrate taxa in rivers', *Annales de Limnologie-International Journal of Limnology* **40**(3), 181–191.

Agostinelli, C. & Rotondi, R. (2003), 'Using Bayesian belief networks to analyse the stochastic dependence between interevent time and size of earthquakes', *Journal of Seismology* **7**(3), 281–299.

Almudevar, A. (2007), 'A graphical approach to relatedness inference', *Theoretical Population Biology* **71**(2), 213–229.

Alonso, A., Martin, P., Albarran, C., Garcia, P., de Simon, L. F., Iturralde, M. J., Fernandez-Rodriguez, A., Atienza, I., Capilla, J., Garcia-Hirschfeld, J., Martinez, P., Vallejo, G., Garcia, O., Garcia, E., Real, P., Alvarez, D., Leon, A. & Sancho, M. (2005), 'Challenges of DNA profiling in mass disaster investigations', *Croatian Medical Journal* **46**(4), 540–548.

Antal, P., Fannes, G., Timmerman, D., Moreau, Y. & De Moor, B. (2003), 'Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection', *Artificial Intelligence in Medicine* **29**(1-2), 39–60.

Bai, C. G. (2005), 'Bayesian network based software reliability prediction with an operational profile', *Journal of Systems and Software* **77**(2), 103–112.

Baldi, P. & Brunak, S. (2001), *Bioinformatics: The Machine Learning Approach*, 2 ed., The MIT Press.

Bandyopadhyay, S., Maulik, U. & Roy, D. (2008), 'Gene identification: Classical and computational intelligence approaches', *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews* **38**(1), 55–68.

Bianchi, L. & Lio, P. (2007), 'Forensic DNA and bioinformatics', *Briefings in Bioinformatics* **8**(2), 117–128.

Booch, G., Rumbaugh, J. & Jacobson, I. (2005), *Unified Modeling Language User Guide*, 2 ed., Addison Wesley.

Bruegge, B. & Dutoit, A. H. (2003), *Object-Oriented Software Engineering: Using UML, Patterns and Java*, 2 ed., Prentice Hall.

Butler, J. M. (2005), *Forensic DNA Typing, Second Edition: Biology, Technology, and Genetics of STR Markers*, 2 ed., Academic Press.

Calabrez, M. (1999), Influência do calor na análise de DNA extraído de sangue e tecidos humanos: importância para a identificação de corpos carbonizados, PhD thesis, Universidade de São Paulo - USP.

Cheng, J., Greiner, R., Kelly, J., Bell, D. & Liu, W. R. (2002), 'Learning Bayesian networks from data: An information-theory based approach', *Artificial Intelligence* **137**(1-2), 43–90.

Costa, J. T. C. (2007), 'Redes Bayesianas: Fundamentação Teórica e Aplicações'. Trabalho de Conclusão de Curso, Instituto de Computação, Universidade Federal de Alagoas – Brasil.

Cowell, R. G. (2003), 'FINEX: a probabilistic expert system for forensic identification', *Forensic Science International* **134**(2-3), 196–206.

da Silva Júnior, C. & Sasson, S. (2002), *Biologia*, Vol. 1, 7 ed., Editora Saraiva. Disponível em `http://biologiacesaresezar.editorasaraiva.com.br/navitacontent_/userFiles/File/Biologia_Cesar_Sezar/BIO1_250.jpg` em 16/01/2009.

Dawid, A. P., Mortera, J., Pascali, V. L. & Van Boxel, D. (2002), 'Probabilistic expert systems for forensic inference from genetic markers', *Scandinavian Journal of Statistics* **29**(4), 577–595.

Egeland, T., Mostad, P. F., Mevag, B. & Stenersen, M. (2000), 'Beyond traditional paternity and identification cases Selecting the most probable pedigree', *Forensic Science International* **110**(1), 47–59.

Elmasri, R. & Navathe, S. B. (2003), *The Fudamentals of Database Systems*, 4 ed., Addison Wesley.

Flores, C. D., Höher, C. L., Ladeira, M. & Vicari, R. M. (2000), 'Uma Experiência do Uso de Redes Probabilísticas no Diagnóstico Médico', *Argentine Symposiun on Healthcare Informatics* pp. 126–128.

Foundation for Neural Networks–SNN and University Medical Centre Utrecht–UMCU (n.d.), '"PROMEDAS" a probabilistic decision support system for medical diagnosis'. URL http://www.promedas.nl/doc/TR_Promedas2002.pdf, última consulta em dezembro de 2006.

Gamma, E., Helm, R., Johnson, R. & Vlissides, J. (2005), *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison Wesley.

Goodwin, W., Linacre, A. & Hadi, S. (2007), *An Introduction to Forensic Genetics*, John Wiley & Sons Ltda.

Guedes, G. T. A. (2008), *UML - Uma Abordagem Prática*, 3 ed., Editora Novatec.

Henry, J. B. (2008), *Diagnósticos Clínicos e Tratamento por Métodos Laboratoriais*, 20 ed., Manole.

James, B. (1981), *Probabilidade: um Curso em Nível Intermediário*, Projeto Euclides, Instituto de Matemática Pura e Aplicada, Rio de Janeiro.

Jeffreys, A. J., Brookfield, J. F. Y. & Semenoff, R. (1985), 'Positive Identification of an Immigration Test-Case using Human DNA Fingerprints', *Nature* **317**(6040), 818–819.

Koten, C. V. & Gray, A. R. (2006), 'An application of Bayesian network for predicting object-oriented software maintainability', *Information and Software Technology* **48**(1), 59–67.

Leclair, B., Fregeau, C. J., Bowen, K. L. & Fourney, R. M. (2004), 'Enhanced kinship analysis and STR-based DNA typing for human identification in mass fatality incidents: The Swissair Flight 111 disaster', *Journal of Forensic Sciences* **49**(5), 939–953.

Leclair, B., Shaler, R., Carmody, G. R., Eliason, K., Hendrickson, B. C., Judkins, T., Norton, M. J., Sears, C. & Scholl, T. (2007), 'Bioinformatics and human identification in mass fatality incidents: The world trade center disaster', *Journal of Forensic Sciences* **52**(4), 806–819.

Lucke, H. (1995), 'Bayesian Belief Networks as a Tool for Stochastic Parsing', *Speech Communication* **16**(1), 89–118.

Luger, G. F. (2004), *Inteligência Artificial: Estruturas e Estratégias para a Resolução de Problemas Complexos*, 4 ed., Bookmann. Traduzido por P. Engel.

Magalhães, M. N. & de Lima, A. C. P. (2002), *Noções de Probabilidade e Estatística*, 2 ed., Editora da Universidade de São Paulo.

Marques, R. L. & Dutra, I. (n.d.), 'Redes Bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações'. URL http://www.cos.ufrj.br/~ines/courses/cos740/leila/cos740/Bayesianas.pdf, última consulta em maio de 2006.

Pearl, J. (1988), *Probabilistic Reasoning in Ingelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Francisco.

Pena, S. D. (2006), 'Thomas Bayes: 'o cara'!', *Ciência Hoje* **38**(228), 22–29.

Pike, W. A. (2004), 'Modeling drinking water quality violations with Bayesian networks', *Journal of the American Water Resources Association* **40**(6), 1563–1578.

Russell, S. J. & Norvig, P. (2004), *Inteligência Artificial*, 2 ed., Campus. Traduzido por V. D. de Souza.

Santos Júnior, J. R., Costa, J. T. C. & Almeida, E. S. (2008), Utilização de Redes Bayesianas na Verificação de Vínculo Genético, *in* 'Encontro Regional de Matemática Aplicada e Computacional', Vol. 8.

Santos Júnior, J. R., Costa, J. T. C. & Almeida, E. S. (2009), Um Sistema de Informação Modelado com Redes Bayesianas para Auxílio na Resolução de Testes de Paternidade, *in* 'Simpósio Brasileiro de Sistemas de Informação', Vol. 5.

Setubal, C. & Meidanis, J. (1997), *Introduction to Computational Molecular Biology*, 1 ed., PWS Publishing.

Silva, M. P. S. (2004), 'Mineração de Dados: Conceitos, Aplicações e Experimentos com Weka', *Livro da Escola Regional de Informática Rio de Janeiro - Espírito Santo (ERI RJ/ES)*.

Smarra, A., Paradela, E. & Figueiredo, A. (2006), 'A Genética Forense no Brasil', *Scientific American Brasil* pp. 51–83. URL http://ceticismo.net/2008/02/26/a-genetica-forense-no-brasil/, última consulta em janeiro de 2009.

SNN Nijmegen (n.d.), 'BayesBuilder (SNN Nijmegen)'. URL http://www.snn.ru.nl/nijmegen/index.php?option=com_content&view=article&id=89&Itemid=212, última consulta em janeiro de 2009.

Somerville, I. (2003), *Engenharia de Software*, 6 ed., Addison Wesley.

Universidade de Brasília (n.d.), 'UnBBayes'. URL http://sourceforge.net/projects/unbbayes, última consulta em janeiro de 2009.

University of Waikato (n.d.), 'Weka 3: Data Mining Software in Java'. URL http://www.cs.waikato.ac.nz/ml/weka, última consulta em janeiro de 2009.

Vieira, B. L., Costa, J. T. C. & Frery, A. C. (2006), Análise do Processo de Desenvolvimento de Software através de Redes Probabilísticas, *in* 'Encontro de Modelagem Computacional (EMC)', Vol. 9.

FAPEAL
Fundação de Amparo à Pesquisa do Estado de Alagoas

ALAGOAS
GOVERNO DO ESTADO

DATA: 22.07.2014

INTERESSADOS: JOSÉ TENÓRIO CÉSAR COSTA/ELIANA SILVA DE ALMEIDA

PROCESS: 20040429671-7

MODALIDADE: BOLSA DE MESTRADO - EAP

Assunto: Apresentar à FAPEAL o Exemplar da Dissertação de 15.03.2009.

Prezado(a) Senhor(a),

A FAPEAL solicita de V. Sa. a apresentação do Exemplar da Dissertação acima referenciado, conforme Termo de Outorga, assinado em 02 de abril de 2008.

Comunicamos-lhe que, lamentavelmente, o não atendimento a essa solicitação no prazo de 15 (quinze) dias, contados a partir da data de expedição deste documento, ficará o Bolsista e Orientador na condição de inadimplente, inviabilizando a concessão de futuros auxílios que venham a ser apresentados a esta Fundação.

Lembramos, ainda, ser indispensável, ao encaminhar-nos qualquer documento, fazer constar o nome completo de V. Sª. e o número do processo correspondente.

Sendo o que se apresenta no momento, subscrevemo-nos,

Atenciosamente,

Mônica Melo Gomes do Nascimento
Diretora da UGC&T

Rua Melo Moraes,354 - Centro _Maceió - AL
CEP:57020-330 -Fone: 0XX82-3315-2200 – Fax : 0XX82-3315-4098 – *e-mail:* fap@fapeal.br

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa Multidisciplinar de Pós-Graduação em Modelagem Computacional de Conhecimento
Avenida Lourival Melo Mota, Km 14, Bloco 09, Cidade Universitária
57.072-900 Maceió AL Brasil CGC: 24.464.109/0001-48
Telefone: (082) 3214-1364

MODELAGEM COMPUTACIONAL DE CONHECIMENTO

Ata da defesa de dissertação do aluno
José Tenório César Costa

Realizou-se no dia 13 de abril de dois mil e nove, a partir das 14h, na sala de aula do Mestrado em Modelagem Computacional de Conhecimento da Universidade Federal de Alagoas, a defesa de dissertação de Mestrado em Modelagem Computacional de Conhecimento, intitulada "THÊMIS: Um sistema para análise forense de DNA utilizando Redes Bayesianas", apresentada por José Tenório César Costa, graduado em Ciência da Computação, como requisito parcial para a obtenção do grau de Mestre em Modelagem Computacional de Conhecimento, perante a seguinte comissão examinadora:

Professora e orientadora Eliana Silva de Almeida
Instituto de Computação – UFAL

Professor e co-orientador Alejandro César Frery
Instituto de Computação – UFAL

Professor Luiz Antônio Ferreira da Silva
Instituto de Ciências Biológicas e da Saúde – UFAL

Professor Maurício Marengoni
Faculdade de Computação e Informática – MACKENZIE

A Comissão indicou melhorias a serem realizadas num prazo máximo de sessenta (60) dias. Isto feito, sob a responsabilidade da professora e orientadora Eliana Silva de Almeida, a dissertação será considerada aprovada. Finalizados os trabalhos, às 16h, lavrou-se a presente ata, que vai assinada por mim e pelos membros da Comissão.

Maceió, 13 de abril de 2009.

José Vitor de Menezes Torres
Secretário

PROGRAMA DE PÓS GRADUAÇÃO - MODELAGEM COMPUTACIONAL DE CONHECIMENTO - Universidade Federal de Alagoas

Confere com o original. 14/05/09

Eliana Silva de Almeida
Doutora em Informática – Pontifícia Universidade Católica do Rio de Janeiro/Brasil

Alejandro César Frery
Doutor em Computação Aplicada – Instituto Nacional de Pesquisas Espaciais/Brasil

Luiz Antônio Ferreira da Silva
Doutor em Ciências Biológicas (Genética) – Universidade Federal do Rio de Janeiro/Brasil

Maurício Marengoni
Doutor em Ciência da Computação – University of Massachusetts at Amherst/Estados Unidos