Um Sistema de Informação Modelado com Redes Bayesianas para Auxílio na Resolução de Testes de Paternidade

João Roberto dos S. Junior , José Tenório Cesar*, Eliana Silva de Almeida

[1] Instituto de Computação
Universidade Federal de Alagoas – Maceió, AL – Brasil

{jroberto.comp, tenoriocesar, eliana.almeida}@gmail.com
* Bolsista da Fundação de Amparo à Pesquisa do Estado de Alagoas (FAPEAL).

Abstract. *The high request of paternity testing made by forensic genetic laboratories has been estimulated the use of informations systems for support on resolutions of these tests. System of this kind need high rate of reliability, since treat of parentage indentification and, therefore, a formal and efficient approach for the studies of genetic linkage among individuals has a great value. In this paper, is proposed a tool that uses the formalism of the Bayesian Networks to support on the resolution of paternity testing. Since the complexity of this test type, which analise the causality relation among paternal genes and the child genes, and the need of realization of accurate statistical calculus, the use of the Bayesian Networks on modeling grants not only the reliabily of the results, but also the system dependability within the society.*

Resumo. *A grande demanda de testes de paternidade feitos por laboratórios de genética forense tem estimulado o uso de sistemas de informações para auxílio na resolução desses testes. Sistemas deste porte requerem alto grau de confiança, já que tratam de identificação de parentesco e, portanto, uma abordagem formal e eficiente para o estudo de vínculo genético entre indivíduos é de grande valia. Neste trabalho, é proposta uma ferramenta que utiliza o formalismo das Redes Bayesianas para auxiliar na resolução de testes de paternidade. Considerando a complexidade deste tipo de testes, onde se analisa a relação de causalidade entre os genes paternos e o gene da criança e a necessidade da realização de cálculos estatísticos precisos, o uso de Redes Bayesianas na sua modelagem garante não apenas a confiabilidade dos resultados, mas também a credibilidade do sistema perante a sociedade.*

1 Introdução

Atualmente, os laboratórios de genética forense realizam intensamente testes de paternidade. De todas as análises feitas pelo Laboratório de Genética Forense da UFAL (`http://www.labdnaforense.br`), 90% delas são testes de paternidade. Basicamente, a resolução destes testes deve confirmar se um suposto pai, mãe e criança compartilham ou não as mesmas características genéticas. Resultados deste tipo, quando comprovados, influenciam em casos de determinação de pensão alimentícia e reclamantes de heranças, por exemplo, que abalam fortemente o ambiente familiar envolvido.

Em um exame de paternidade alguns passos são realizados. Inicialmente, deve ser feita a coleta dos dados necessários para a análise a partir das amostras de DNA de

cada indivíduo envolvido (suposto pai, mãe e criança). Se esses dados são coletados e interpretados corretamente, o estudo do DNA provê um potencial informativo muito grande no teste de paternidade [Morling et al. 2002].

Em seguida, a análise do material genético é feita em regiões específicas (*loci*) das amostras de DNA do suposto pai e da criança que são comparadas entre si. Ou seja, para cada região é relizada a comparação entre as duas amostras e, caso ocorra a igualdade (*matching*), então a criança pode ter herdado um dos dois alelos do suposto pai naquela região. Para inferir se há ou não o vínculo genético deve ser levado em conta a frequência que cada alelo possui em uma dada população, frequência esta chamada de frequência alélica. Logo, para que seja possível provar a existência de vínculo genético, além de realizar as comparações entre as regiões das amostras, deve-se tomar como base as frequências alélicas, que são adotadas como conhecimento *a priori* para o estudo. Além disso, considera-se ainda as configurações possíveis para os alelos (heterozigótico ou homozigótico). Todo este processo requer confiabilidade e precisão para ser a base de um sistema crítico desta natureza.

Existem diversas abordagens estocásticas para o tratamento dessas situações incertas. Uma delas, trabalha por meio de cálculos algébricos. Ela é eficiente, porém, tediosa, uma vez que é necessário aplicar uma fórmula específica para cada configuração possível dos alelos (heterozigótico ou homozigótico) de todos os indivíduos envolvidos no teste (para mais detalhes sobre a referida abordagem ver [Butler 2005]).

Em [Dawid et al. 2002] é apresentada uma abordagem alternativa que utiliza redes Bayesianas para o tratamento de incerteza na resolução de casos de parentesco. Uma grande vantagem é que, uma vez montada a rede, basta apenas identificar quais os alelos encontrados nos indivíduos durante a análise e, a própria rede, por fim, fornece o resultado (chamado de Índice de Paternidade - IP) para cada região analisada do DNA.

De fato, o uso de métodos formais durante o estudo de paternidade garante a sua confiabilidade. Algumas vezes, o resultado final é levado como prova para tribunais. Qualquer eventual equívoco pode implicar a tomada de decisões drásticas por parte da corte, uma vez que trata-se de julgamento de pessoas que podem ser condenadas ou inocentadas erroneamente.

O presente trabalho apresenta a modelagem de um sistema de informação para auxílio na resolução de teste de paternidade considerando-se o caso padrão, onde somente tem-se os materiais genéticos do suposto pai, mãe e criança, dando ênfase na utilização das redes Bayesianas como método estocástico para o tratamento de incerteza. O sistema aqui proposto vem sendo testado com êxito, utilizando dados reais do Laboratório de DNA Forense da UFAL.

Logo a seguir, na seção 2, será dada uma breve explanação acerca do formalismo de redes Bayesianas. A metodologia envolvida nos testes de paternidade é abordada na seção 3, bem como os cálculos necessários para resolver os testes. Na seção 4, a modelagem do sistema é apresentada, assim como as ferramentas utilizadas no seu desenvolvimento. Por fim, na seção 5, alguns comentários sobre o trabalho serão feitos.

2 Redes Bayesianas

Uma rede bayesiana pode ser vista como um grafo direcionado 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. Esse tipo de rede pode ser especificado como segue:

1. Os nós da rede representam variáveis aleatórias (booleanas, discretas, contínuas, mistas ou singulares) que estão inseridas dentro de um domínio de interesse.
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). E, caso P não tenha pais, P é dito ser uma raiz.
3. Um nó raiz possui uma tabela contendo probabilidades incondicionais (ou probabilidades *a priori*) denotadas por $Pr(A)$.
4. Cada nó X_i, $1 \leq i \leq número\ de\ nós$, tem uma probabilidade condicional, dada por $Pr(X_i|Pais(X_i))$, que quantifica o efeito dos pais sobre o nó filho.
5. A distribuição da variável aleatória X_i, dados todos os nós que a precedem, só depende dos seus pais.

A relação entre pais e filhos é de causalidade, diferentemente de outros modelos estatísticos como, por exemplo, o de regressão [Fenton et al. 2002], cuja utilização não é interessante em domínios com rica estrutura causal. A Figura 1 mostra uma rede bayesiana com apenas três nós, onde as variáveis E_1 e E_2 são causas de um problema e C, um efeito dessas causas.

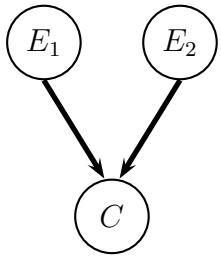

Figura 1. $Pais(C) = \{E_1, E_2\}$

Para obter conclusões acerca de uma variável pertencente à rede, é necessário que sejam consideradas evidências. As evidências são observações relativas a algum evento dentro do domínio de interesse. Diante disso, para um conjunto de eventos $A = (A_1, \ldots, A_n)$, é possível calcular o comportamento de qualquer variável em A através do teorema de Bayes:

$$Pr(A_k|B) = \frac{Pr(B|A_k) \times Pr(A_k)}{\sum_{i=1}^{n} [Pr(B|A_i) \times Pr(A_i)]}, para\ 1 \leq k \leq n$$

onde:

- A_k representa a variável de interesse que deseja-se saber sua probabilidade, considerando-se algum estado.
- B é um conjunto de evidências observadas no domínio em estudo.

No contexto das redes Bayesianas, para o mesmo conjunto A também é possível calcular a probabilidade de ocorrência mútua de uma série de eventos através da regra da cadeia [Taroni et al. 2004]:

$$Pr(A_1, \ldots, A_n) = \prod_{i}^{n} Pr(A_i|Pais(A_i))$$

Para a construção de uma rede bayesiana, é importante lembrar que ela não representa fluxo de informação, mas serve como uma representação direta de parte de um mundo real [Jensen 2001].

3 Teste de Paternidade

Em um teste de paternidade, o DNA de cada indivíduo envolvido é analisado por meio de métodos e ferramentas da biologia molecular. A grosso modo, a tarefa envolve comparações entre os materiais genéticos dos indivíduos. É razoável pensar que a longa cadeia de DNA é analisada inteiramente. Porém, este método não é utilizado por laboratórios, pois não provê vantagens em relação ao tempo (mais que 3,2 bilhões comparações devem ser realizadas [Goodwin et al. 2007]). Por outro lado, há um método alternativo em que apenas algumas regiões da cadeia (denominadas *lócus*) são selecionadas e analisadas para simplificar o trabalho.

Em cada lócu analisado, são encontrados dois alelos sendo um proveniente do pai e o outro, da mãe (embora não seja possível identificar a origem de cada um). Se os alelos encontrados em um lócu são diferentes, então o indivíduo é dito ser heterozigótico para esse lócu. Caso contrário, então ele é homozigótico. Nem todo o DNA é composto por regiões que contêm pares de alelos. Para que seja possível trabalhar com essas regiões alélicas, os laboratórios de genética forense utilizam marcadores genéticos. Os marcadores são enzimas que têm a capacidade de identificar lócus específicos na cadeia de DNA. A Figura 2 mostra um possível configuração alélica para um marcador X, onde a criança herda do pai o alelo 11 e da mãe, o alelo 12.

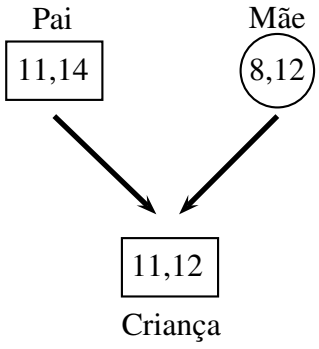

Figura 2. Configurações de alelos em um lócu comum para um marcador X

Grande parte dos laboratórios de genética forense utiliza 17 marcadores nas análises, a fim de se obter, justamente, 17 pares de alelos de cada indivíduo envolvido no teste (na realidade, este número pode variar entre cada laboratório). Um fato importante é que quanto mais marcadores são usados, mais preciso fica o resultado.

Cada alelo possui um número de ocorrências dentro de uma população de indivíduos, considerando-se um conjunto de marcadores genéticos. Esta frequência, denominada frequência alélica, é única para cada população e é a base de informações para os cálculos estatísticos feitos no estudo de vínculo genético. A Tabela 1 exibe parte da frequência alélica da população do estado de Alagoas.

3.1 Modelagem de redes bayesianas no estudo de paternidade

Na modelagem bayesiana do estudo de paternidade é necessário considerar o pai e a mãe

Marcador	Alelo	Frequência
D5S818	7	0.015
	8	0.015
	9	0.026
TH01	5	0.001
	6	0.223
⋮	⋮	⋮

Tabela 1. Frequência alélica de uma população de indivíduos

como indivíduos que não possuem vínculo genético entre si. Além disso, também deve-se levar em conta os eventos $PB = \texttt{sim}$, correspondente à hipótese do alegado pai ser o pai biológico da criança, e $PB = \texttt{não}$, significando que existe um outro homem na população que é o pai biológico da criança. Os dois eventos têm probabilidades *a priori* de 0.5, ou seja, $Pr(PB = \texttt{sim}) = Pr(PB = \texttt{não}) = 0.5$. O objetivo do estudo é determinar a probabilidade *a posteriori* de $PB = \texttt{sim}$ e $PB = \texttt{não}$, dados os genótipos do trio de indivíduos.

Sejam os seguintes eventos correspondentes aos alelos de cada indivíduo:

- ***pppg***: alelo paterno do suposto pai;
- ***ppmg***: alelo materno do suposto pai;
- ***mpg***: alelo paterno da mãe;
- ***mmg***: alelo materno da mãe;
- ***cpg***: alelo paterno da criança;
- ***cmg***: alelo materno da criança.

Os eventos *pppg*, *ppmg*, *mpg* e *mmg* são independentes, pois se referem aos alelos dos progenitores. Uma vez que o pai deve ser condiderado como um indivíduo selecionado ao acaso, a probabilidade *a priori* desses eventos é igual à frequência de ocorrência do alelo na população. Por exemplo, na frequência alélica da Tabela 1, para o marcador TH01, $Pr(pppg = 5) = 0.001$ e $Pr(ppmg = 6) = 0.223$.

O genótipo da criança é condicionado pelos genótipos dos seus pais e por ***PB***, portanto, deve-se fazer uma análise mais detalhada e ver a probabilidade de transmissão do alelo para a criança. As condições de transmissão são as mesmas para as variáveis *cpg* e *cmg*, exceto pelo fato da primeira ser condicionada por `PB`. Se o progenitor é homozigoto (A_1,A_1), a probabilidade de transmissão do alelo (A_1) é 1, já que não há outra possibilidade. Se o progenitor é heterozigoto (A_1,A_2), a probabilidade de transmissão (A_1 ou A_2) é 0.5. Quando assume-se $PB = \texttt{não}$ em `cpg`, a probabilidade de transmissão do alelo é igual à sua frequência na população. Para os alelos A_1 e A_2, com frequências, respectivamente, F_1 e F_2, obtém-se a tabela de probabilidades condicionais da variável *cpg*:

pppg	A_1				A_2			
ppmg	A_1		A_2		A_1		A_2	
PB	`sim`	`não`	`sim`	`não`	`sim`	`não`	`sim`	`não`
A_1	1	F_1	0.5	F_1	0.5	F_1	0	F_1
A_2	0	F_2	0.5	F_2	0.5	F_2	1	F_2

Tabela 2. Tabela de probabilidades da variável *cpg*

A estrutura da tabela de probabilidades da variável *cmg* difere um pouco da tabela de *cpg*, pois *mpg* não é condicionada por *PB*.

Com base nas informações expostas sobre as probabilidades de transmissão de alelos, já é possível montar a rede bayesiana que modela o problema do estudo do caso padrão de paternidade. A rede é apresentada a seguir:

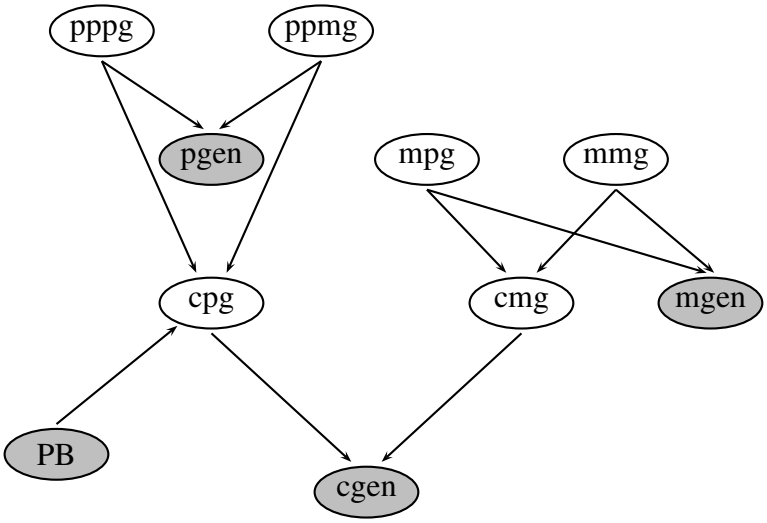

Figura 3. Rede bayesiana para o estudo de paternidade padrão

No grafo apresentado acima, o nó *cgen* tem o propósito de manter a rede conexa. Bem como os nós *pgen* e *mgen*, os seus estados são apenas a concatenação dos dois alelos contidos nos nós pais. Para os alelos A_1 e A_2, tem-se a seguinte tabela de probabilidades condicionais para a variável *cgen*:

cpg	A_1		A_2	
cmg	A_1	A_2	A_1	A_2
$A_1 - A_1$	1	0	0	0
$A_1 - A_2$	0	1	0	0
$A_2 - A_1$	0	0	1	0
$A_2 - A_2$	0	0	0	1

Tabela 3. Tabela de probabilidades da variável *cgen*

Com a rede montada e os genótipos dos três indivíduos conhecidos, o próximo passo limita-se em calcular as probabilidades *a posteriori* da variável *PB* para cada lócu. Na abordagem adotada há a necessidade de descobrir qual o alelo paterno obrigatório do filho e, através do par de alelos da mãe, isso torna-se possível [Lee et al. 2001]. Por exemplo, se o genótipo do suposto pai é $A_1 - A_2$, da mãe é $A_1 - A_3$ e do filho é $A_1 - A_3$, então o alelo paterno A_1 do filho é obrigatório, por que a mãe transmitiu certamente A_3. Portanto, os alelos maternos não estão relacionados nesse cálculo.

A probabilidade de PB dados o genótipo do suposto pai $[A_1, A_2]$ e o alelo paterno A_1 da criança (ou a probabilidade *a posteriori* de PB) é dada por:

$$Pr(PB|pppg = A_1, ppmg = A_2, cpg = A_1) =$$

$$= \frac{Pr(pppg{=}A_1, ppmg{=}A_2, cpg{=}A_1, PB)}{\sum_{pb} Pr(pppg{=}A_1, ppmg{=}A_2, cpg{=}A_1, PB{=}pb)},\ para\ pb = \texttt{sim}, \texttt{não} \quad (1)$$

Na divisão acima, o numerador e o denominador são um conjunção de eventos, portanto, é possível calculá-los através da regra da cadeia (seção 2):

$$\frac{Pr(pppg{=}A_1, ppmg{=}A_2, cpg{=}A_1, PB)}{\sum_{pb} Pr(pppg{=}A_1, ppmg{=}A_2, cpg{=}A_1, PB{=}pb)} =$$
$$= \frac{Pr(pppg{=}A_1) \times Pr(ppmg{=}A_2) \times Pr(cpg{=}A_1 | pppg{=}A_1, ppmg{=}A_2, PB) \times Pr(PB)}{\sum_{pb} [Pr(pppg{=}A_1) \times Pr(ppmg{=}A_2) \times Pr(cpg{=}A_1 | pppg{=}A_1, ppmg{=}A_2, PB) \times Pr(PB)]} \quad (2)$$

O fator que mensura se o suposto pai é o pai biológico da criança para um marcador específico é o Índice de Paternidade (IP). Para cada marcador usado, há um IP específico associado. Este índice é dado pelo quociente:

$$IP = \frac{Pr(PB = \texttt{sim} | pppg = A_1, ppmg = A_2, cpg = A_1)}{Pr(PB = \texttt{não} | pppg = A_1, ppmg = A_2, cpg = A_1)}$$

A informação definitiva na determinação do resultado da análise, o IPC (Índice de Paternidade Combinado), é o produto dos IP's calculados para cada marcador. O padrão mínimo aceito para o IPC é igual ou superior à 100 [Evett and Weir 1998]. Quando $IPC = 100$, significa que o suposto pai tem chance de 99 para 1 de ser o pai biológico da criança.

Como exemplo, sejam as probabilidades 0.206 e 0.222 para os alelos A_1 e A_2, respectivamente, para um determinado marcador. Utilizando a equação 2:

$$Pr(PB = \texttt{sim} | pppg = A_1, ppmg = A_2, cpg = A_1) =$$
$$= \frac{0.206 \times 0.222 \times 0.5 \times 0.5}{(0.206 \times 0.222 \times 0.5 \times 0.5) + (0.206 \times 0.222 \times 0.206 \times 0.5)} = \frac{0.011433}{0.016143396} = 0.708215297$$

Para obter o valor de $PB = \texttt{não}$, basta apenas fazer:

$$Pr(PB = \texttt{não} | pppg = A_1, ppmg = A_2, cpg = A_1) =$$
$$= 1 - Pr(PB = \texttt{sim} | pppg = A_1, ppmg = A_2, cpg = A_1) = 0.291784703$$

Uma vez que tem-se as probabilidades *a posteriori* de PB, pode-se calcular o IP para o mesmo marcador escolhido:

$$IP = \frac{0.708215297}{0.291784703} = 2.427184461$$

Para obter o resultado definitivo (IPC), basta apenas calcular e multiplicar todos os outros IP's para os respectivos marcadores escolhidos na análise.

4 Modelo do sistema

Um sistema de informação pode ser definido tecnicamente como um conjunto de componentes inter-relacionados que coleta (ou recupera), processa, armazena e distribui informações destinadas a apoiar a tomada de decisões, a coordenação e o controle de uma organização [Laudon and Laudon 2004].

Sistemas de informação que tratam questões judiciais requerem alto grau de confiança. A utilização de métodos formais e automatização da entrada dos dados genéticos dos indivíduos são tarefas cruciais para a conclusão do teste. O modelo do sistema aqui apresentado foi baseado na análise dos requisitos fundamentados nos processos realizados no Laboratório de Genética Forense da UFAL.

Na Figura 4, observa-se o modelo do sistema de informação com características voltadas para o problema de gerenciamento dos dados inerentes aos perfis genéticos de DNA. A entrada envolve a captação das amostras de DNA para o estudo de paternidade. Essas amostras são processadas com técnicas biotecnológicas, nos quais são gerados perfis genéticos, que são analisados através de modelos estatísticos. A saída envolve a transferência dos resultados comparativos dos perfis genéticos para seus respectivos interessados.

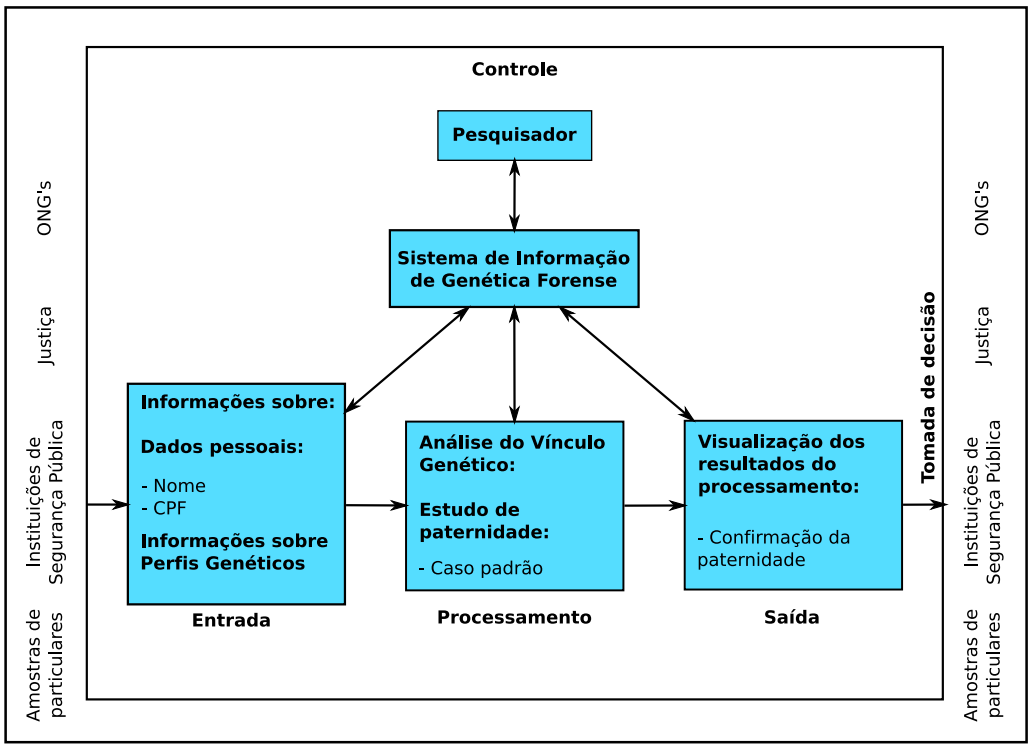

Figura 4. Visão do ambiente organizacional de um sistema de informação voltado ao estudo de paternidade padrão

Os dados de entrada referentes ao perfis genéticos dos indivíduos são gerados por uma máquina de PCR (*Polimerase Chain Reaction*). O método de PCR permite criar milhares de cópias de amostras de DNA. Após a aplicação desse método, a máquina de PCR armazena os perfis genéticos dos indivíduos envolvidos em um arquivo (em formato `csv`). Na Tabela 4, é exibida uma saída da máquina de PCR para a análise dos DNAs da mãe (M), suposto pai (SP) e criança (C) para o marcador CSF1PO.

4.1 Tecnologias usadas

O principal objetivo do presente trabalho é a modelagem e implementação de um ambiente para estudo do DNA em práticas forenses. A maior tarefa se restringe ao entendimento de como é realizado os estudos sobre vínculo genético e como esse estudo poderia

Sample Name	Marker	Allele 1	Allele 2
224-A-M	CSF1PO	10	11
224-A-SP	CSF1PO	10	12
224-A-C	CSF1PO	11	12

Tabela 4. Saída gerada pela máquina de PCR

ser automatizado. Para tanto, a escolha das seguintes tecnologias foi extremamente criteriosa.

Padrões de projetos: Os padrões de projeto (*Design Patterns*) fornecem uma metodologia para reutilização de softwares. Estes padrões tornam mais fácil o reuso de projetos e arquiteturas bem sucedidas. Os seguintes padrões foram utilizados no ambiente: Façade, Singleton e Data Acess Objects – DAOFactory.

Linguagem de programação: A linguagem de programação selecionada para o projeto foi a J2SDK (*Java 2 Software Development Kit*) que é uma linguagem que dá suporte à orientação a objetos e independe de plataforma. A programação orientada a objetos (OOP - *Object Oriented Programming*) é uma metodologia de desenvolvimento de software em que um programa é percebido como um grupo de objetos que trabalham juntos [Deitel and Deitel 2005].

PostgreSQL: O que deve ser levado em consideração na determinação de qual SGBD (Sistema de Gerenciamento de Banco de Dados) usar é: eficiência na execução de consultas a tabelas grandes, e na atualização de tabelas, granulação de *locks* para controle de transações concorrentes, recursos para *backup* e recuperação de dados *restore* e eficiência do *driver* de comunicação da aplicação com o SGBD [Guimarães 2003]. Duas ferramentas foram testadas, MySQL (ver `http://www.mysql.com`) e PostgreSQL (ver `http://www.postgresql.org`). Concluiu-se que o MySQL tem como foco principal a agilidade e o PostgreSQL busca oferecer recursos avançados a banco de dados de grande porte. Portanto, a escolha do PostgreSQL se fez necessária devido às operações futuras que o sistema poderá realizar.

Hibernate: Um framework é uma aplicação reusável e "semi-completa" que pode ser especializada para produzir aplicações personalizadas. O framework Hibernate (`http://www.hibernate.org`) é implementado na linguagem Java e permite mapear objetos da camada de domínio da aplicação para SGBD's específicos de forma muito simples.

UnBBayes: UnBBayes (`http://unbbayes.sourceforge.net`) é um software para modelagem, aprendizado e raciocínio utilizando o formalismo de redes bayesianas. Sua API (*Application Programming Interface*) permite que os cálculos de probabilidades sejam feitos. Este framework também foi escrito na liguagem Java.

4.2 Casos de uso

O sistema é composto por três grupos de usuários (atores) – secretário, perito e perito máster – que irão interagir diretamente com ele, sendo essa interação limitada às permissões concedidas a cada tipo de ator.

O secretário tem as seguintes permissões:

- cadastrar um novo processo de estudo de paternidade, sendo necessário informar o tipo de estudo e a localização das pessoas vinculadas ao processo. Essas informações são de fundamental importância para as atividades subsequentes, uma vez que os algoritmos utilizados para o cálculo, assim como as interfaces utilizadas para a inserção dos dados, estão vinculados ao tipo de estudo de paternidade. Além disso, as frequências que são usadas nos cálculos estão vinculadas à localização das pessoas envolvidas no processo como, por exemplo, as frequências utilizadas no cálculo de paternidade de pessoas do Estado de Alagoas devem ser as frequências alélicas do Estado de Alagoas que são diferentes das frequências de outro Estado qualquer;
- inserir os dados pessoais das pessoas vinculadas ao processo;
- excluir um processo de estudo de paternidade, sendo necessário informar o número do processo.

O perito tem as seguintes permissões:

- inserir os perfis genéticos das pessoas vinculadas ao processo;
- excluir os perfis genéticos das pessoas vinculadas ao processo.

O perito máster, além das permissões dos outros dois atores, pode:

- efetuar o cálculo do estudo de paternidade;
- verificar o resultado do cálculo do estudo de paternidade;
- excluir o resultado do cálculo do estudo de paternidade;
- inserir/excluir marcador, alelo, localização, frequência alélica, classificação de pessoa e tipo de estudo de paternidade.

Como se pôde observar, há uma hierarquia entre os atores, imposta pelas restrições de integridade do próprio banco de dados, como por exemplo, a exclusão de um processo (atividade realizada pelo secretário) só é possível se não houver perfis associados às pessoas do processo (sendo a manipulação dos perfis atividade do perito), assim como, a exclusão dos perfis (atividade realizada pelo perito) só é possível se não houver cálculo associado ao processo (sendo a manipulação do cálculo atividade do perito máster). Nessa hierarquia, o perito máster está no topo, tendo acesso irrestrito a todo o sistema, vindo logo abaixo o perito, cujas atividades de excluir e inserir perfis estão limitadas às ações do perito máster e do secretário respectivamente, e na base encontra-se o secretário, cuja atividade de excluir processos está limitada às ações do perito e do perito máster.

4.3 Arquitetura do sistema

O sistema é composto por cinco camadas:

View Contém todas as interfaces do sistema por meio das quais os usuários irão interagir com o software.

Controller Como o próprio nome sugere, é o responsável por coordenar (controlar) a interação entre a camada de apresentação (View) e as demais camadas (Model, Persistence e Util).

Model É a camada de negócio do sistema (domínio do sistema).

Persistence É a camada de persistência que contém os objetos necessários à manipulação dos dados do sistema.

Util É a camada de utilitários, onde se encontram as classes que irão efetuar os cálculos de paternidade e o carregamento dinâmico de tabelas de frequências (em formato `csv`).

A arquitetura do sistema foi modelada dessa forma com o intuito de diminuir o acoplamento, a fim de que os módulos pudessem ser desenvolvidos de maneira independente. A Figura 5 corresponde à arquitetura do sistema.

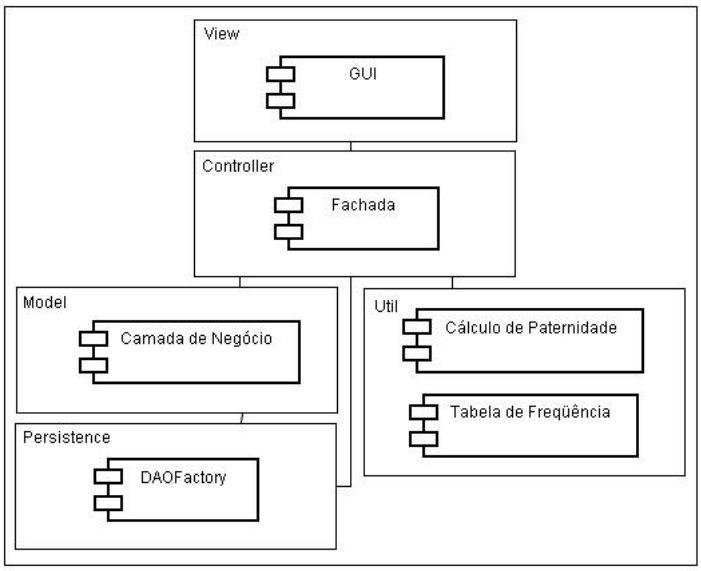

Figura 5. Arquitetura do sistema

5 Conclusão

Neste trabalho foi desenvolvido um ambiente para análise forense de DNA no que tange ao estudo de paternidade, atividade esta comum em laboratórios de genética forense, tais como o Laboratório de DNA Forense da UFAL, no qual o software foi validado através de dados reais do mesmo.

Modelar um sistema com essas características, requer um conhecimento multidisciplinar, envolvendo conceitos de genética forense e computação. Por ser um sistema crítico, uma vez que lida com identificação humana, exige um entendimento profundo do processo de análise de DNA para fins forenses, o que o diferencia dos sistemas de softwares encontrados no mercado, evidenciando assim a importância em se utilizar boas práticas de Engenharia de Software no seu projeto. Dentre os sistemas existentes atualmente que realizam estudos forenses envolvendo DNA, tem-se o CODIS do FBI que realiza estudos criminais e o FAMILIAS (`http://www.math.chalmers.se/~mostad/familias/`.

Como já mencionado, o desenvolvimento deste sistema foi guiado pelos requisitos dos biólogos do Laboratório de Genética Forense da UFAL. Um requisito de grande importância em sistemas de análise forense de DNA é a inserção automática dos dados de saída contendo os perfis genéticos dos indivíduos gerados pela máquina de PCR. Em grande parte dos software disponíveis no mercado, esses dados devem ser inseridos manualmente, causando suscetibilidade a erros e exigindo muito trabalho por parte do biólogo. No ambiente aqui apresentado, este problema é solucionado através da interpretação automática (*parsing*) do arquivo (em formato `csv`) gerado pela máquina de PCR.

Os trabalhos futuros estão direcionados para o tratamento de casos de disputa de paternidade mais complexos. Por exemplo, deseja-se saber se um homem falecido é o pai

de um criança e não se dispõe de seu perfil genético, porém este mesmo homem possui irmãos. As redes bayesianas proporcionam a resolução destes tipos de casos complexos devido ao seu poder de tratamento de situações de incerteza.

Referências

Butler, J. M. (2005). *Forensic DNA Typing*. Elsevier Academic Press.

Dawid, A. P., Mortera, J., Pascali, V. L., and Boxel, D. V. (2002). Probabilistic Expert Systems for Forensic Inference from Genetic Markers. *Scandinavian Journal of Statistics*, 29:577–595.

Deitel, H. M. and Deitel, P. J. (2005). *Java: como programar*. Prentice-Hall, Porto Alegre.

Evett, I. W. and Weir, B. S. (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates.

Fenton, N., Krause, P., and Neil, M. (2002). Software Measurement: Uncertainty and Causal Modelling. *IEEE Software*, 19(4):116–122.

Goodwin, W., Linacre, A., and Hadi, S. (2007). *An Introduction to Forensic Genetics*. Wiley.

Guimarães, C. C. (2003). *Fundamentos de bancos de dados: modelagem, projeto e linguagem SQL*. UNICAMP, Campinas, SP.

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*. Springer.

Laudon, K. C. and Laudon, J. P. (2004). *Sistemas de Informações Gerenciais*. Pretince-Hall.

Lee, J. W., Lee, H.-S., Park, M., and Hwang, J.-J. (2001). Paternity determination when the alleged father's genotypes are unavailable. *Forensic Science International*, 123:202–210.

Morling, N., Alen, R. W., Carracedo, A., Geada, H., Guidet, F., Hallenberg, C., Martin, W., Mayr, W. R., Olaisen, B., Pascali, V. L., and Schneider, P. M. (2002). Paternity Testing Comission of the International Society of Forensic Genetics: recomendations on genetic investigations in paternity cases. *Forensic Science International*, 129:148–157.

Taroni, F., Biedermann, A., Garbolino, P., and Aitken, C. G. C. (2004). A general approach to Bayesian networks for the interpretation of evidence. *Forensic Science International*, 139:5–16.