Universidade Federal de Alagoas
Instituto de Computação

Dissertação de Mestrado

Modelagem de um Ambiente para Análise de DNA em Genética Forense

Felipe José de Queiroz Sarmento

Orientadora:
Profª. Drª. Eliana Silva de Almeida

Co-orientador:
Prof. Dr. Luiz Antonio Ferreira da Silva

Maceió
2006

Felipe José de Queiroz Sarmento

Modelagem de um Ambiente para Análise de DNA em Genética Forense

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência, área de concentração em Bioinformática, pelo Programa Multidisciplinar de Pós-Graduação em Modelagem Computacional de Conhecimento da Universidade Federal de Alagoas.

Orientadora:

Profª. Drª. Eliana Silva de Almeida

Co-orientador:
Prof. Dr. Luiz Antonio Ferreira da Silva

Maceió
2006

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência, área de concentração em Bioinformática, 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 - Orientador
Instituto de Computação
Universidade Federal de Alagoas

Prof. Dr. Luiz Antonio Ferreira da Silva - Orientador
Instituto de Ciências Biológicas e da Saúde
Universidade Federal de Alagoas

Prof. Dr. Alejandro César Frery Orgambide - Examinador
Instituto de Computação
Universidade Federal de Alagoas

Prof. Dr. Alexandre Plastino de Carvalho - Examinador
Departamento de Ciência da Computação
Universidade Federal Fluminense

Maceió, Maio de 2006

Resumo

Os avanços da biologia molecular vêm favorecendo a geração de uma enorme quantidade de informações genéticas em um tempo cada vez menor. Essa capacidade de geração de dados permite que os pesquisadores acelerem o ritmo de suas pesquisas, exigindo a utilização de ferramentas eficientes para o gerenciamento desses dados. Outra necessidade está relacionada com o desenvolvimento de ferramentas computacionais com capacidade de auxiliar na tarefa de análisar e dar um significado biológico a estes dados em um breve espaço de tempo para os pesquisadores. Este trabalho propõe a modelagem de um ambiente de apoio à análise e ao estudo do DNA Forense, cujo principal repositório seja o DNA autossômico. Este ambiente visa dar suporte a identificação de pessoas condenadas ou suspeitas de ter realizado algum tipo de crime contra a sociedade, bem como auxiliar no estudo de paternidade e na busca de pessoas desaparecidas. Este ambiente irá atender ao Laboratório de DNA Forense, da UFAL, que vêm realizando estas atividades. O modelo do ambiente aqui proposto, possui quatro módulos, "estudo de paternidade", "criminal", "desaparecido" e o "banco de freqüência das populações". Os módulos foram modelados de forma que funcionem independentemente, atendendo as especificações inerentes à análise sobre vínculo genético. O sistema foi desenvolvido na linguagem de programação JAVA com banco de dados PostgreSQL. Ambas as ferramentas possuem característica de software aberto e uma relação custo/benefício excelentes.

Palavras-chave: Bioinformática. Sistema de recuperação da informação. Banco de dados. Banco de dados populacional. DNA autossômico - Perfil. Genética forense. DNA forense. Identificação humana.

Abstract

The advances in molecular biology have increased the production of enormous amount of genetic information in a small period of time. This capacity of data production motivated the researchers to increase the rhythm of their researches. This necessity demands the use of efficient softwares in order to manage these data. Besides this, it also demands the development of good softwares in order to assist the researchers in the task of analyzing the data and giving them a biological meaning in a brief space of time. This work proposes a software model that will support the study of Forensic DNA, whose main repository is the autossomic DNA. This software intends to support the researchers in the identification of condemned persons or persons that are suspected of a crime. It also intends to assist the researchers in the study of paternity and the search for disappeared persons. The results of this work will be applied in the Forensic DNA Laboratory of UFAL. The software modeled here has four modules "study of paternity", "criminal", "disappeared people" and the "bank of populational frequencies". The modules were modeled independently from each other, considering the specifications related to the analysis of genetic links. The software was developed using the JAVA programming language together with PostgreSQL database. Both are free software and have an excellent relationship between cost and benefit usage.

Keywords: Bioinformatic. System recovery of the information. Data base. Population data base. Autossomic DNA - Profile. Forensic genetics. Forensic DNA. Identification human.

Agradecimentos

A Deus.

À minha adorável esposa Patrícia, por estar sempre presente com tamanha paciência, incentivo, carinho e amor.

À minha orientadora Profa. Eliana Silva de Almeida por, além de outras coisas, toda a dedicação, a paciência, a compreensão e por ter acreditado logo de início na capacidade de um biólogo no mundo da computação.

Ao Prof.Luiz Antonio Ferreira da Silva e Msc.Dalmo Azevedo, ambos do Laboratório de DNA Forense da Universidade Federal de Alagoas.

Aos professores do Programa Multidisciplinar em Modelagem Computacional de Conhecimento.

Aos funcionário do Instituto de Computação em especial o secretário Vitor M. Torres, por sempre estar cobrando os livros a serem devolvidos para a biblioteca do mestrado e pela organização e difusão de informações.

A todos os companheiros do Programa Multidisciplinar em Modelagem Computacional de Conhecimento-UFAL, em especial às figuras Fábio, Alan e Liliane, por compartilharem diversas situações importantes.

Um agradecimento especial aos alunos do curso regular de graduação em Ciência da Computanção e em especial Tenório César, por colocar em prática e discutir a modelagem do sistema.

À banca examinadora pelas correções e sugestões.

À FAPEAL pelo suporte financeiro.

Sumário

1 Introdução **1**

1.1 Motivação do trabalho 2

1.2 Organização da dissertação 4

2 Fundamentação teórica **5**

2.1 A Estrutura do DNA 5

2.2 Cromosssomos, genes e genótipos 7

2.3 O genoma humano 8

2.4 A reação em cadeia de polimerase - PCR 11

2.4.1 Polimorfismos STRs mediante a técnica de PCR 13

2.5 Genética de populações 15

2.5.1 Modelo de Hardy-Weinberg 16

2.5.2 Proporções alélicas para dois alelos 17

2.6 Softwares para estudo sobre vínculo genético 19

2.6.1 Sistemas para estudos de genealogias 21

3 Estudo da probabilidade de vínculo genético **23**

3.1 Estudo de paternidade 23

3.1.1 Paternidade com um genitor 27

3.2 Exclusões de paternidade 29

3.3 Paternidade em subpopulações 30

3.4 Identificação de pessoas desaparecidas 32

3.5 Probabilidade de coincidência de perfil genético 33

4 O modelo do ambiente para análise de DNA forense 36

4.1 Sistemas de informação 36

4.2 Modelagem do Sistema 37

4.3 Tecnologias usadas 38

4.3.1 Padrões de projetos 38

4.3.2 Linguagem de programação 40

4.3.3 O sistema de gerenciamento de banco dados 40

4.4 O modelo do banco de dados 41

4.5 Arquitetura do sistema 44

4.5.1 Funções gerais 44

4.5.2 Módulo de paternidade 49

4.5.3 Módulo de desaparecidos 53

4.5.4 Módulo criminal 58

4.5.5 Módulo de freqüências alélicas 58

5 Estudo de caso 63

5.1 Dados para o estudo 63

5.2 Módulo de estudo de paternidade 67

5.3 Módulo de busca de desaparecidos 71

5.4 Módulo inserção da tabela de freqüência 74

6 Conclusões 78

Referências Bibliográficas 80

Lista de Figuras

2.1 Organização do DNA no núcleo celular (Modificado de Strachan & Read (2002)). 6

2.2 Visualização de uma cariótipo cromossômico (Modificado de Butler (2005)). 7

2.3 Reação em cadeia da polimerase (Modificado de Brown (2003)). 12

2.4 Demonstração de uma seqüenciamento de estudo de paternidade, com os marcadores D8S1179, D21S11 e D7S820. No primeiro eletroferograma, observamos os *ladders* de cada marcador. De cima para baixo, observamos o perfil do suposto pai, da mãe e da criança, respectivamente.(Fonte: Laboratório DNA Forense - UFAL). 14

4.1 Visão do ambiente organizacional de um sistema de informação voltado para análise do estudo sobre vínculo genético 37

4.2 Modelagem do banco de dados e dos quatro módulos do sistema. 43

4.3 Visão geral da modelagem do sistema, com visualização dos módulos funcionais de estudo de paternidade, pessoas desaparecidas e coincidência de perfil genético. 45

4.4 Diagrama de casos de uso 47

4.5 Arquitetura do sistema 48

4.6 Módulo de estudo de paternidade e sua relação funcional com o pesquisador. 49

4.7 Diagrama de classe do módulo de paternidade. 50

4.8 Diagrama de seqüência para inclusão de dados pessoais. 53

4.9 Diagrama de seqüência para exclusão de dados pessoais. 54

4.10 Diagrama de seqüência para inclusão de perfil genético. 54

4.11 Diagrama de seqüência para exclusão de peril genético. 55

4.12 Diagrama de seqüência do estudo de paternidade. 55

4.13 Modelagem do módulo de busca de desaparecido e suas funcionalidades de inserção e remoção de dados pessoais, inserção e remoção dos perfis genéticos e busca de vínculo genético do perfil genético. 56

4.14 Diagrama de classes do módulo de desaparecidos 57

4.15 Módulo de análise de coincidência de perfil genético em casos criminais. 59

4.16 Diagrama de classe do módulo criminal. 60

4.17 Módulo de inserção das freqüências alélicas de acordo com a localidade a ser usada no estudo de paternidade, criminal ou desaparecidos. 62

5.1 Janela geral do sistema. 67

5.2 Janela com as atividades para o teste de paternidade. 67

5.3 Janela para escolha do tipo de estudo de paternidade. 68

5.4 Entrada de dados pessoais para o caso padrão, mãe, filho e suposto pai. 68

5.5 Busca do tipo e em qual caso será inserido o perfil genético. 69

5.6 Seleção dos marcadores que serão usado no estudo. 69

5.7 Entrada dos perfis genéticos de DNA para mãe, filho e suposto pai. 70

5.8 Busca do processo para a realização do cálculo de paternidade. 70

5.9 Visualização do cálculo de paternidade. 71

5.10 Visão geral da janela do módulo de busca de pessoas desaparecidas. 72

5.11 Janela de cadastro de um nova pessoa desaparecida. 72

5.12 Inserção do perfil genético de desaparecidos no sistema e escolha da origem da amostra. 73

5.13 Janela demonstrando a escolha dos marcadores genéticos que serão armazenados para determinado indivíduo. 73

5.14 Perfil genético do suposto pai sendo inserido para realização da busca no banco de dados. 74

5.15 Resultado da busca no banco de dados de DNA, na qual visualizamos o IP de cada loco e a Probabilidde de Paternidade. 75

5.16 Janela de exclusão de desaparecido. 75

5.17 Entrada da tabela de freqüência dos marcadores genéticos que poderão ser utilizados, de acordo com a localidade, no estudo sobre vínculo genético. 76

5.18 Janela de verificação e inclusão no sistema da tabela e freqüências alélicas. 77

Lista de Tabelas

2.1 Proporções dos tipos de acasalamento e prole de uma população em EHW com genótipos dos genitores nas proporções $p^2 : 2pq : q^2$. 17

3.1 Construção e probabilidade de transmissão dos alelos para o caso padrão do estudo de paternidade. Observamos as duas hipótese, X o suposto pai é o pai da criança e em Y outra pessoa. 28

3.2 Resumo das relações de parentesco Filho, Mãe e Suposto Pai e suas equações. 28

3.3 Resumo das relações de parentesco criança e um progenitor e suas equações. 29

3.4 Probabilidade de transmissão de caracteres em estudo que envolvam subpopulações. G_C, G_M e G_SP, são os respectivos genótipos da criança, da mãe e do suposto pai. IP^1 paternidade para o caso padrão, como descrito na Seção 3.1. IP^2 o suposto pai, o pai e a mãe pertence a mesma subpopulação. IP^3 o suposto pai está intimamente relacionado com o pai. θ é o coeficiente de ancestria. θ_{AT} é o coeficiente de co-ancestria quando o suposto pai está relacionado com o pai. 31

3.5 Relação de parentesco usando o perfil genético dos pais em casos de desaparecidos. . . 33

3.6 Relação de parentesco usando o perfil genético somente com o suposto pai em casos de desaparecidos. 33

3.7 Probabilidade de coincidência de perfil entre o suspeito e a população de referência. . . 35

5.1 Perfis genéticos para o estudo de paternidade envolvendo mãe, criança e suposto pai. IP - Índce de Paternidade; PE - Probabilidade de Exclusão do locos; IPCom. - Índice de Paternidade Combinado; PP - Probabilidade de Paternidade; PEC - Probabilidade de Exclusão Cumulativa. 64

Capítulo 1

Introdução

O avanço biotecnológico nas últimas décadas fez com que a Genética Forense obtivesse grandes êxitos na temática referente à identificação humana (Corte-Real 2004). É indiscutível que a investigação biológica de paternidade, prova clássica para todos os laboratórios de Genética Forense, tenha presenciado progressos significativos, satisfazendo cada vez mais e melhor as necessidades da nossa complexa sociedade (Inman & Rudin 2002).

As bases de dados de DNA para fins de investigação criminal são atualmente bancos de grande interesse dos laboratórios forenses. Devido à experiência acumulada por diversos países em todo o mundo, que desenvolveram uma legislação específica para o gerenciamento de suas bases de dados de DNA Forense. Assim o tratamento automatizado dos perfis de DNA. Este tratamento consiste na comparação sistemática desses perfis, visando encontrar uma prova que auxilie no processo de redução do índice de criminalidade de determinados delitos, em especial aqueles que possuem reincidência (Weir 1996, Primorac et al. 2000).

A importância da análise do DNA em casos forenses não somente se faz presente para identificação de criminosos. O estudo sobre vínculo genético é uma ferramenta de grande importância nos casos de compravação da paternidade. Nesses casos o estudo é caracterizado como reconstituição do perfil genético do suposto pai. Outro caso relacionado com o estudo de paternidade são os estudos de desaparecidos, os quais podem possuir dados de DNA associados a projetos de identificação de crianças de rua, ou de organizações contra o tráfico de menores.

As bases de dados populacionais de marcadores de DNA humanos utilizados na genética forense hoje existentes, em sua maioria, são públicas e acessíveis pela Internet e são de grande importância para a realização da análise estatística dos dados de DNA como prova, a exemplo, os marcadores STRs autossômicos (Ruitberg et al. 2001).

1.1 Motivação do trabalho

O DNA como prova forense, pode ser extraido de uma pequena amostra biológica, como algumas gotas de sangue, que pode ser analisada e conseqüentemente usada para determinar o perfil genético de uma indivíduo. Um perfil genético de DNA usado como uma prova forense consiste em comparar um perfil genético de DNA conhecido (suspeito) com um perfil genético de DNA desconhecido (criminoso, ou de um suposto pai, por exemplo). Se forem iguais então o suspeito é culpado, senão prova-se sua inocência.

A molécula de DNA é muito estável e pode resistir a uma significativa degradação ambiental, o que permite que cientistas forenses obtenham informação de evidências biológicas muito antigas (como ossadas, através do DNA mitocôndrial). A estabilidade da molécula, combinada com as características distintivas do DNA de cada indivíduo e a precisão técnica da análise de DNA atuais, enriquecem ainda mais a tecnologia forense de identificação humana, sendo um componente vital da maioria das investigações policiais.

Os Banco de Dados de DNA seguem diretrizes rígidas em vários países, como EUA, Canadá e Inglaterra, através de especificações legislativas. Só podem ser usadas as amostras biológicas coletadas dos ofensores/criminosos condenados e os perfis de DNA resultantes para propósitos de execução de lei (Council 2001). Nestes países, os Banco de Dados de DNA são usados em conjunto com os procedimentos da justiça e, para segurança de suas populações, asseguram que os criminsos sejam identificados com uma maior eficiência por todas as jurisdições policiais.

O DNA é utilizado como instrumento de identificação e prova. Em estudo de identificação, é possível relacionar os restos mortais de um pessoa e seus familiares, utilizando para isso o DNA nuclear, o Cromossomo Y ou o DNA Mitocondrial. A contribuição para cidadania, está relacionada na busca de um parente que está desaparecido ou que foi seqüestrado. Em grandes desastres, no qual uma grande quantidade de pessoas é vítimada, o estudo sobre o vínculo genético se faz necessário, bem como os casos descritos por Birus et al. (2003) relacionados a vítimas de guerra.

O DNA em casos criminais pode ser utlizado para resolver crimes onde não há suspeito; possibilitando a identificação de suspeitos através das amostras de uma cena do crime ou, eliminando os suspeitos onde não há ligação entre o DNA da cena do crime e um perfil do DNA na base de dados específica para este fim (Primorac et al. 2000, Weir 2004). Só em estudos de paternidade apresentados por diferentes autores são baseados na comparação entre os perfis de DNA da criança e de seus parentes mais próximos, como mãe, pai, tios, primos, etc.(Bernal 1999). Para isso são necessárias pesquisas com os marcadores genéticos utilizados na determinação dos perfis de DNA, sejam RFLP (*Restriction Fragment Length Polymorphisms*), VNTR (*Variable Number of Tandem Repeats*) ou os STR (*Short Tandem Repeats*), esses últimos

serão objeto do nosso estudo.

Dado o alto poder de identificação e discriminação do DNA, um ambiente computacional que auxilie o processo de análise de DNA Forense torna-se um importante instrumento de investigação e prova, auxiliando no combate ao crime e a impunidade. A modelagem de tal ambiente possibilitaria aos poderes judiciário e executivo o uso do DNA como evidência, para vincular suspeitos aos crimes.

O laboratório de DNA Forense de Alagoas da UFAL, que trabalha em parceria com a secretaria de segurança, tem como principal atuação a resolução de testes de paternidade e a identificação de criminosos e pessoas desaparecidas. Até o momento, este laboratório trabalha e realiza todos os seus experimentos manualmente, o que diminui seu ritmo de produção, e dificulta o armazenamento dos seus dados biológicos.

A presente proposta tem por objetivo apresentar a modelagem de um ambiente para análise de DNA em genética forense. Este ambiente trará um avanço nas práticas atuais através da automatização do processo de controle de qualidade e armazenamento desses dados gerados em laboratórios forenses. O modelo do ambiente aqui proposto obedece as especificações técnicas exigidas pela pesquisa forense realizada no Laboratório de DNA Forense de Alagoas na UFAL. A realização deste projeto contribuirá para o desenvolvimento e geração de tecnologia própria, no avanço biotecnológico, para o estado, como também, para o resto do país.

O ambiente computacional proposto conterá quatro módulos que são descritos a seguir:

Investigação Policial e Crimes Sexuais: O Laboratório de DNA Forense de Alagoas desde 1995, vem desenvolvendo atividades em parceiras com o poder judiciário e executivo, enfatizando a resolução de crimes e estudos de casos de paternidade. Este do módulo possibilita, atender melhor a demanda que hoje se apresenta para estes casos. Outra vertente é o estudo sobre crimes sexuais. A maioria desse tipo de delito é realizado por criminosos que agem em série. Este módulo terá função de dar suporte ao processo de identificação do indivíduo vinculado ao material colhido de vítimas que denunciem a agressão e passem por exame de corpo-delito.

Identificação de Corpos e Pessoas Desaparecidas: Este módulo terá a função de armazenamento de dados de pessoas desaparecidas ou carbonizadas, onde não foi possível a sua identificação por vestígios físicos ou de objetos pessoais. O sistema terá como principal repositório o material biológico doado por familiares que reclamem os seus parentes e compareçam como voluntários para doarem material biológico para a determinação do parentesco com os desaparecidos.

Estudo de Paternidade: Este módulo em questão atenderá às damandas de reconhecimento de paternidade, principal atividade do Laboratório de DNA Forense,

em parceria com o sistema judiciário. Os casos previstos relacionados com o referido estudo são: paternidade padrão (criança, mãe e o suposto pai), além de outras relações de parentesco.

Freqüências Alélicas da População: Para a realização do estudo sobre vínculo genético se faz necessário a constituição das freqüências alélicas das populações que serão alvo do estudo sobre o estudo de vínculo genético, seja este um caso criminal, de paternidade ou de desaparecidos.

1.2 Organização da dissertação

Esta dissertação está organizado da seguinte forma: no segundo capítulo serão descritos os principais conceitos da Biologia Molecular e da Genética de Populações, importantes para o entendimento do ambiente proposto; no terceiro capítulo é descrito como a genética de populações trabalha com a probabilidade e a estatística para a determinação da transmissão genética; no quarto capítulo são expostos os principais conceitos necessários para modelar o ambiente e algumas técnicas utilizadas; o quinto capítulo é apresentado um estudo de caso para os módulos de paternidade, desaparecidos e de freqüências alélicas; e seguimos, no sexto capítulo, com as conclusões desta dissertação.

Capítulo 2

Fundamentação teórica

Neste capítulo são descritos os principais conceitos da Biologia Molecular e da Genética de Populações, importantes para o entendimento do ambiente desenvolvido. Na Seção 2.1, é descrita a estrutura do DNA e seus constituintes. Na Seção 2.2, descrevemos como os cromossomos estão organizados no núcleo celular e sua importância no estudo da transmissão das características hereditárias. Um breve estudo do Genoma Humano é realizado na Seção 2.3. A técnica de Reação em Cadeia de Polimerase, procedimento biotecnológico usado na obeção das informações necessárias para o estudo de vínculo genético, é concisamente descrita na Seção 2.4. Na Seção 2.5 são detalhados os conceitos referentes à genética de populações, como o Equilíbrio de Hardy-Weinberg e a caracterização estatística da transferência de caracteres hereditários. Por fim, na Seção 2.6, apresentamos softwares existentes para análise de perfis genéticos que utilizam marcadores moleculares STRs.

2.1 A Estrutura do DNA

O material genético dos seres humanos, conhecido como DNA (ácido desoxirribonucléico) está presente no núcleo de todas as células dos organismos (Figura 2.1). É o DNA que define a constituição genética e especifica a função de cada uma das células que constituem o organismo, ou seja, é o DNA que determina todas as características do indivíduo, as quais podem ou não manifestar-se ao longo de sua vida (Strachan & Read 2002).

O DNA é formado por duas cadeias de nucleotídeos que se enrolam formando uma dupla hélice. Os nucleotídeos são unidades moleculares compostos por um grupo fosfato, um açúcar e uma base nitrogenada. O açúcar e o fosfato são componentes invariáveis nos nucleotídeos e apenas fazem parte da estrutura da molécula de DNA. Já as bases nitrogenadas - citosina (C), timina (T), adenina (A) e guanina (G) - são responsáveis por armazenar toda a informação importante para a síntese proteíca.

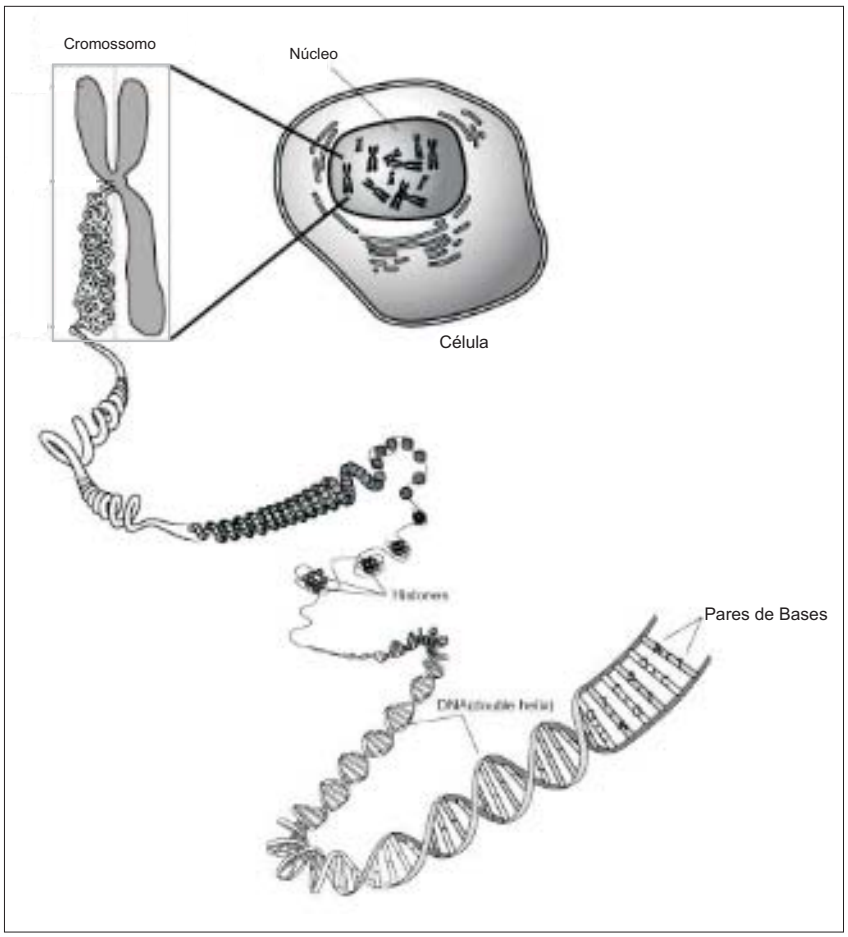

Figura 2.1: Organização do DNA no núcleo celular (Modificado de Strachan & Read (2002)).

A estrutura da molécula de DNA tem sido comparada a uma escada enrolada sobre si mesma em forma de espiral, sendo os corrimãos formados pela parte invariável da molécula - ou seja, o grupo fosfato e o açúcar - e cada degrau por duas bases nitrogenadas ligadas por pontes de hidrogênio. No caso humano, os únicos pares de bases possíveis são A-T e C-G, ou seja, a adenina sempre se liga à timina e a citosina sempre se liga à guanina. Desta maneira, as duas cadeias de nucleotídeos que fazem parte da molécula de DNA são complementares. Por exemplo, se numa cadeia aparece a seqüência de bases AATCCGGT, na cadeia complementar aparecerá a seqüência TTAGGCCA.

2.2 Cromosssomos, genes e genótipos

O DNA de uma célula humana apresenta um comprimento total de quase dois metros, e provavelmente, para facilitar sua organização dentro do núcleo de cada célula é dividido em vários elementos distintos chamados cromossomos. Existem 46 cromossomos na espécie humana, formando 23 pares, dos quais 44 são chamados de autossômicos e dois chamados cromossomos sexuais, por estarem envolvidos na determinação do sexo (Strachan & Read 2002) (Figura 2.2).

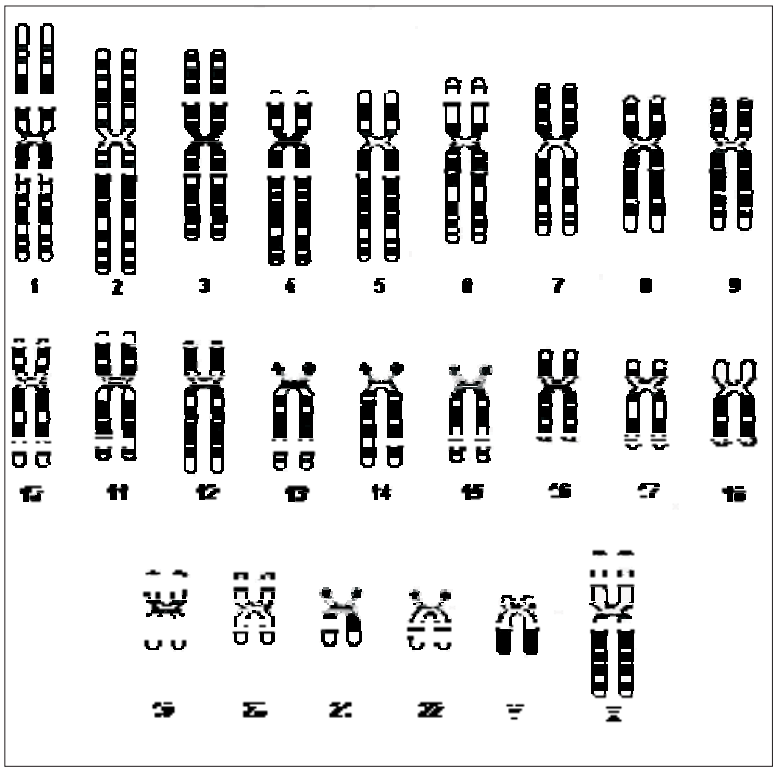

Figura 2.2: Visualização de uma cariótipo cromossômico (Modificado de Butler (2005)).

Nos cromossomo existem regiões específicas contendo seqüências de nucleotídeos

denominados lócus[1]. Em cada lócu pode ser encontrado um gene ou uma seqüência de nucleotídeos não-codificador. Posições ou lócus correspondentes em cromossomos homólogos contém genes responsáveis pela mesma característica genética, embora muitas vezes sejam considerados genes alelos ou simplesmente alelos[2] Também são denominados alelos as seqüências não codificadoras que ocupam lócus correspondentes nos cromossomos homólogos. Quando um indivíduo possui alelos diferentes para um dado lócus, dizemos que ele é heterozigoto com respeito àquela característica genética ou para aquele gene particular. Se tais alelos são idênticos dizemos que ele é homozigoto.

Nos cromossomos homólogos, os alelos presentes em um determinado lócus são o genótipo deste lócu. O termo genótipo é também utilizado para designar o conjunto de todos os alelos presentes em um conjunto de lócus de um mesmo indivíduo. A manifestação morfológica ou bioquímica de qualquer genótipo é denominado de fenótipo. Quando genótipos diferentes manifestam-se como fenótipos diferentes dizemos que os alelos são co-dominantes.

Como nosso sistema de reprodução é sexuado, a formação de um indivíduo provém da fusão de um gameta feminino (óvulo) como um gameta masculino (espermatozóide). Para manter o número cromossômico da espécie humana, cada um dos gametas deve contribuir com a metade do número de cromossomos presentes na outra célula gamética. Por esta razão os gametas são células haplóides - isto é, são células que possuem somente um cromossomo de cada lócu.s Este lócus é dito monomórfico. O conjunto formado pelos alelos de um ou mais lócu, transferidos a um indivíduo por um de seus pais, é chamado de haplotipo. As outras células de nosso corpo são chamadas somáticas e todas elas são diplóides por possuírem dois conjuntos dos cromossomos presentes nos gametas (Strachan & Read 2002).

Assim, em cada par cromossômico encontrado em um indivíduo adulto, cada genótipo de cada lócus, é formado por um alelo proveniente da mãe e outro do pai. Constituindo desta maneira, o DNA representante de todo o material genético que um indivíduo herda de seus pais e por isso é chamando de material genético ou hereditário da célula.

2.3 O genoma humano

O termo Genoma é usada para designar o conjunto de genes ou seqüências de nucleotídeos não codificadoras presentes em uma célula, em um indivíduo ou em uma

[1]Localização cromossômica única que define a posição de um gene individual ou de uma seqüência de DNA.

[2]Umas das várias formas alternativas de um gene ou seqüência de DNA em uma posição específica do cromossomo (lócus). Em cada lócus autossômico um indivíduo possui dois alelos, um herdado do pai e um da mãe.

espécie. A informação genética na célula humana está organizada em dois grandes genomas: o mitocondrial e o nuclear (Brown 2003).

O genoma mitocondrial é constituído por alguns genes extranucleares e tem herança exclusivamente materna. Este tipo de genoma é, em geral, idêntico ao herdado da mãe e às vezes pode ser usado na identificação de indivíduos, por exemplo, quando dispomos apenas de amostras muito degradas, carbonizadas ou em que não é possível obter o DNA nuclear.

O genoma nuclear de uma célula humana haplóide contém um total de 3×10^9 pares de bases que aparecem distribuídos nos 23 cromossomos e cada cromossomo consiste de uma única molécula de DNA de tamanho variado. Aproximadamente 75% do genoma nuclear é constituído por seqüências de nucleotídeos que não se repetem ou que aparecem poucas vezes representadas no genoma humano, as quais são chamadas seqüências simples de DNA. O DNA restante é composto de seqüências que se repetem de centenas a milhões de vezes no genoma, compondo o DNA repetido (rDNA). O total de DNA codificador (genes) compõe somente cerca de $2,5\%$ do genoma humano e encontra-se principalmente entre as seqüências de nucleotídeos que não se repetem na molécula de DNA. Estima-se que no genoma humano existam aproximadamente 100.000 genes.

O DNA simples não codificador pode ser encontrado dentro dos genes, formando os *íntrons*, ou pode ser genes que, se acredita, foram um dia ativos mas que perderam sua atividade ao longo da evolução (os pseudogenes), ou ainda pode ser uma seqüência dispersa entre os genes - o DNA extragênico.

O DNA repetido, que compõe aproximadamente 25% do genoma nuclear, pode ser classificado em DNA codificador, formador das famílias de multigenes, e em DNA não codificador, que compõe DNA extragênico. Este último tipo de DNA não inclui genes funcionais e é composto por seqüências que se repetem em *tandem* (ou seja, uma após a outra) e por seqüências dispersas no genoma que se repetem individualmente. As seqüências que se repetem em *tandem* são classificadas de acordo com o tamanho médio das unidades de repetição. Segundo este critério podem ser classificadas como DNA satélite, minissatélite ou microssatélite.

O DNA satélite compreende seqüências relativamente grandes - mais de 60 nucleotídeos - e sem atividades de transcrição. Ainda pouco se sabe sobre este tipo de DNA. No DNA minissatélite, a seqüência repetida é de tamanho entre 15 a 20 mil pares de bases (base pair - bp). As famílias de DNA microssatélites incluem seqüências muito pequenas, repetidas em *tandem*, contendo entre 1bp e 4bp que aparecem distribuídas ao longo do genoma.

As seqüências de DNA microssatélites é um tipo de DNA com maior interesse para nosso trabalho, serem uma ferramenta de grande importância para Biologia Molecular. Através do estudo destas seqüências, algumas áreas da genética como a locali-

zação de genes, o diagnóstico de doenças genéticas, a identificação de indivíduos e a determinação do vínculo genético, têm-se beneficiado enormemente, já que podemos identificar o perfil genético ou *fingerprinting* molecular de cada indivíduo utilizando uma amostra qualquer de tecido. Esta identificação só é possível porque, salvo em casos de gêmeos idênticos, não existem dois indivíduos com o mesmo genótipo e também porque o DNA de um indivíduo é igual em qualquer célula de seu organismos e em qualquer tempo.

Os perfis de DNA humano fornecem uma poderosa ferramenta na comparação de amostras biológicas tais como manchas de sêmen ou sangue, fios de cabelo, etc., que aparecem como possíveis provas na cena de um crime. Estes perfis de DNA também são utilizados para determinar relações familiares em disputa de paternidade, casos de imigração, e na identificação de ossadas.

Especificamente, o perfil genético refere-se somente às análises de hibridização das seqüências repetidas dispersas. essa técnica, embora valiosa no trabalho de investigação biológica. A técnica de PCR com STR evita problemas como:

- Uma quantidade relativamente grande de DNA é necessária, pois algumas técnicas anteriores a STR depende da análises por hibridização. A técnica datiloscopia genética (*genetic fingerprinting*) não pode ser utilizada com as quantidades ínfimas de DNA de cabelos e manchas de sangue.
- A interpretação da datiloscopia genética pode ser difícil, em decorrência de variações nas intensidades dos sinais de hibridização. Em um processo judicial, pequenas diferenças na intensidade das bandas entre uma perfil-teste e outra realizada de um suspeito podem ser suficientes para que este seja inocentado.
- Embora os sítios de inserção das seqüências repetidas sejam variáveis, existe um limite dessa variabilidade e, portanto, uma pequena chance de que dois indivíduos não-relacionados possam ter as mesmas, ou pelo menos muito semelhantes, os perfis, ocorrendo como no caso descrito anteriormente.

Perfis utilizam seqüências polimórficas chamadas STR (a sigla vem de *Short Tandem Repeats*, ou sejam Pequenas Repetições em *Tandem*). Uma STR é uma seqüência curta de 1 a 13 nucleotídeos de comprimento, que é repetida várias vezes em um arranjo $[CA]_n$, na qual n, é o número de repetições, está normalmente entre 5 e 20.

O número de repetições em uma determinada STR é variável, pois repetições podem ser adicionadas ou, menos freqüentemente removidas por erros que ocorrem durante a replicação do DNA. Na população como um todo, devem existir em torno de 10 versões diferentes de uma determinada STR, cada um dos alelos caracterizados por um número diferente de repetições. No perfil de DNA, os alelos de um número selecionado de STRs diferentes são determinados, isso pode ser rapidamente obtido

e com quantidades muito pequenas de DNA por meio de PCRs com iniciadores que se anelam às seqüências de DNA em ambos os lados de uma repetição, como será melhor explicado nas próximas seções. Uma breve descrição desses marcadores pode ser encontrada em Weir (1996) e em Butler (2005).

2.4 A reação em cadeia de polimerase - PCR

O conhecimento acerca da técnica de PCR que expomos nessa seção demonstrará como é possível obter as informações genéticas necessárias para identificar amostras de DNA.

A técnica de PCR resulta na amplificação seletiva de uma região escolhida de uma molécula de DNA (ver Figura 2.3). Qualquer região de qualquer molécula de DNA pode ser selecionada, desde que as seqüências nas extremidades dessa região sejam conhecidas, pois para realizar uma PCR, dois pequenos oligonucleotídeos[3] devem hibridizar[4] com a molécula de DNA, um com cada uma das fitas da hélice dupla. Esse oligonucleotídeos, que atuam como iniciadores para as reações de síntese de DNA, delimitam a região que será amplificada (Figura 2.3 - momento 1, 2 e 3).

Via de regra, a amplificação é realizada pela enzima DNA-polimerase[5] I de um organismo conhecido como *Thermus aquaticus*. Esse tipo de organismo vive em ambientes quentes e muitas das suas enzimas, incluindo a polimerase de *Taq*, são termo-estáveis, o que significa resistência à desnaturação pelo calor. Como se tornará evidente a seguir, a termo-estabilidade da polimerase de *Taq* é essencial para a metodologia da PCR.

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 (Figura 2.3 - momento 4). A mistura é então aquecida a 94^oC, para que as fitas recém-sintetizadas separem-se do molde (Figura 2.3 - momento 5) e, posteriormente resfriadas, permitindo que mais iniciadores hibridizem com suas respectivas posições, incluindo aquelas das novas fitas sintetizadas, a seguir, a polimerase de *Taq* realiza uma segunda rodada de síntese de DNA (Figura 2.3 - momento 6). O ciclo desnaturação-hibridização-síntese é repetido, geralmente 25 a 30 vezes, resultando, ao final, na síntese de centenas de milhões de cópias do fragmento de DNA amplificado (Figura 2.3 - momento 7).

Ao término de uma PCR, uma amostra da mistura resultante é geralmente ana-

[3]Molécula de DNA sintética, curta e de fita simples, tal como aquela utilizada como iniciador no seqüenciamento de DNA ou na reação PCR.

[4]Formação de uma molécula de fita dupla, por meio do pareamento de bases não-pareadas, que pode ser formada em um polinucleotídeo.

[5]Enzima que sintetiza DNA a partir de um molde de DNA ou RNA.

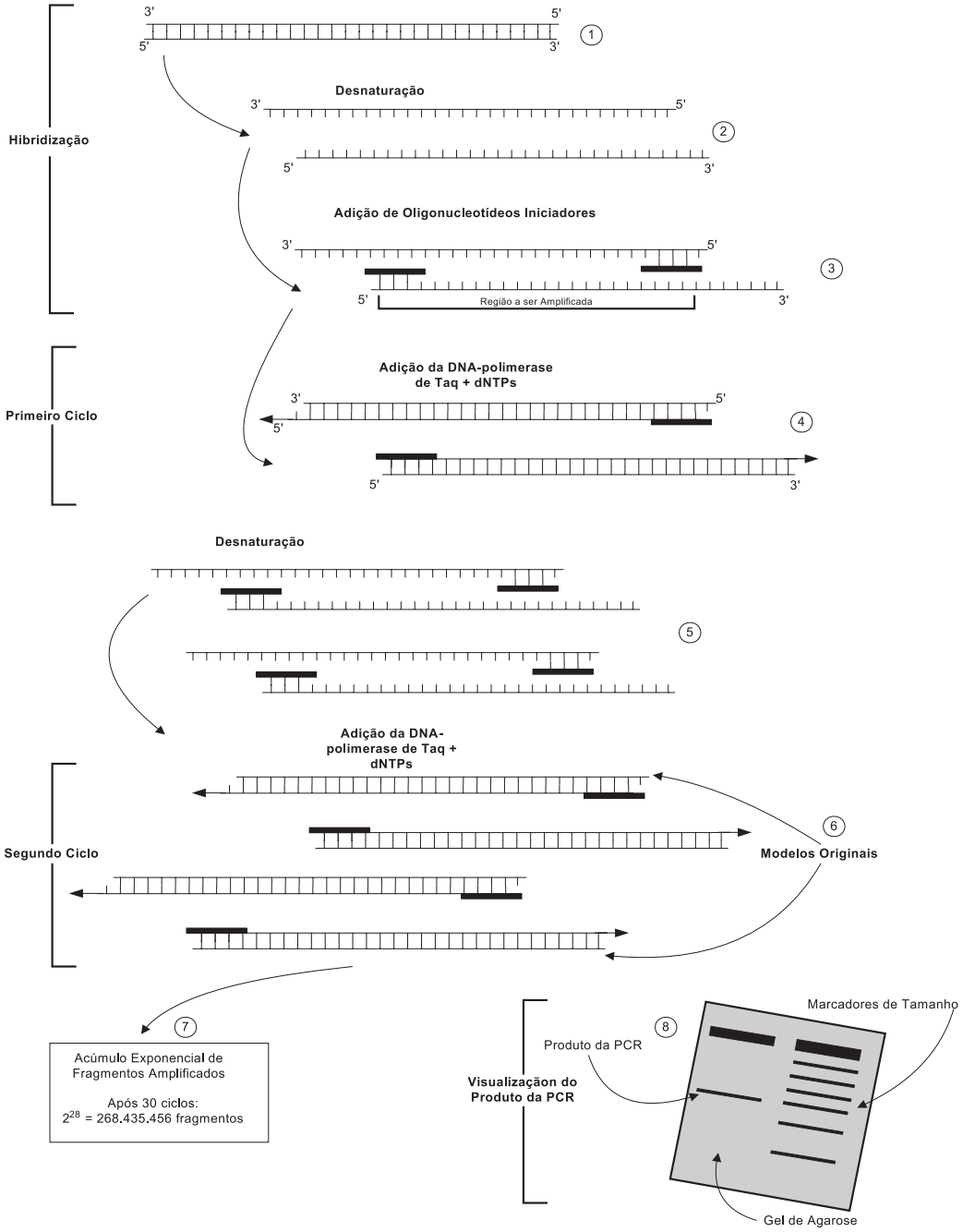

Figura 2.3: Reação em cadeia da polimerase (Modificado de Brown (2003)).

lisada por eletroforese em gel de agarose[6]. Se o DNA foi amplificado com sucesso o fragmento será visível como uma banda discreta após a coloração com brometo de etídio (Figura 2.3 - momento 8).

2.4.1 Polimorfismos STRs mediante a técnica de PCR

Uma vez ealizada a PCR seu resultado é utilizado para analisar os polimorfismos[7] de STRs. A escolha dos STRs em detrimento a outros técnicas, por exemplo VNTRs[8], é devida a facilidade de encontrá-las em pequenas quantidades de DNA e em restos de DNA de baixa qualidade, como ossadas envelhecidas.

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 de forma simultânea, com até 16 lócus de uma vez só (PCR multiplex). Os produtos amplificados, já marcados e desnaturalizados, são separados de acordo com o tamanho, e visualizadas com maior eficiência (diferenciação de uma única base pelo tamanho, ou seja, pelo seu peso molecular). Além disso, são detectados com um sistema a *laser* incorporado a um sistema semi-automático chamando "seqüênciador de eletroforese capilar".

Na Figura 2.4 observamos os resultados de uma eletroforese, reunidas de um seqüênciador de eletroforese capilar, demonstrando os respectivos alelos de cada marcador de acordo com seu peso molecular. este figura ilustra o resultado da análise dos marcadores D8S1179, D21S11 e D7S820 que comprovam a transmissão dos caracteres hereditários dos pai e da mãe para o filho. Observa-se que para o marcador D8S1179 o suposto pai possui $13-14$ repetições, a mãe $13-14$ repetições. A criança herda um alelo da mãe e outro do pai, para esse marcador ele herdou os dois alelos 14, que é visualizada somente uma vez, devido a homozigosidade. Essa mesma dedução pode ser vista para os demais marcadores (Tabbada et al. 2004).

Com relação ao uso do DNA na genética forense, Bernal (1999) realizou uma revisão sobre os lócus, as técnicas atuais e sua utilização nesses casos, bem como discutiu a necessidade de se estabelecer algumas premissas para a seleção desses marcadores, na identificação de criminosos e em estudos de paternidade.

As principais características que os marcadores STRs devem possuir para serem incluídos na reação de PCR multiplex de aplicação no campo da genética forense são:

- Alto poder de exclusão *a priori*, geralmente maior que 90%;

[6]A eletroforese é a separação de moléculas com base na relação entre carga e massa das mesmas. A eletroforese em gel é executada em uma matriz gelatinosa, a qual permite que moléculas de cargas elétricas similares possam ser separadas com base em seus tamanhos.

[7]Refere-se a um lócus presente em um número variável de alelos diferentes ou outras variações na população como um todo.

[8]Polimorfismo de número variável de repetições em *tandem*, ou seja, um após o outro.

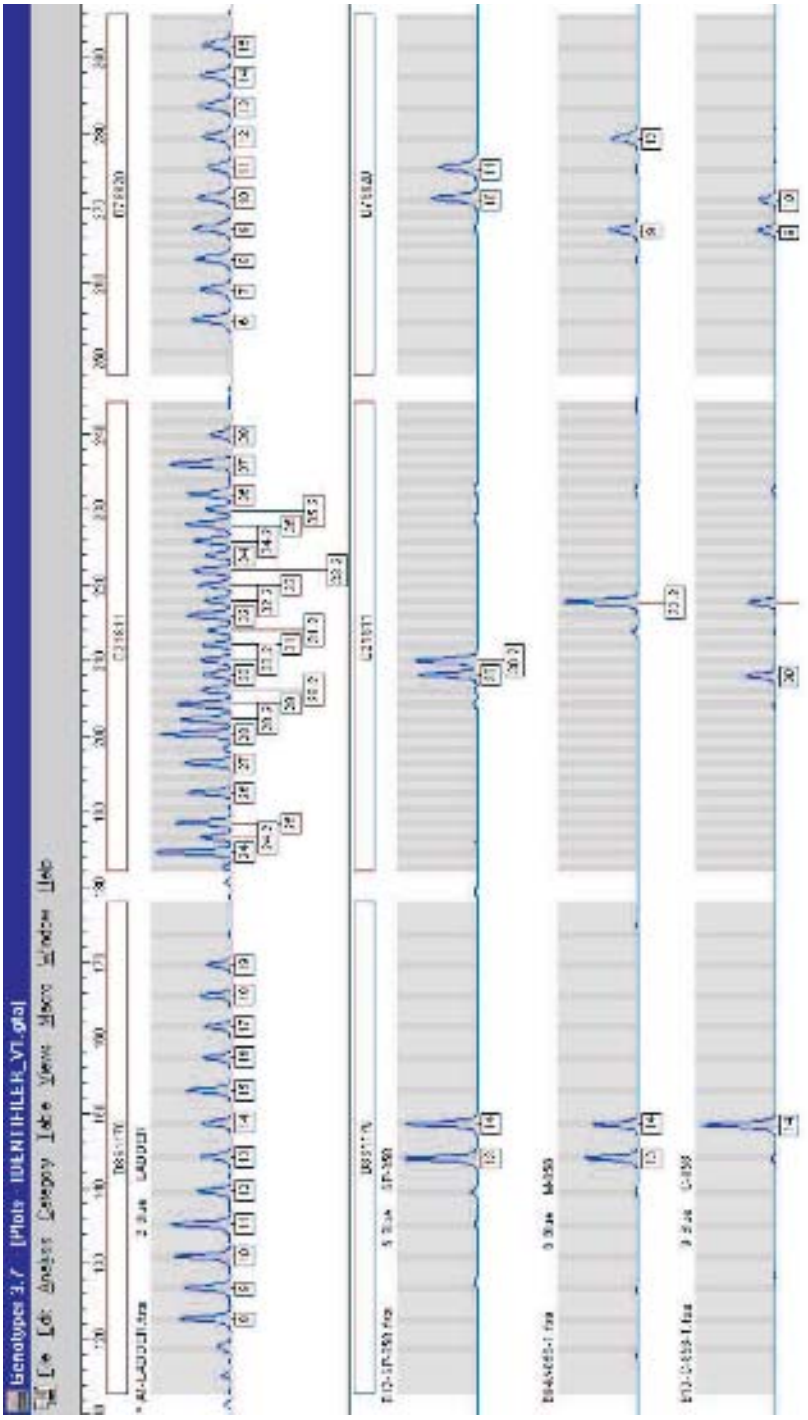

Figura 2.4: Demonstração de uma seqüenciamento de estudo de paternidade, com os marcadores D8S1179, D21S11 e D7S820. No primeiro eletroferograma, observamos os *ladders* de cada marcador. De cima para baixo, observamos o perfil do suposto pai, da mãe e da criança, respectivamente.(Fonte: Laboratório DNA Forense - UFAL).

- Heterozigosidade observada maior que 70%;
- Resultados precisos e reproduzíveis nas reações de PCR Multiplex;
- Baixa taxa de mutação;
- Localização cromossômica dos marcadores sem ligação entre seus lócus;
- Bandas *stutter* baixas;

O "deslizamento" (*slippage*) da polimerase é considerado o mecanismo originador de mutações em microssatélites bem como de artefatos, durante a PCR, que contém uma repetição a menos que o alelo principal, conhecidos como "bandas de repetição" ou *stutter*. O deslizamento da polimerase ocorre quando esta se dissocia do DNA, durante uma pausa na replicação, o filamento em síntese se separa do filamento molde e volta a parear com outra repetição para a frente ou para traz da posição correta (erro de pareamento). A replicação é então retomada completando o processo no qual o comprimento do filamento recém sintetizado pode ser maior ou menor do que o do filamento molde. Este pareamento fora de registro é uma propriedade intrínseca das seqüências de microssatélites e ocorre a frequências elevadas sendo o mecanismo responsável pela grande variabilidade em lócus de microssatélites.

2.5 Genética de populações

Em 1908 o matemático inglês Godfrey Harold Hardy e o médico alemão Wilhelm Weinberg descobriram, independentemente, o princípio relativo às freqüências alélicas numa população (Andrade & Pinheiro 2002, Strachan & Read 2002). Este princípio é a base da teoria da genética populacional. Eles estudaram o efeito do cruzamento entre as freqüências dos alelos numa população em sucessivas gerações, supondo estar ausente de seleção[9] ou de outra força evolutiva. No caso de dois alelos, Hardy e Weinberg demonstraram que, quando os cruzamentos ocorrem de forma aleatória, as freqüências alélicas e genotípicas seguem uma distribuição binomial nas populações de organismos diplóides ($2n$ = conjuntos gênicos[10]).

Neste caso, os alelos A_1 e A_2 possuem freqüências p_{A_1} e $p_{A_2} = 1 - p_{A_1}$, respectivamente. No binômio de Newton $(p_{A_1} + p_{A_2})^2 = 1$ ou $p_{A_1}^2 + 2p_{A_1}p_{A_2} + p_{A_2}^2 = 1$, cada termo está associado a um dos três genótipos possíveis (A_1A_1, A_1A_2, A_2A_2) que ocorrem na população nas seguintes proporções:

- $P_{A_1A_1} = p_{A_1}^2$: proporção do genótipo homozigoto $A_1(A_1A_1)$;
- $P_{A_1A_2} = 2p_{A_1}p_{A_2}$: proporção do genótipo heterozigoto (A_1A_2);

[9] Ação do ambiente que leva à sobrevivência e reprodução diferenciais de uma população ou espécie.
[10] Informação genética de uma população; conjunto de diferentes subpopulações de uma dada espécie.

- $P_{A_2A_2} = p^2_{A_2}$: proporção do genótipo homozigoto $A_2(A_2A_2)$.

2.5.1 Modelo de Hardy-Weinberg

Produção de descendência com variação é a base da evolução. Estas variações podem ser fenotípicas e genotípicas. As fenotípicas decorrem de adaptações a variações ambientais tais como temperatura, concentração de uma substância, etc, sem que a descendência seja afetada. As variações genotípicas são as mais importantes, uma vez que se transmitem às gerações e, além disso, a maioria das variações fenotípicas encontradas na natureza não se deve a adaptações, mas à variações genotípicas.

Considerando que a natureza permanece estável por longo tempo, os organismos que conhecemos em seus ambientes naturais podem ser considerados como estando bem adaptados ao seu meio. Deste modo, a maioria das mutações que ocorrem na descendência, que é devida inteiramente ao acaso, não é essencial ou então é deletéria para a espécie dependendo do genoma em que ocorre.

Alelos não essenciais ou deletérios são introduzidos nas populações de espécies através de mutação ou de introgressão (entrada de organismos externos dentro da população). A reprodução sexuada confere uma vantagem em retardar o acúmulo destas mutações e reordenar constantemente os genes deletérios. Entretanto, os genes deletérios ou não essenciais só sofrem ação da seleção natural quando estão em homozigose, mas não em heterozigose, a menos que esta condição confira um valor pouco adaptativo à espécie. Deste modo, se compara a presença de alelos deletérios à um *iceberg*: somente a ponta, a menor proporção, manifesta-se como homozigose, enquanto a maioria permanece não visível sob a forma de heterozigose.

O Equilíbrio de Hardy-Weinberg nos diz que *em uma população muito grande (uma espécie) com cruzamentos ocorrendo ao acaso, não haverá mudanças na freqüência de genes a menos que mutações sejam introduzidas por introgressão ou aconteça variação da pressão seletiva.*

As proporções alélicas em uma população estão em *Equilíbrio Alélico* quando não há mudança nas mesmas de geração para geração. Em uma população suficientemente grande e na ausência de seleção, migração e mutação, o equilíbrio é atingido após uma geração de acasalamento ao acaso.

O modelo de Hardy-Weinberg define em equilíbrio que é atingido quando são verificadas as seguintes hipóteses:

- ausência de fatores evolutivos como seleção natural[11], migração, mutação[12] e

[11] Ação do ambiente que leva à sobrevivência e reprodução diferenciais de uma população ou espécie.
[12] Alteração na seqüência de bases do DNA, quer seja por substituição, deleção ou inserção de nucleotídeos.

deriva genética[13];

- os indivíduos devem ser diplóides (dois conjuntos gênicos);
- os indivíduos devem ter reprodução sexuada;
- a população analisada deve ter um número grande de indivíduos;
- os cruzamentos devem ocorrer ao acaso, não havendo preferência de acasalamento.

2.5.2 Proporções alélicas para dois alelos

Considerando o caso de dois alelos A_1 e A_2 em um dado lócu, é possível encontrar três genótipos diferentes: A_1A_1, A_1A_2 e A_2A_2. Considerando os três genótipos possíveis da mãe e os três respectivos do pai, um cruzamento resulta em nove tipos possíveis de acasalamento, como é mostrado em Andrade e Pinheiro (2002). A probabilidade associada a cada tipo é facilmente calculada supondo que a população esteja em Equilíbrio de Hardy-Weinberg. Estes cálculos são listados na Tabela 2.1.

Suponha que a probabilidade de encontrar o alelo A_1 na população seja $p = p_1$, e do alelo A_2 seja $q = p_2$. Na Tabela 2.1, o primeiro tipo de acasalamento é o caso em que a mãe e o pai são homozigotos para o alelo A_1. A probabilidade do genótipo A_1A_1 é p^2, tanto para a mãe quanto para o pai. Logo, a probabilidade associada ao cruzamento é $p^2 \times p^2 = p^4$. A probabilidade de ocorrência da prole será p^4 para o genótipo A_1A_1, já que os pais somente podem transmitir o alelo A_1 para a sua prole.

Mãe	Pai	Probabilidade	A_1A_1	A_1A_2	A_2A_2
A_1A_1	A_1A_1	$p^2 \times p^2 = p^4$	p^4	0	0
A_1A_1	A_1A_2	$p^2 \times 2pq = 2p^3q$	$1/2(2p^3q)$	$1/2(2p^3q)$	0
A_1A_2	A_1A_1	$2pq \times p^2 = 2p^3$	$1/2(2p^3q)$	$1/2(2p^3q)$	0
A_1A_1	A_2A_2	$p^2 \times q^2 = p^2q^2$	0	p^2q^2	0
A_2A_2	A_1A_1	$q^2 \times p^2 = p^2q^2$	0	p^2q^2	0
A_1A_2	A_1A_2	$2pq \times 2pq = 4p^2q^2$	$1/4(4p^2q^2)$	$1/2(4p^2q^2)$	$1/4(4p^2q^2)$
A_1A_2	A_2A_2	$2pq \times q^2 = 2pq^3$	0	$1/2(2pq^3)$	$1/2(2pq^3)$
A_2A_2	A_1A_2	$2pq \times q^2 = 2pq^3$	0	$1/2(2pq^3)$	$1/2(2pq^3)$
A_2A_2	A_2A_2	$q^2 \times q^2 = q^3$	0	0	q^4
		Soma da Prole	P^2	$2pq$	q^2

Tabela 2.1: Proporções dos tipos de acasalamento e prole de uma população em EHW com genótipos dos genitores nas proporções $p^2 : 2pq : q^2$.

No caso do pai e da mãe serem heterozigotos, A_1A_2, a proporção genotípica de A_1A_2 é $2pq$. Portanto, este tipo de cruzamento ocorre com a probabilidade de $4p^2q^2$.

[13]Mecanismo evolutivo que resulta da oscilação nas freqüências alélicas de uma geração a outra em uma população.

Para um cruzamento do tipo $A_1A_2 \times A_1A_2$, como tanto o pai quanto a mãe transmitem cada um de seus dois alelos com probabilidade 1/2, os pares (A_1,A_1), (A_1,A_2) ou (A_2,A_2) de alelos (resultado em genótipos) são transmitidos com probabilidades de 1/4, 1/2 e 1/4, respectivamente (Presciuttini et al. 2002).

Naturalmente, a probabilidade da ocorrência da prole ter genótipo A_1A_1, considerando cruzamentos qualquer, é dada por:

$$P(A_1A_1) = P(A_1A_1|A_1A_1 \times A_1A_1) \times P(A_1A_1 \times A_1A_1) \quad (2.1)$$
$$+ P(A_1A_1|A_1A_1 \times A_1A_2) \times P(A_1A_1 \times A_1A_2) \quad (2.2)$$
$$+ P(A_1A_1|A_2A_1 \times A_1A_1) \times P(A_1A_2 \times A_1A_1) \quad (2.3)$$
$$+ P(A_1A_1|A_1A_2 \times A_1A_2) \times P(A_1A_2 \times A_1A_2) \quad (2.4)$$

A Equação (2.1) é usada para o cálculo de transmissão de caracteres hereditários, quando filho e os pais são homozigotos. Já na Equação (2.2) calcula a probabilidade da mãe homozigota e o pai heterozigoto ter um filho homizigoto. A Equação (2.3) analisa o caso contrário descrito anteriormente e com Equação (2.4) verifica a probabilidade dos pais heterozigotos terem filhos homozigotos.

Logo, para o cruzamento $A_1A_2 \times A_1A_2$, o genótipo A_1A_1 da prole ocorre com probabilidade:

$$P(A_1A_1|A_1A_2 \times A_1A_2)P(A_1A_2 \times A_1A_2) = \frac{1}{2}4p^2q^2 = p^2q^2$$

Analogamente, para o cruzamento $A_1A_2 \times A_1A_2$, os genótipos A_1A_2 e A_2A_2 da prole ocorrem com probabilidade $\frac{1}{2}4p^2q^2$ e $\frac{1}{4}4p^2q^2$, respectivamente.

O somatório das probabilidades de cada genótipo da prole observando todos os cruzamentos possíveis é de p^2 para o genótipo A_1A_1, e $2pq$ para o A_1A_2 e q^2 para o A_2A_2, isto é:

1. Genótipo A_1A_1:

$$P_{11} = p^4\frac{1}{2}(2p^3q) + \frac{1}{2}(2p^3q) + \frac{1}{4}(4p^2q^2)$$
$$P_{11} = p^4 + 2p^3q + p^2q^2$$
$$P_{11} = p^2(p^2 + 2pq + q^2)$$
$$P_{11} = p^2$$

2. Genótipo A_1A_2:

$$P_{12} = \frac{1}{2}(2p^3q) + \frac{1}{2}(2p^3q) + p^2q^2 + p^2q^2 + \frac{1}{2}(4p^2q^2) + \frac{1}{2}(pq^3) + \frac{1}{2}(2pq^3)$$
$$P_{12} = 2p^3q + 4p^2q^2 + 2pq^3$$
$$P_{12} = 2pq(p^2 + 2pq + q^2)$$
$$P_{12} = 2pq$$

3. Genótipo A_2A_2:

$$P_{22} = \frac{1}{4}(4p^2q^2) + \frac{1}{2}(2p^3) + \frac{1}{2}(2pq^3) + q^4$$
$$P_{22} = p^2q^2 + 2pq^3 + q^4$$
$$P_{22} = q^2(p^2 + 2pq + q^2)$$
$$P_{22} = q^2$$

Observa-se que as proporções dos genótipos da prole foram as mesmas dos genitores. Tal resultado reforça o conceito de EHW[14], no qual as proporções dos genótipos não mudam ao longo das gerações e se distribuem de acordo com $p^2 : 2pq : q^2$.

De acordo com que foi visto, vamos descrever na próxima seção alguns softwares que utilizam de alguma forma os conceitos exposto nesta seção, embora para diversos fins.

2.6 Softwares para estudo sobre vínculo genético

No início da utilização dos bancos de dados de DNA para fins de estudo sobre vínculo genético ou de identificação, um grupo de cientistas europeus da INTERPOL selecionaram somente 4 marcadores para estudos na Europa. Na mesma época a Rede Européia de Institutos sobre Ciência Forenses (ENFSI) recomendou 7 a serem acrescentados aos 4 já recomendados, totalizando 11 marcadores. Já o Grupo Iberoamericano de Trabalhos de Análises de DNA (GITAD) recomendou um conjunto de 6 marcadores genéticos. Em meados de 2000 foi estabelecido o *Interpol Standard of Loci* (ISSOL), que considera 7 marcadores, mais a *Amelogenina*[15] para determinação do sexo, totalizando para a constituição de um perfil genético 16 marcadores (Díaz et al. 2002).

[14]Refere-se à população cujas freqüências genotípicas podem ser estimadas por $p^2 = (AA), 2pq = (Aa)$ e $q^2 = (aa)$. Estas freqüências são uma indicação de que o acasalamento está acontecendo aleatoriamente e que, portanto, não há endogamia, ou seja, homozigose de uma população que é decorrente de autofecundação ou cruzamento entre indivíduos aparentados.

[15]Gene homólogo do cromossoma X e Y usado na determinação do sexo.

O Sistema de Indexação Combinado do Laboratório de FBI (CODIS) realiza o seu estudo de coincidência de perfil genético com 12 marcadores. O CODIS começou como um projeto piloto em 1990 servindo 14 Estados e laboratórios para uso em crimes com importância local. O Ato de Identificação Criminal de 1994, formalizou a autoridade do FBI para estabelecer um índice de DNA nacional para propósitos de execução da lei, semelhante ao índice de impressão digital que tinha sido implementado em 1924, nos Estados Unidos da América. Em outubro de 1998, o Sistema de Indexação de DNA Nacional do FBI (NDIS) ficou operacional. O CODIS foi implementado como um banco de dados distribuído com três níveis hierárquicos - por local, por estado e o nacional. O NDIS é o nível mais alto na hierarquia do CODIS, e habilita os laboratórios que participam do Programa do CODIS a trocar e a comparar perfis de DNA em nível nacional. Todos os perfis de DNA tem sua origem através dos laboratórios municipais (LDIS). Em seguida é direcionado para o nível estadual (SDIS) e posteriormente para o nível nacional (Council 2001, Díaz et al. 2002).

As informações são armazenadas seguindo duas classificações: um sistema de indexação para casos forenses e outro para criminosos. O primeiro registra perfis genéticos obtidos na cena do crime, e o segundo armazena perfis de indivíduos condenados. O sistema americano se aproxima da nossa proposta, pois realiza o armazenamento e o processamento de informações em resposta à existência ou não de um perfil que já esteja cadastrado no banco. A aproximação da identificação permite que os órgãos de interesse operem os próprios bancos de dados de acordo com sua própria legislação ou com agências legais.

Em 10 de dezembro de 1998, o Governo Canadense criou a sua legislação para dar suporte à identificação do DNA e desenvolveu um banco de dados de DNA para agregar ao código criminal. Juízes passaram então a ter um mecanismo para ordenar que pessoas condenadas por qualquer tipo de crime, sejam designadas obrigatoriamente à fornecer material biológico, como sangue, células bucais ou cabelo, do qual serão armazenados em um banco de dados os perfis genéticos de DNA. Esta legislação foi oficializada em 30 de junho de 2000. O uso da análise de DNA Forense na resolução de crimes está sendo utilizado nos tribunais canadenses com sucesso, sendo comparado e até superando a impressão digital introduzida há mais de um século como prova investigativa.

No Brasil, foi criada em 1994 a Divisão de Pesquisa de DNA Forense (DPDNA) da Polícia Civil do Distrito Federal, a instituição policial pioneira, em nosso país, a realizar, rotineiramente, exames forenses criminais através da análise de material genético. Atualmente, alguns estados da federação iniciaram suas atividades com laboratórios especialmente montados para este fim. Os resultados dos exames são utilizados apenas no âmbito dos laboratórios que os realizou, sendo sempre necessária a utilização de amostra de referência. A produção de informações tem aumentado bastante, havendo a necessidade de um mecanismo eficiente de armazenamento (SENASP 2002).

2.6.1 Sistemas para estudos de genealogias

O uso dos lócus de microssatélite para a identificação humana permitem o desenvolvimento de métodos estatísticos para a dedução das relações familiares a partir de dados genéticos. Hoje, com as ferramentas já disponíveis, é possível deduzir informações genealógicas detalhadas de populações naturais[16], apesar destas técnicas existentes demonstrarem somente aproximações da realidade observada.

Pesquisadores de diferentes áreas de conhecimento têm desenvolvido métodos para extrair informação de genealogia de populações naturais. A habilidade para deduzir as relações genealógicas entre indivíduos em uma população propiciou o surgimento de muitas áreas de pesquisa relacionadas diretamente com o estudo do comportamento humano, sua evolução, conservação e da estrutura populacional. Alguns exemplos destes estão expostos em Thomas & Hill (2000) onde é descrito uma proposta para o cálculo de parâmetros relacionados com a genética quantitativa, usando os marcadores genéticos para a reconstrução da linhagem familiar.

Muitos estudos genéticos requerem análise de ligação genética a qual estuda à transmissão das características genéticas. Esta análise é conhecida como *linkage* e tem como função à detecção do posicionamento de genes responsáveis pela transmissão de doenças hereditárias, ou mesmo detectar a probabilidade de um membro da genealogia manifestar tal doença. O *PedHunter*, descrito por Agarwala et al. (1998), é um pacote de software que facilita a criação e verificação de genealogias dentro de grandes genealogias. Este software utiliza a teoria dos grafos para resolver o problema de *linkage* de uma genealogia com outra. O *PedHunter* utiliza um banco de dados relacional, que armazena as características de uma genealogia. Este software não realiza buscas sobre vínculos genéticos específicos em genealogias previamente conhecidas, ou seja, o seu objetivo é estimar a transmissão de doenças genéticas em uma árvore genealógica, que contém dados de várias gerações armazenadas.

Em Egeland et al. (2000), encontramos uma técnica direcionada para os casos de estudo sobre vínculo genético mais complexos implementada pelo programa *familias*. Este programa é usado quando a reconstituição do perfil genético em estudo é duvidosa, ou seja, quando não temos perfil para o estudo de reconstituição. O procedimento para reconstituir tal perfil está fundamentado nas genealogias, que são estruturadas no ambiente pelo proprio pesquisador. Neste caso, o *familias* trabalha com duas hipótese, ou seja, duas genealogias alternativas. A primeira considera que o suposto pai seja realmente o pai da criança, e a segunda que ele não o seja. A robustez do *familias* é proporcionar ao pesquisador o ordenamento das relações de parentesco considerando a linhagem paterna, como tios, primos, avós, etc. Assim, trabalha-se em uma única análise com múltiplas hipóteses. Alguns sistemas que usam o perfil genético não consideram as mutações que possam ocorrer entre gerações. Se a

[16]População na qual os acasalamentos ocorrem ao acaso.

possibilidade de mutação não for considerada, o pesquisador poderá excluir uma genealogia erradamente. O *familias* possui um tratamento especial, fazendo com que o pesquisador leve em consideração diferentes taxas de mutação. O *familias* também não possue um repositório dos perfis genéticos para estudos futuros. As comparações são direcionadas para a genealogia estruturada pelo pesquisador. Análises de coincidências de perfis genéticos, como as utilizadas em casos criminais ou na busca de desaparecidos em uma base específica não podem ser realizadas neste caso.

As técnicas moleculares que utilizam marcadores genéticos estão cada vez mais sendo usadas para estudos referentes à relação de parentesco. O programa descrito por Goodnight & Queller (1999) utiliza a teoria da probabilidade na verificação do vínculo genético utilizando os marcadores moleculares. O *Kinship* estima a probabilidade de parentesco, considerando a hipótese de se ter duas genealogias; uma de ser a verdadeira genealogia e a outra em ser uma falsa genealogia, e para isso, faz uso de cálculos probabilísticos. *Kinship* realiza estudos de perfis genéticos dos casos mais simples, no qual temos um suposto pai, e queremos verificar a probabilidade de ser realmente o pai ou de ser outra pessoa não vinculada à estrutura familiar. Nos casos complexos ele executa a análise seguindo os métodos descritos por Egeland et al. (2000). Como resposta, o programa fornece informação suficiente para realizar a exclusão, quando nenhum pai for detectado na genealogia.

No próximo capítulo iremos descrever como são realizados os estudos sobre paternidade, coincidência de perfil genético e estudos sobre vínculos genético. Para os casos descritos, veremos a importância da probabilidade e da estatística na estimativa das relações de parentesco.

Capítulo 3

Estudo da probabilidade de vínculo genético

Este capítulo, inicia com a descrição de como é realizado o estudo do vínculo genético com o caso padrão, no qual temos o perfil genético da criança, do suposto pai e da mãe na Seção 3.1. É a partir desse estudo que as outras relações de parentesco são estimadas. Na Seção 3.2 o caso da exclusão de paternidade ou maternidade é tratada. O estudo de subpopulações é abordado na Seção 3.3, na qual são descritos os casos complexos como o estudo de paternidade com perfis genéticos da linhagem paterna ou materna, meio-irmãos, tios, etc. Na Seção 3.4 abordamos os casos de pessoas desaparecidas de forma à diferenciar o estudo de paternidade, pois tratamos da incerteza da transmissão dos caracteres hereditários, tanto quando temos o perfil genético do suposto pai, quando o da suposta mãe. Finalizamos o presente capítulo com o estudo dos casos criminais (Seção 3.5), no qual temos uma amostra obtida no local do crime e uma de um suspeito identificado pela vítima caracterizado como estudo da coincidência de perfil genético.

3.1 Estudo de paternidade

A investigação biológica de paternidade está sendo beneficiada com os avanços adquiridos no campo da genética forense, em especial com a técnica de PCR, já que atualmente se dispõe de um conjunto de instrumentos de marcadores[1] muito amplos e altamente polimórficos[2] e padronizados. Efetivamente é possível abordar com êxito a maioria das investigações de paternidade. Contudo, a análise estatística-genética-populacional é fundamental para a consolidação da prova, como descrito por Bernal (1999).

[1] Um gene ou região cromossômica facilmente identificável, usado para identificação.

[2] A presença de mais de alelo em um lócu. Em lócus forenses, o alelo mais comum geralmente tem a freqüência menor que $0,6$.

A prova biológica necessita de estudos Genético e da Genética de Populações da população de referência (ou seja, estudo da freqüência populacional) e a comprovação do equilíbrio de *Hard-Weinberg* com relação à independência do lócu. Durante o estudo, é necessário obter uma probabilidade de exclusão *a priori* alta (tipicamente superior à $99,9\%$) para assegurarmos que o sistema seja suficientemente eficiente. A análise final da prova mediante o cálculo da probabilidade de paternidade é o índice de paternidade (IP) que é a informação definitiva na determinação do resultado da análise.

A abordagem estatística no cálculo de paternidade apresentada será eminentemente prática. Porém haverá uma necessidade de abordarmos alguns fundamentos teóricos para se descrever outros tipos de casos e análise de evidências (Ayres & Balding 2004).

Quando consideramos a transferência de evidência no estudo de paternidade, no qual as proporsições H_p e H_d refere-se respectivamente que o suspeito é o verdadeiro pai da criança e o segundo, caso contrário. Para um caso de paternidade escreveremos M para a mãe, C para a criança e SP para suposto pai. Seus genótipos serão denotados por G_M, G_C e G_{SP}, respectivamente. As duas proposições para o estudo de paternidade serão:

1. H_p: o suposto pai é realmente pai da criança.
2. H_d: existem um outro homem na população que é o pai da criança.

Estas duas evidências tem uma probabilidade em função de informações de evidência I, que será justificada mais adiante, e da análise genética E. Através do teorema de Bayes podemos expressa-las em forma de "chances":

$$RV = \frac{Pr(H_p|E,I)}{Pr(H_d|E,I)} = \frac{Pr(E|H_p,I)}{Pr(E|H_d,I)} \times \frac{Pr(H_p|I)}{Pr(H_d|I)} \tag{3.1}$$

Direcionamos à atenção para a avaliação da razão de verossimilhança, e precisaremos citar mais dois termos que são usados no teste de parentesco. O primeiro termo é o Índice de Paternidade (IP), o que é simplesmente outro nome para a razão de verossimilhança (RV) na Equação (3.1). Em casos de paternidade simples os termos de RV e IP são intercambiáveis. O segundo termo é a probabilidade de paternidade significando a probabilidade *a posteriori* de paternidade.

Para a probabilidade de paternidade observamos que:

$$Pr(H_d|E,I) = 1 - Pr(H_r|E,I) \tag{3.2}$$

Como não usaremos a evidência I a Equação (3.2) ficará assim,

$$Pr(H_d|E) = 1 - Pr(H_r|E) \tag{3.3}$$

Desse modo a Equação (3.1), poderá ser reescrita em termos das probabilidades *a posteriori* e *a priori* de H_p e rearranjando os termos para:

$$Pr(H_p|E) = \frac{RV \times Pr(H_p|E)}{RV \times Pr(H_p|E) + [1 - Pr(H_p|E)]} \tag{3.4}$$

Se as "chances" (*odds*) *a priori* são iguais a um, significando que a probabilidade de paternidade, *a priori*, é $0,5$, a probabilidade de paternidade, a *a posteriori*, é:

$$Pr(H_p|E) = \frac{RV}{RV + 1} \tag{3.5}$$

E isto é a quantidade que é referida como probabilidade de paternidade.

Segundo Bernal (1999), o trabalho de valorização das provas por parte dos laboratórios consiste em calcular o valor do IP adequado para cada caso. Começaremos pelos casos mais simples, direcionando para um aprofundamento das relações de parentesco. As premissas que utilizaremos no caso mais simples são:

- Ambos os progenitores biológicos (Pai e Mãe) não estão relacionados genéticamente.
- Existe o equilíbrio de Hardy-Weinberg na população de referência (no que implica que não há subestruturação populacional (acasalamento entre parentes), e que os marcadores são independentes e que não há alelos ocultos.
- Não ocorrem mutações.
- A herença é mendeliana com sistema co-dominante, ou seja, a herança dos dois alelos presentes nos marcadores estudados.

Segundo Shoemaker et al. (1998), com exceção da última premissa, temos que ter à certeza do conhecimento e das circunstâncias, pois essas informações podem alterar os resultados.

Para calcular o IP é necessário a tipagem genética do suposto-pai, da mãe e da criança e um número adequado de marcadores. Para o caso padrão, é suficiente entre 12 a 15 marcadores STRs, altamente polimórficos, para alcançar uma probabilidade de exclusão *a priori* de 99%, se não existir ninguém com relação familiar com o suposto-pai.

Partindo do pré-suposto que a maternidade é certa, e que *a priori* se exclui qualquer parente próximo do suposto pai, o IP total será o produto dos IP para cada

marcador. Tradicionalmente expressa-se o IP como o quociente X/Y (denominador/-numerador) sendo $X = P_r(E \mid H_p, I)$, ou seja, consideramos o suposto pai como pai da criança e $Y = P_r(E \mid H_d, I)$ o suposto pai é outra pessoa na população. A evidência genética que temos (E) são os genótipos do suposto pai, mãe e filho, respectivamente representados por (G_{SP}, G_M, G_C). Portanto:

$$IP = \frac{P_r(E|H_p, I)}{P_r(E|H_d, I)} = \frac{P_r(G_{SP}, G_M, G_C|H_p, I)}{P_r(G_{SP}, G_M, G_C|H_d, I)} \tag{3.6}$$

Onde aplicando a Regra do produto de probabilidades, temos:

$$IP = \frac{P_r(G_C|G_{SP}, G_M, H_p, I)}{P_r(G_C|G_{SP}, G_M, H_d, I)} \times \frac{P_r(G_{SP}, G_M, |H_p, I)}{P_r(G_{SP}, G_M|H_d, I)} \tag{3.7}$$

Nem H_p nem H_d incluem informação que afete nossa incerteza em relação a G_M ou G_{SP} de tal modo que a segunda relação é igual a um. Então:

$$IP = \frac{Pr(G_C|G_M, G_{SP}, H_p, I)}{Pr(G_C|G_M, G_{SP}, H_d, I)} \tag{3.8}$$

Haverá muitas possibilidades (G_C, G_M, G_{SP}) para as quais o numerador da relação de máxima verossimilhança seja zero. por exemplo, se $G_M = A_iA_j$ e $G_C = A_iA_k$ com $k \neq j$ e a criança (C) tem que ter o alelo paterno A_k e o suposto pai (SP) não pode ser o pai de C (por exemplo, devido a mutações) se ele tem o genótipo $G_{SP} = A_lA_m$ e $l, m \neq k$.

Abandonaremos a notação estatística a partir daqui e usaremos uma notação mais genética. O desenvolvimento completo pode ser encontrado em vários textos, como no livro de Weir (1996) já mencionado.

Como temos visto, o problema consiste em calcular a probabilidade do genótipo do filho condicionado ao genótipo dos pais sob ambas as hipóteses, ou seja, de o casal ser os genitores e de não ser. Assim, temos que diferenciar quando o filho é homozigoto ou heterozigoto.

Filho ser Homozigoto - A_iA_i**:**

$$P_r \text{ Genótipo Filho} = Pr \text{ Pai transmitir } A_i \times Pr \text{ Mãe transmitir } A_i$$

ou seja,

$$Pr(G_F) = Pr(G_P)^{A_i} \times Pr(G_M)^{A_i} \tag{3.9}$$

Filho ser Heterozigoto - A_iA_j**:**

$$\text{Pr Genótipo Filho} = \text{Pr Pai transmitir } A_i \times \text{Pr Mãe transmitir } A_j + \text{Pr Pai transmitir } A_j \times \text{Pr Mãe transmitir } A_i$$

ou seja,

$$Pr(G_F) = Pr(G_P)^{A_i} \times Pr(G_M)^{A_j} + Pr(G_P)^{A_j} \times Pr(G_M)^{A_i} \tag{3.10}$$

Nos dois casos acima, temos que fazer uma análise mais detalhada e ver qual é a probabilidade de transmissão. Se o progenitor for homozigoto a probabilidade de transmissão do alelo é 1, já que não existe outra opção. Se o progenitor for heterozigoto, a probabilidade de transmissão de cada alelo será a metade, ou seja $0,5$. Se o progenitor não tem alelo, a probabilidade de transmissão será 0.

No caso do "denominador" (X) citado anteriormente, consideramos que o pai não seja o suposto pai, e sim outro indivíduo não relacionado, o que se considera como pai um indivíduo da população, escolhido ao acaso.

Em alguns casos abordados por Lee et al. (2001), o pai é incompatível com a combinação Mãe-Filho, porque o valor de "numerador" (Y)é 0, e portanto o IP será 0, havendo assim, uma exclusão. O tema das exclusões será abordado mais adiante.

Outra forma de abordarmos os cálculos é determinar qual é o alelo obrigatório paterno do filho. Este alelo será único tanto em homozigotos como heterozigotos já que não é transmitido pela mãe (salvo nos casos em que ambos são A_iA_j e em que os alelos obrigatórios serão ambos) (Lee et al. 2001).

No "numerador" o valor será o número de alelos obrigatórios dividido por 2:

- Homizigoto para o alelo obrigatório = $2/2 = 1$
- Heterozigoto para uma alelo obrigatório = $1/2 = 0,5$
- Heterozigoto para ambos os alelos obrigatórios = $2/2 = 2$ ($G_{SP} = G_M = G_C = G = A_iA_j$)

O valor de "denominador" será a freqüência na população do alelo obrigatório e a soma de ambos os alelos. De acordo com o tipo de características herdadas, as combinações geradas são estimadas seguindo as relações exposta na Tabela 3.1 e resumida na Tabela 3.2

3.1.1 Paternidade com um genitor

Esta situação ocorre quando queremos realizar uma investigação de maternidade ou paternidade, sem termos a amostra biológica da suposta mãe ou do suposto pai.

N°	Filho	Mãe	Pai	IP	M_1	M_2	P_1	P_2	Soma	IP
1	A_iA_i	A_iA_i	A_iA_i	X	1	-	-	1	1	
	A_iA_i	A_iA_i	A_iA_i	Y	1	-	-	p_i	p_i	$\frac{1}{p_i}$
2	A_iA_i	A_iA_i	A_iA_j	X	1	-	-	0,5	0,5	
	A_iA_i	A_iA_i	A_iA_j	Y	1	-	-	p_i	p_i	$\frac{1}{2p_i}$
3	A_iA_i	A_iA_j	A_iA_i	X	0,5	-	-	1	0,5	
	A_iA_i	A_iA_j	A_iA_i	Y	0,5	-	-	p_i	$0,5p_i$	$\frac{1}{p_i}$
4	A_iA_i	A_iA_j	A_iA_j	X	0,5	-	-	0,5	0,25	
	A_iA_i	A_iA_j	A_iA_j	Y	0,5	-	-	p_i	$0,5p_i$	$\frac{1}{2p_i}$
5	A_iA_j	A_iA_i	A_jA_j	X	1	1	0	0	1	
	A_iA_j	A_iA_i	A_jA_j	Y	1	p_j	0	p_i	p_j	$\frac{1}{p_j}$
6	A_iA_j	A_iA_i	A_iA_j	X	1	0,5	0	0,5	0,5	
	A_iA_j	A_iA_i	A_iA_j	Y	1	p_j	0	p_i	p_j	$\frac{1}{2p_j}$
7	A_iA_j	A_iA_i	A_iA_i	X	1	0,5	0	0	0,5	
	A_iA_j	A_iA_i	A_iA_i	Y	1	p_j	0	p_i	p_j	$\frac{1}{2p_j}$
8	A_iA_j	A_iA_k	A_iA_i	X	0,5	1	0	0	0,5	
	A_iA_j	A_iA_k	A_iA_i	Y	0,5	p_j	0	p_i	$0,5p_j$	$\frac{1}{p_j}$
9	A_iA_j	A_iA_k	A_iA_i	X	0,5	0,5	0	0,5	0,25	
	A_iA_j	A_iA_k	A_iA_i	Y	0,5	p_j	0	p_i	$0,5p_j$	$\frac{1}{2p_j}$
10	A_iA_j	A_iA_k	A_iA_i	X	0,5	0,5	0	0	0,25	
	A_iA_j	A_iA_k	A_iA_i	Y	0,5	p_j	0	p_i	$0,5p_j$	$\frac{1}{2p_j}$
11	A_iA_j	A_iA_j	A_iA_i	X	0,5	0	0,5	1	0,5	
	A_iA_j	A_iA_j	A_iA_i	Y	0,5	p_j	0,5	p_i	$0,5(p_i+p_j)$	$\frac{1}{(p_ip_j)}$
12	A_iA_j	A_iA_j	A_iA_i	X	0,5	1	0,5	0	0,5	
	A_iA_j	A_iA_j	A_iA_i	Y	0,5	p_j	0,5	p_i	$0,5(p_i+p_j)$	$\frac{1}{(p_ip_j)}$
13	A_iA_j	A_iA_j	A_iA_i	X	0,5	0	0,5	0,5	0,25	
	A_iA_j	A_iA_j	A_iA_i	Y	0,5	p_j	0,5	p_i	$0,5(p_i+p_j)$	$\frac{1}{2(p_ip_j)}$
14	A_iA_j	A_iA_j	A_iA_i	X	0,5	0,5	0,5	0	0,25	
	A_iA_j	A_iA_j	A_iA_i	Y	0,5	p_j	0,5	p_i	$0,5(p_i+p_j)$	$\frac{1}{2(p_ip_j)}$
15	A_iA_j	A_iA_j	A_iA_i	X	0,5	0,5	0,5	0,5	0,5	
	A_iA_j	A_iA_j	A_iA_i	Y	0,5	p_j	0,5	p_i	$0,5(p_i+p_j)$	$\frac{1}{(p_ip_j)}$

Tabela 3.1: Construção e probabilidade de transmissão dos alelos para o caso padrão do estudo de paternidade. Observamos as duas hipótese, X o suposto pai é o pai da criança e em Y outra pessoa.

Filho	Mãe	Suposto Pai	IP
A_iA_i	A_iA_i, A_iA_j	A_iA_i	$\frac{1}{p_i}$
		A_iA_j, A_iA_k	$\frac{1}{p_i}$
A_iA_j	A_iA_i, A_iA_k	A_jA_j	$\frac{1}{p_j}$
		A_iA_j, A_jA_k, A_jA_l	$\frac{1}{2p_j}$
A_iA_j	A_iA_j	A_iA_i, A_jA_j, A_iA_j	$\frac{1}{(p_i+p_j)}$
		A_iA_j	$\frac{1}{2(p_i+p_j)}$

Tabela 3.2: Resumo das relações de parentesco Filho, Mãe e Suposto Pai e suas equações.

Nestes casos, já que temos a estimação probabilística de exclusão *a priori*, temos que utilizar o maior número de marcadores genéticos, já que não temos toda a informação genética do progenitor ausente.

O desenvolvimento é similar ao caso descrito, quando temos a criança, a mãe e o pai, sendo que neste caso a probabilidade de transmissão do genitor é substituída pela freqüência na população do alelo correspondente, tanto no "numerador" como no "denominador". Assim, o "denominador" é a freqüência populacional do genótipo do filho (p_i^2 para homozigoto e $2p_ip_j$ para heterozigoto). No numerador, será 1 se o pai for homozigoto (necessariamente quando for compatível com a criança) e $0,5$ se for heterozigoto (exceto se a criança for A_iA_j, nesse caso será $0,5(p_i + p_j)$).

Filho	Suposta-Mãe ou Suposto-Pai	IP
A_iA_i	A_iA_i	$\frac{1}{p_i}$
	A_iA_j	$\frac{1}{2p_i}$
A_iA_j	A_iA_i	$\frac{1}{2p_i}$
	A_iA_j	$\frac{(p_i+p_j)}{4p_ip_j}$
	A_iA_k	$\frac{1}{4p_i}$

Tabela 3.3: Resumo das relações de parentesco criança e um progenitor e suas equações.

Como podemos observar na Tabela 3.3, se a criança for homozigota e o progenitor também, o alelo obrigatório será evidente. Mais se a criança for heterozigota o IP será a metade quando temos o progenitor, já que não sabemos qual dos alelos será o obrigatório no estudo.

3.2 Exclusões de paternidade

Quando existe uma incompatibilidade genética entre o pai e a criança, falamos de exclusão de paternidade. Existem duas formas de determinarmos uma exclusão, ou seja, existem exclusões de primeira ordem e de segunda ordem.

A exclusão de primeira ordem é caracterizada pela observação direta da presença ou da ausência dos alelos estudados, nos seguintes casos:

- Quando a criança possuir um alelo que está ausente no perfil genético do suposto pai e da mãe.
- Quando não se detecta na criança nenhum dos alelos que estão presentes no suposto pai, mesmo a criança e o suposto pai sendo heterozigotos.

A exclusão de segunda ordem é caracterizada no estado de homozigose direcionando a um resultado negativo. Quando a criança e o suposto pai são homozigotos para os alelos em estudo.

Em ambos os casos descritos, as exclusões podem ser explicadas devido o suposto pai não ser realmente o pai biológico da criança. Diversas circunstâncias genéticas produzem uma aparente exclusão, basicamente as mutações, e neste caso as exclusões de segundo grau apresentam alelos protegidos.

No caso padrão, no qual temos o suposto pai, mãe e a criança, com 13 a 15 marcadores genéticos analisados, encontra-se comumente entre 7 a 10 exclusões. Em casos de paternidade com um único genitor, o número de exclusões típico está entre 4 a 5 para o mesmo número de marcadores. Se realizamos o estudo de paternidade com uma quantidade menor de marcadores genéticos, estaremos diferenciando a análise e reduzindo a informação genética, e o número de exclusões que podemos esperar se reduz. Para termos a certeza da estimação sobre a paternidade, temos que aumentar o número de marcadores genéticos investigados, em concordância com a probabilidade de exclusão *a priori* para cada caso. O método para determinar a exclusão da paternidade seria uma exclusão de primeira ordem e mais de duas exclusões de segunda ordem. Com esses resultados, considera-se que a exclusão está provada. É necessário esclarecer qual seria o valor próximo do ideal para Índice de Paternidade. Segundo (Weir 1996) um valor do IP a partir de 99,73% (equivalente a um IP de 400), considera-se a paternidade como praticamente provada.

3.3 Paternidade em subpopulações

Para populações que não seguem os princípios de Hardy-Wienberg, constata-se baixo nível de relacionamento entre todos os membros de uma mesma subpopulação. Esta situação está bem definida em Lu et al. (2004), que demonstra algumas implicações para o teste de parentesco. A mãe, o suposto pai e o verdadeiro pai, embora não sejam da mesma família, possuem alguma relação em virtude de pertencer a mesma subpopulação. Se as proporções alélicas são conhecidas para esta subpopulação poderemos usar os resultados da Tabela 3.1, mas se a informação existente for as proporções alélicas da população total, teremos que levar consideração a variabilidade genética entre as subpopulações.

Poderemos direcionar as estimações, considerando que não existe mudança para $Pr(G_C|G_M, G_{SP}, H_d)$, uma vez que os genótipos de M e do SP direcionam a probabilidade do genótipo de C. Com relação a H_d não podemos mais supor que os alelos materno e o paterno são independentes.

Neste caso a razão de verossimilhança para H_p, sabendo que este é o pai de C, e H_d sabendo que o pai não está relacionado com C. Depende do tipo de alelo dos genótipos da mãe e do suposto pai em vez do alelo paterno e do genótipo do suposto pai.

Nesse momento se faz necessário expor o uso da correção θ. Este é uma correção

matemática aplicada ao cálculo da freqüência, quando ambos os alelos em um mesmo locus estão correlacionados. Esta correção ajusta a freqüência para comprovarmos a paternidade, na presença de casos em subpopulações(Li et al. 2003).

Assim, em indivíduos heterozigotos A_iA_j em uma população com proporções alélicas $p_i = p_j = p_l = p$ e com estrutura caracterizada por θ ($0,01$ e $0,05$), a razão de verossimilhança é dada por,

$$LR = \frac{(1+\theta)(1+2\theta)}{2[\theta+(1-\theta)p]^2} \tag{3.11}$$

Para indivíduos homozigotos A_iA_i, com $p_i = p$ os efeitos tendem a ser ligeiramente maiores. A razão de verossimilhança é dada por,

$$LR = \frac{(1+\theta)(1+2\theta)}{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]} \tag{3.12}$$

Aplicamos a correção de θ nos casos de estudo de paternidade como descrito na Seção 3.1.1, e neste caso estaremos realizando um estudo de paternidade em subpopulações A Tabela 3.1 descreve o estudo padrão da paternidade, com os respectivos Índices de Paternidade, assim para realizarmos o estudo de paternidade o IP é modificado como mostramos na Tabela 3.4.

G_C	G_M	G_{SP}	IP^1	IP^2	IP^3
A_iA_i	A_iA_i	A_iA_i	$\frac{1}{p_i}$	$\frac{1+3\theta}{4\theta+(1-\theta)p_i}$	$\frac{1}{p_i(1-2\theta_{AT})+2\theta_{AT}}$
		A_iA_j	$\frac{1}{2p_i}$	$\frac{1+3\theta}{2(3\theta+(1-\theta)p_i)}$	$\frac{1}{2p_i(1-2\theta_{AT})+2\theta_{AT}}$
	A_iA_j	A_iA_i	$\frac{1}{p_i}$	$\frac{1+3\theta}{3\theta+(1\theta)p_i}$	$\frac{1}{p_i(1-2\theta_{AT})+2\theta_{AT}}$
		A_iA_j	$\frac{1}{2p_i}$	$\frac{1+3\theta}{2(2\theta+(1-\theta)p_i)}$	$\frac{1}{2p_i(1-2\theta_{AT})+2\theta_{AT}}$
A_iA_j	A_iA_i	A_jA_j	$\frac{1}{p_j}$	$\frac{1+3\theta}{2\theta+(1-\theta)p_j}$	$\frac{1}{p_i(1-2\theta_{AT})+2\theta_{AT}}$
		A_iA_j	$\frac{1}{2p_j}$	$\frac{1+3\theta}{2(\theta+(1-\theta)p_j)}$	$\frac{1}{2p_j(1-2\theta_{AT})+2\theta_{AT}}$
		A_jA_k	$\frac{1}{2p_i}$	$\frac{1+3\theta}{2(\theta+(1-\theta)p_j)}$	$\frac{1}{2p_j(1-2\theta_{AT})+2\theta_{AT}}$
	A_iA_j	A_iA_i	$\frac{1}{p_i+p_j}$	$\frac{1+3\theta}{4\theta+(1-\theta)(p_i+p_j)}$	$\frac{1}{(p_i+p_j)(1-\theta_{AT})+2\theta_{AT}}$
		A_iA_j	$\frac{1}{p_i+p_j}$	$\frac{1+3\theta}{4\theta+(1-\theta)(p_i+p_j)}$	$\frac{1}{(p_i+p_j)(1-\theta_{AT})+2\theta_{AT}}$
		A_iA_k	$\frac{1}{2(p_i+p_j)}$	$\frac{1+3\theta}{2(3\theta+(1-\theta)(p_i+p_j))}$	$\frac{1}{2(p_i+p_j)(1-\theta_{AT})+2\theta_{AT}}$
	A_iA_k	A_jA_j	$\frac{1}{p_j}$	$\frac{1+3\theta}{2\theta+(1-\theta)p_j}$	$\frac{1}{p_j(1-2\theta_{AT})+2\theta_{AT}}$
		A_jA_l	$\frac{1}{2p_j}$	$\frac{1+3\theta}{2(\theta+(1-\theta)p_j)}$	$\frac{1}{2p_j(1-2\theta_{AT})+2\theta_{AT}}$

Tabela 3.4: Probabilidade de transmissão de caracteres em estudo que envolvam subpopulações. G_C, G_M e G_SP, são os respectivos genótipos da criança, da mãe e do suposto pai. IP^1 paternidade para o caso padrão, como descrito na Seção 3.1. IP^2 o suposto pai, o pai e a mãe pertence a mesma subpopulação. IP^3 o suposto pai está intimamente relacionado com o pai. θ é o coeficiente de ancestria. θ_{AT} é o coeficiente de co-ancestria quando o suposto pai está relacionado com o pai.

Além disso, o método que listamos na coluna IP^3 (suposto pai está relacionado com o verdadeiro pai e este é seu parente, por exemplo, são irmãos ou primos) atribui um valor associado a θ_{AT}, usado no estudo de outras relações de parentesco, como descrito por Macan et al. (2003). Se tivermos o perfil genético de irmãos no qual o genitor foi o mesmo pai, atribuímos um valor para $\theta_{AT} = 0,25$. Nos casos de estudo sobre o vínculo genético entre o suposto pai no qual temos somente os perfis genéticos dos meio-irmãos ou dos tios, $\theta = 0,125$. E nos casos nos quais temos primos de primeiro grau, $\theta_{AT} = 0,0625$.

O coeficiente de coancestria θ refere-se a pares de alelos em diferentes indivíduos na mesma subpopulação em relação a pares de alelos da população total. A estimação requer dados de mais de uma subpopulação. Assim, não haveria uma base para a comparação, tão pouco, não haveria conhecimento da variação das proporções alélicas entre as populações. Se existir acasalamentos aleatórios dentro das subpopulações dos alelos, estes terão a mesma relação se eles estão no mesmo ou em diferentes indivíduos.

3.4 Identificação de pessoas desaparecidas

Diferentemente o caso de paternidade descrito na Seção 3.1 existe uma série de casos em que não se pode assumir a confiança da maternidade. Esses são casos típicos de identificação humana através de material biológico em decomposição, como os cadáveres. Também aplica-se em casos de tráfico ou seqüestro de bebês. Nesse tipo de estudo trabalha-se com dois tipos de hipótese: a primeira do casal ser realmente os pais da criança e a segunda de não serem, como foi descrito por Hochmeister et al. (1996) e Wenk & Chiafari (2000).

O problema pode surgir quando no casal a mãe é a mãe biológica, e o pai não é (ou não se sabe até o início da análise). Observa-se então a inconsistência e o estudo direciona-se para a análise de a criança ser realmente descendente dos supostos pais. Algumas questões são levadas em consideração direcionando os estudos mais acurados sobre a população de referência, como aqueles relacionados a constituição da estrutura populacional dos marcadores genéticos, já que em alguns casos os restos mortais são antigos, impossibilitando inferir qual seria a população de origem, ou se podem pertencer a outra população (Stephen et al. 2001).

A forma de tratar essa análise segue os procedimentos descritos na Seção 3.1, para o "numerador", descrito no estudo de paternidade, sendo igual a 1 se ambos os pais forem homozigotos, e $0,5$ se foram heterozigotos e $0,25$ se ambos forem heterozigotos (com exceção se os três forem heterozigotos e iguais, neste casos "numerador" será igual a $0,5$).

O "denominador" será a freqüência do genótipo na população, já que tanto o pai

como a mãe seriam escolhidas ao acaso para a hipótese de não paternidade (Tabela 3.5).

Criança	Mãe - Pai	PI
A_iA_i	A_iA_i-A_iA_i	$\frac{1}{p_i^2}$
	A_iA_j-A_iA_j	$\frac{1}{(2p_i^2)}$
	A_iA_j-A_iA_j, A_iA_j-A_iA_k	$\frac{1}{(4p_i^2)}$
A_iA_j	A_iA_i-A_jA_j	$\frac{1}{2p_i\,p_j}$
	A_iA_i-A_iA_j	$\frac{1}{4p_i\,p_j}$
	A_iA_j-A_iA_j	$\frac{1}{4p_i\,p_j}$
	A_iA_j-A_iA_k, A_iA_k-A_jA_k, A_iA_k-A_jA_l	$\frac{1}{8p_i\,p_j}$

Tabela 3.5: Relação de parentesco usando o perfil genético dos pais em casos de desaparecidos.

A Tabela 3.5 Se o suposto pai for realmente pai da criança e estes forem heterozigotos (A_iA_j), então:

$$\frac{p_1}{2} + \frac{p_j}{2} = \frac{(p_i + p_j)}{2} \tag{3.13}$$

O "denominador" será a freqüência da população do genótipo do filho (o pai poderá ser qualquer indivíduo). Por tanto p_i^2 e $2p_ip_j$, para homozigotos e heterozigotos respectivamente. As relações para esse caso estão sumarizadas na Tabela 3.6.

Criança	Pai	PI
A_iA_i	A_iA_i	$\frac{1}{p_i}$
	A_iA_j	$\frac{1}{2p_i}$
A_iA_j	A_iA_i	$\frac{1}{2p_i}$
	A_iA_j	$\frac{(p_i+p_j)}{4p_i\,p_j}$
	A_iA_k	$\frac{1}{4p_i}$

Tabela 3.6: Relação de parentesco usando o perfil genético somente com o suposto pai em casos de desaparecidos.

3.5 Probabilidade de coincidência de perfil genético

Até o momento tratamos da interpretação dos perfis genéticos relacionados com testes de parentescos. Nesta seção são analisados os procedimentos para o cálculo de probabilidade de coincidência de perfil genético, utilizado na identificação de criminosos. O desenvolvimento e a análise dos cálculos que serão descritos nessa seção são sumarizada de Weir (1996), Hochmeister et al. (1996) e Weir (2004).

Denotamos o genótipo do local do crime como G_L e o genótipo do suspeito como

G_S. Quando os dois genótipos coincidem, denotaremos $G_L = G_S = G$. Consideramos duas proposições,

- H_p, o suspeito deixou a amostra no local do crime; e
- H_d, alguma outra pessoa deixou a amostra no local do crime.

A avaliação da RV depende substancialmente do que queremos estimar, com "alguma outra pessoa", que é o verdadeiro criminoso dado que H_d é verdade. Se for dado H_d, podemos supor que o suspeito e o criminoso são não relacionados, então podemos elimar G_S, como condicionante no denominador. A RV é, entao, o recíproco de $Pr(G_C|H_d, I)$, como na Equação (3.14), que pode ser analisada como a proporção genotípica da população total, ou seja, $Pr(G_C|H_d, I) = 1/\text{freqüência do genótipo}$.

$$LR = \frac{1}{Pr(G_C|H_d, i)} \tag{3.14}$$

A necessidade de se supor a independência entre o criminoso e o suspeito é removida quando procedemos, como se sempre existisse um grau de associação entre os genótipos do suspeito e do criminoso, e calculamos a probabilidade de coincidência

As Equações (3.15) e (3.16) seriam usadas no caso geral em que as hipóteses $Pr(G_C|G_S, H_d, I) = Pr(G_C|H_d, I)$ for duvidosa. Isto significa que esta suposição é usada quando duas pessoas, o suspeito e a pessoa que deixou a mancha no local do crime, pertencessem a mesma subpopulação, mas as proporções alélicas não estão disponíveis para aquela subpopulação. Esta equação forma a base de uma das recomendações do NRC (National Research Council), órgão regulador sobre a novas tecnologias relacionadas com o DNA, descrita por Balding & Nichols (1994).

O valor de θ descreve o grau de relação entre os pares de alelos dentro da subpopulação em relação à população total. Supõe-se mantida a independência alélica, por exemplo, o Equilíbrio de Hard-Weinberg dentro das subpopulações. Mas quanto maior à diferença nas proporções alélicas entre subpopulações, significa que há uma maior independência na população total. Em outras palavras, este tratamento, explicitamente, leva em conta o endocruzamento e a coancestria para todos os indivíduos da população, de modo que o Equilíbrio de Hardy-Weinberg não se ajusta ao nível populacional. Contudo, os níveis de endocruzamento e de coancestria para indivíduos em diferentes famílias serão, geralmente, muito baixos.

O maior efeito da correção de θ decorre da inclusão da informação sobre o genótipo do suspeito na determinação da probabilidade do criminoso ter o mesmo tipo de genótipo. Os efeitos numéricos são pequenos a menos que as proporções alélicas sejam pequenas e θ seja grande.

$$\Pr(A_iA_i|A_iA_i) = \frac{[2\theta + (1-\theta)p_i][3\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)} \tag{3.15}$$

$$\Pr(A_iA_j|A_iA_j) = \frac{2[\theta + (1-\theta)p_i][\theta + (1-\theta)p_j]}{(1+\theta)(1+2\theta)} \tag{3.16}$$

A Tabela 3.7 descreve os possíveis genótipos e sua relação com as Equações (3.15) e (3.16). O objetivo final desta tabela e resumir as relações entre o perfil genético do suspeito e o de uma outra pessoa na população de referência.

G_S	G_A	N° Identidade	Probabilidade
A_iA_i	A_iA_i	2	$\frac{p_i[3\theta+(1-\theta)p_i][2\theta+(1-\theta)p_i][\theta+(1-\theta)p_i]}{(1+\theta)(1+2\theta)}$
A_iA_i	A_jA_j	0	$\frac{2(1-\theta)p_ip_j[\theta+(1-\theta)p_i][\theta+(1-\theta)p_i]}{(1+\theta)(1+2\theta)}$
A_iA_i	A_iA_j	1	$\frac{4(1-\theta)p_ip_j[2\theta+(1-\theta)p_i][\theta+(1-\theta)p_i]}{(1+\theta)(1+2\theta)}$
A_iA_i	A_jA_k	0	$\frac{4(1-\theta)^2p_ip_jp_k[\theta+(1-\theta)p_i]}{(1+\theta)(1+2\theta)}$
A_iA_j	A_iA_j	2	$\frac{4(1-\theta)p_ip_j[\theta+(1-\theta)p_i][\theta+(1-\theta)p_j]}{(1+\theta)(1+2\theta)}$
A_iA_j	A_iA_k	1	$\frac{4(1-\theta)^2p_ip_jp_k[\theta+(1-\theta)p_i]}{(1+\theta)(1+2\theta)}$
A_iA_j	A_kA_l	0	$\frac{(1-\theta)^3p_ip_jp_kp_l}{(1+\theta)(1+2\theta)}$

Tabela 3.7: Probabilidade de coincidência de perfil entre o suspeito e a população de referência.

No próximo capítulo serão descritas as tecnologias usadas e como o sistema foi modelado de acordo com as informações exposta nos Capítulos 2 e 3.