Università degli Studi di Firenze

DOTTORATO DI RICERCA IN
"Statistica Applicata"

CICLO XXV

COORDINATORE PROF. CORRADI FABIO

Pre-Experimental Assessment of Forensic DNA Identification Systems

Settore Scientifico Disciplinare SECS-S/01

Dottorando	**Tutore**
Dott. Ricciardi Federico	Prof. Corradi Fabio
_______________________________	_______________________________
(firma)	*(firma)*

Anni 2010/2012

to S.

Probability is the very guide of life.

Cicero

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read or write.

H.G. Wells

While it is easy to lie with statistics, it is even easier to lie without them.

Frederick Mosteller

Abstract

In this thesis we want to produce a methodology to evaluate a kinship identification system, i.e. the set of models and data used to ascertain the identity of an individual, a probabilistic tool devoted to obtain the likelihood ratio supporting (or contradicting) the hypothesis that an individual, the candidate for identification, is a specific member of a family, conditional on the available familial DNA evidence. The thesis considers the likelihood ratio as a random variable and focuses on the evaluation of the probability that a candidate for identification would be correctly classified exploiting the likelihood ratio distributions conditional on each hypothesis.

The aim of this work is thus to show how it is possible to make statements about the goodness of an identification system and to demonstrate how this can be applied in a great variety of cases. As secondary objective, we want to show how it is possible to obtain the distributions for the likelihood ratio, finding efficient computational methods to cope with the their huge state space.

The proposed system evaluation is specific for each case, does not require any additional laboratory costs, and should be carried out before the identification trial is performed. In a pre-experimental perspective, we want to evaluate whether a system fulfils the requirements of the parties involved.

A further objective is to consider and find a solution for some complicating issues affecting the estimation of mutation rates for STR markers.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 The need for the identification systems' evaluation

This work stems from a collaboration with the International Organization for Migration (IOM), one of the world's leading organizations providing services and advice to governments and migrants. Our contribution consisted in helping migrants to rejoin their relatives still living in the country of origin. The identification procedure usually started with a person, the *Sponsor*, asking for the identification of a *Candidate*, living abroad, as a well specified relative. Usually, no other option about the relationship between the sponsor and the candidate was provided and for this reason, in alternative, the candidate was considered unrelated with the family. Often the sponsor and the candidate were not on a direct familial lineage (e.g. parent-child, grandparent-grandchild) and many results did not strongly support one of the hypotheses. At that time we realized that not only the specific alleged relationship between the sponsor and candidate was relevant in determining the strength of the result, but also that there was a case-specific effect that deserved to be investigated. For these reasons we were persuaded to produce a contribution able to predict whether a forensic identification based solely on the use of the sponsor(s) evidence would be able to produce decisive answers once the candidate data were to be available.

1.1.1 Introduction to the identification problem

The identification of individuals by using genetic data has been adopted in an increasing manner in the last decades. The contexts are quite varied, but can be summarized in two kinds of problems. We call *direct identification* the use of genetic profiles to compare biological samples, such as blood or semen stains collected as evidence from a crime scene, with the profiles from people suspected of having contributed to those samples. Even if matching of profiles does not guarantee without any doubt (as we will show along the lines of this thesis) a common source, this result can be used as decisive evidence in a trial.

The other major use of genetic evidence is to perform an *indirect identification*, the most established version of which being that of paternity test: genetic profiles of mother, child and alleged father are used to make statements about the probability of the evidence if the alleged father would be the actual father. Similar reasoning can be used to identify mothers or to reunite families separated for different reasons. Also disputes about inheritance or identification of remains from deceased people take advantage of indirect identifications based on DNA evidence.

It is commonly acknowledged that the reliability of an identification system increases with the degree of kinship between the family member(s) who require the identification and provide their DNA and the alleged relative's hypothesized position in the pedigree (i.e. the set of kinship relationships among a group of individuals). At the same time the task becomes more difficult if the two alternative hypotheses are very close each other. Finally, the number of family's DNA donors and the number and the specific fragments of the observed DNA have an influence on the effectiveness of the results. For these reasons, we believe that each kinship identification has its own characteristics which deserve to be investigated.

Introducing in advance some of the argument that will be treated along the chapters, the identification procedure is addressed on the basis of the computation of a Likelihood Ratio (LR) supporting the hypothesis that an individual, the candidate for the identification, is a specific member of a family, conditional on the available familial DNA evidence. In standard practice, the LR is almost invariably obtained by forensic laboratories without any consideration of the specific characteristics of each different identification.

One of the main consequences of this lack of specificity in producing the results is that a common methodology to asses the effectiveness of the probabilistic tools for forensic cases, the importance of which has also been recognized by Cook et al. (1998), is not available yet. This reflects also on the activity of whom are called to take decisions, in civil or criminal cases, based on a likelihood ratio. In fact, when scientific results are brought into a court, the problem of the admissibility of expert witnesses' testimony during legal proceedings. In particular in the United States this issue is ruled according to the so called "Daubert standard", named after the U.S. Supreme Court Daubert v. Merrell Dow Pharmaceuticals sentence in 1993 and then amended twice since then, which states the following:

A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:

1. The expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;

2. The testimony is based on sufficient facts or data;

3. The testimony is the product of reliable principles and methods;

4. The expert has reliably applied the principles and methods to the facts of the case.

These principles have also had some international influences: the Canadian Supreme Court expressly adopted them in two cases since 2000 and the creation of a Forensic Science Advisory Council to regulate forensic evidence in the United Kingdom has been suggested since 2005.

In particular, for the purposes of this thesis, the second and the third principles are the most interesting. The former requires a sufficient amount of data, and this request could not always be guaranteed when indirect identification is performed using poorly informative evidence; the latter explicitly mentions the reliability of the methods that can be questioned in some cases if an insufficient amount of information has been used, as we will show ahead in the thesis. With this work we would like to give guidelines to help in judging when satisfactory standard of performance are reached.

1.2 Literature overview

In the previous Section we acknowledged the claim for the performance assessment of a methodology employed for forensic purposes. Despite this, in the literature there are a limited number of contributions in the field of identification through DNA and they focus on the LR obtained using datasets of real or simulated cases for which the identification of the candidate was already known.

As an example, in Mayor and Balding (2006) the authors describe the characteristics of a kinship analysis attempting to determine whether two individuals are half-siblings or are unrelated. Results are obtained simulating a large sample of pairs of individuals: in the half-sibling hypothesis an individual is drawn from the population, then two of her offspring are sampled; otherwise, if the hypothesis of no relatedness holds, two individuals are sampled from the population. Finally, for each sampled pair, the LR and the number of cases for which it does not support the hypothesis used are computed. There, the LR uncertainty concerns all the possible LR arising for the half-sibling problem, but there is no indication on how the identification system is expected to perform in a specific case where some familial DNA donors are actually observed. In the same vein, Evett and Buckleton (1996), using a database of 1401 different individuals on four loci, evaluated the likelihood ratio's empirical distribution originating from criminal identification cases occurring when two traces, both observed, are queried as to whether they belong to the same person. Taroni et al. (2007) obtained the LR distributions for the criminal (i.e direct) identification issue by simulating 100,000 genetic profiles on 16 loci. In the field of kinship analysis, Brenner and Staub (2003) evaluated the LR distribution, only under the identification hypothesis, for 19 different pedigrees simulating for each of them the genetic evidence of 100 familial groups. Lauritzen and Mazumder (2008), using an information-theoretic approach, proposed a measure to evaluate the informativeness of different loci for some different kinship identifications.

1.3 Goals and contributions

Our work is meant to achieve a methodology to evaluate a kinship identification system and to suggest improvements in case of unsatisfactory results. We aim to modify the current practice of treating any kind of indirect identification using a single standard procedure.

Our proposal shares one of the goals of the Design of Experiments (DOE): the purpose of improving the statistical inference by appropriately selecting the conditions under which a crucial unobserved random variable has certain desirable characteristics.

The focus of this thesis is on ascertaining the LR distributions conditionally on all the relevant alternative identification hypotheses before carrying out the identification for specific identification cases, aiming to identify a candidate as an (unavailable and therefore) unobserved member of a family, exploiting the knowledge of some of the family members' genetic profiles, and the familial relationships.

The main contribution of the thesis is to recognise the large variety of behaviour of kinship identification systems related to specific cases. Another result of the work is reaching the computational feasibility of the LR distributions calculation, under both hypotheses in a pre-experimental phase, for a specific, well defined kinship case. Once these distributions are obtained, they can be used in different ways, some of which will be illustrated, to give an evaluation of the system by means of different approaches.

The procedure precedes the usual kinship analysis and it uses only a subset of the data required for the post experimental assessment of the hypotheses under debate. The final LR computation, including the candidate evidence, is recommended only after the system has shown a satisfactory behaviour.

1.4 Outline of the thesis

Chapter 1 gives a brief overview of the topic relevant for this thesis, acknowledges some reference that are known in the literature and includes the motivations, the aims and the contributions given to the field of personal identification by this work.

Chapter 2 provides a preliminary introduction to the genetics topics relevant for comprehension of the following Chapters of the thesis. It starts from the description of the structure of the DNA, then gives the reader an overview on how DNA data are used in forensic applications and finally provides elements of population genetics.

Chapter 3 introduces the fundamental concepts behind the theory of Bayesian networks (BN) and shows how to apply this probabilistic tool to forensic genetics, with special focus on personal identification, also when more complicating features are considered.

Chapter 4 is about the main proposal of the research, that is how to correctly assess the value of a kinship identification system in the pre-experimental phase. First we give an overview of what kinship identification is, then we introduce the models adopted for the work and finally the methodology by which the evaluation of the system is obtained, based on the LR distributions under two competitive hypotheses, is revealed. We also consider alternative approaches for the evaluation and formulate some comments about the strengths and weaknesses of these alternatives.

Chapter 5 describes the computational issues arising while treating the high dimensional spaces of the LR distributions, also giving some information on the software employed to perform the analysis and the relevant Bayesian networks to handle the involved models.

Chapter 6 gives a collection of cases, based on real data, for which the proposed methodology reveals helpful. At first an overview on 71 cases is given, then we will analyse some of them in further details, also performing sensitivity analysis on the adopted models.

Chapter 7 elaborates some insights on how to cope with some difficulties with mutation rates data from various sources and how to handle them in order to obtain corrected mutation rates.

Chapter 8 finally gives a general discussion about the main findings reached with this research.

CHAPTER 2

Genetic background

In the current forensic practice, DNA represents a powerful and reliable source of information capable to help in the treatment of a great variety of identification cases, e.g. paternity testing, criminal investigations, natural disasters and so on. In order to correctly use this important source on data, understand of some basic notions of genetics is required.

2.1 DNA structure and the genome.

Deoxyriboobucleic Acid (DNA) is a nucleic acid containing the genetic instructions used in the development and functioning of all known living organisms. The well known double helix structure of the DNA, suggested for the first time by

Watson and Crick (1953), is the representation of two long polymers composed of simple units called nucleotides. Attached to each of them is one of four types - adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T) - of molecules called nucleobases. It is the sequence of these four bases along that encodes the genetic information.

DNA is contained within the nucleus of a cell in long structures called *chromosomes*. In humans, chromosomes can be divided into two types: *autosomes* and *sex chromosomes*. Gender related traits, which depends on a person's sex, are inherited through the sex chromosomes. The autosomes contain the rest of the genetic hereditary information. Human cells have 23 pairs of chromosomes (22 pairs of autosomes and one pair of sex chromosomes, these latter named XX in women and XY in men), giving a total of 46 per cell. Each chromosome contains a unique strand of DNA (the largest of which, chromosome 1, in about 73mm in length), and the 46 chromosomes are orderly shown in Figure 2.1.

A *gene* is a fundamental unit of heritable genetic information, located in a specific position along the chromosome that is called *locus* (plural, *loci*). The word *allele* indicates the variant form of a gene observed in a given locus, often represented by numeric values. At every locus humans contains two alleles, one inherited from the father and the other one from the mother. The set of alleles owned by an individual at one locus is named *genotype*. In particular if the two alleles in a locus are equal the genotype is *homozygous* and if they differ the genotype is *heterozygous*.

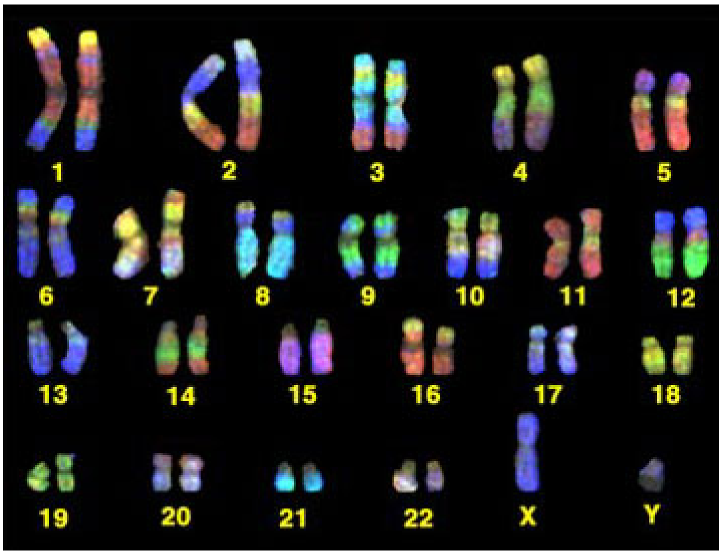

Figure 2.1: The male human karyotype pictured contains 22 pairs of autosomes and the X and Y sex chromosomes (the female karyotype has two X chromosomes). The chromosomes have been labelled with fluorescent probes allowing them to be identified.

2.2 The use of DNA for forensic purposes

The fundamental aim of using DNA data in forensic casework is to obtain an individual profile, a sort of genetic passport, or fingerprint, that is highly discriminating. The ideal situation would be that in which the DNA profile is unique to each individual. However, even people that looks very different to each others, are in fact very similar at genetic level. Comparing the human genome with that of our closest animal relative, the chimpanzee, sharing with us a common ancestors about 6 million years ago, we find that our genomes share 95% of the DNA sequence.

Modern humans have a much more recent common history (fossils analysis helped in dating back to 150000 years ago our common human progenitors), thus approximately 99.9% of human DNA sequences are the same in every person. This is what makes us human being rather than oaks or foxes, but from a forensic point of view there is very little rationale in analysing this part of DNA shared between individuals. Fortunately, the other 0.1% is truly unique and distinguishes one individual from another, unless they are monozygotic twins. These fragments of DNA are of true interest to the forensic scientists. Jeffreys et al. (1984) first reported the DNA profiling technique, also known as "genetic fingerprinting", based upon these high variable non-coding sequences of DNA, called *variable number of tandem repeat* (VNTR) or *minisatellites.*

2.2.1 Short Tandem Repeats

Short Tandem Repeats (STRs) are, nowadays, the most commonly used type of VNTR for DNA profiling. A short tandem repeat in DNA occurs when a pattern of two or more nucleotides are repeated and the repeated sequences are directly adjacent to each other. The pattern can typically range in length from 2 to 5 base pairs (bp) and shows a high level of polymorphism. The allelic state is simply determined by the number of repeats present at the selected locus, so that the following sequence

CATATTGGGCATGCATGCATGCATGCATGCATGCATGAATTCAG

is associated with an allelic value of 7, since the 4 bases CATG are repeated 7 times (in blue), preceded and followed by random sequences.

Currently the process of acquiring DNA evidence is a standard practice in every genetics laboratory and it's mainly based on a biochemical process named *Polymerase Chain Reaction* (PCR). PCR was developed in 1983 by Kary Mullis (who was given the 1993 Nobel Prize® in chemistry for this invention) and consists in a process allowing the amplification of specific DNA sequences by means of

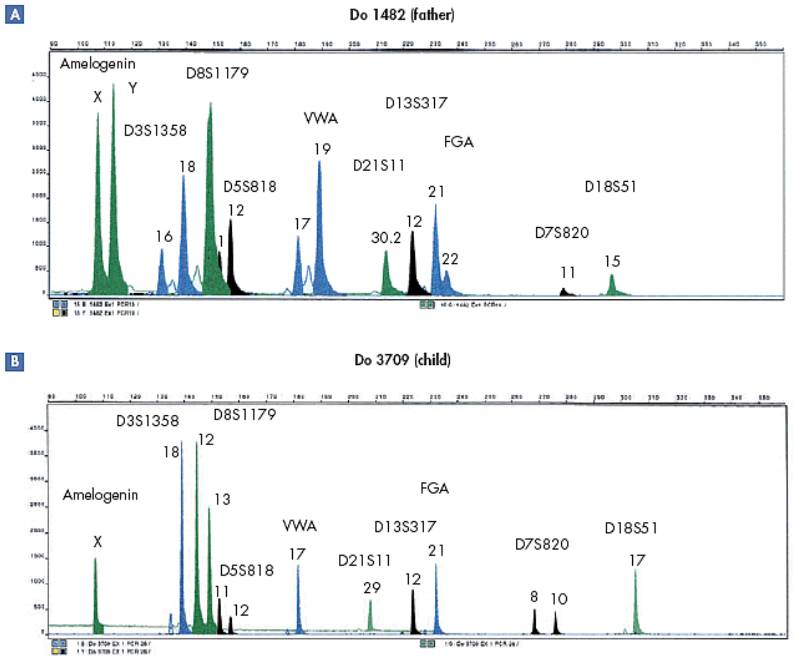

Figure 2.2: Two partial human STR profile on 9 loci belonging to a father and to his daughter.

successive cycle of DNA replication. In theory, a single molecule can be amplified 1 billion-fold by 30 cycles of amplification; in practice, the PCR is not totally efficient but does still produce tens of millions of copies of the target sequence. Anyway, its high sensitivity has a dramatic effect on the types of forensic sample that can be used, making possible to analyse even highly degraded samples successfully. Once these sequences have been amplified, they are separated either through gel or capillary electrophoresis, an analytical technique used to separate DNA fragments by means of an electric field that induces the nucleic acids to migrate toward the anode, exploiting the mobilities with which different sized molecules are able to pass through a viscous medium, the gel, or move in the interior of a small capillary filled with an electrolyte. Finally they can be visualized using bands of different length indicating the variant number of repeats for the considered loci, as in Figure 2.2.

Each STR (there are currently over 10000 published STR sequences in the human genome) is polymorphic, but the number of alleles can be very small for some locus. The power of STR analysis comes from looking at multiple STR loci simultaneously, and this is the reason why a large variety of kits of primers are available to handle forensic identifications (Butler (2006)). They consist in a different number of markers on different loci - typically from 8 to 16 - selected in order to guarantee several features including among other (Goodwin et al. (2007)):

- discrete and distinguishable alleles;
- robust amplification of the locus;
- high power of discrimination, i.e. high polymorphism;
- absence of genetic linkage with other loci being analysed, which means that independence between loci belonging to the same kit can be assumed.

In addiction to the STR loci, some kits (like the one in Figure 2.2) include *amelogenin*, present in the sexual X and Y chromosomes and used for sex determination. STR loci selected to be used in human DNA profiling generally exhibit Hardy-Weinberg expected genotype frequencies and there is evidence that the loci meet the other assumptions of Hardy-Weinberg (more details will be given Section 2.3).

2.2.2 Other kinds of genetic data: Y-DNA, mtDNA, SNPs

Even if in the present work they are not exploited, other kinds of genetic data can be collected and used for forensic purposes: *Y chromosome*, *mitochondrial*

DNA (mtDNA) and *Single-nucleotide polymorphisms* (SNPs). The first two are mainly used as lineage markers; even if their power of discrimination is lower than that of autosomal markers, some features make them valuable forensic tools. The use of SNPs in forensic activity is currently limited to some specialist cases, but may play an increasingly role in the future: their potentiality consists in the large number of SNPs in the human genome, on the other hand their limit is the low degree of polymorphism of SNPs loci.

Y chromosome

The Y chromosome is one of the two sex-determining chromosomes in humans (and in mammals in general). It contains the gene SRY, which triggers testis development if present. DNA in the Y chromosome is passed from father to son unchanged, so Y-DNA analysis is thus used in genealogy research. A Y chromosome contains a large number of polymorphisms including more than 100 known STR markers. The technique used for profiling is the same as for autosomal STRs, i.e. PCR.

The Y chromosome is helpful in a series of forensic cases such as paternity testing and sexual assaults when differential DNA extraction is not possible, and it also simplifies the sorting of material following mass disasters. Finally, due to the widespread practice of partilocality (where the female moves to the male's birth place/residence after marriage), Y chromosome has a non-random distribution among global population, making it a useful tool for inferring the geographical origin of recovered biological material.

mtDNA

Mitochondrial DNA (mtDNA) is the DNA located in organelles called *mitochondria*, it is maternally inherited and it is present in multiple copies. In fact, differently from the nuclear genome that compares with two copies per cell, individual cells can contain hundreds of mitochondria which in turn can contain several copies of the genome.

Mitochondrial DNA shows polymorphic sites concentrated within relatively small regions of the genome and can be analysed using PCR amplification. In particular two main regions belonging to the so *called control* region of the mtDNA sequence represent the focus of most forensic studies concerning identifications, and are known as hypervariable sequence regions I and II (HV-I and HV-II). It is also of interest for forensic purposes that mutation rates shown by mtDNA sequences are generally higher than nuclear genome.

Mitochondrial DNA is a valuable genetic marker in a number of scenarios , and this is mainly due to two properties of mtDNA: the high copy number and the maternal inheritance. The first property is important when the amount of cellular material available for the analysis is small or when the DNA is highly degraded for standard STR typing. The latter characteristic is a useful trait for human identification when there are no direct relatives to use as a reference sample. mtDNA is also a powerful tool for tracking ancestry through females (matrilineage) and has been used in this role to track the ancestry of many species back hundreds of generations.

SNPs

A single-nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide (A, T, C or G) in the genome differs between different individuals. For example, two sequenced DNA fragments, $\boxed{\text{AAGC}\textbf{C}\text{TA}}$ and $\boxed{\text{AAGC}\textbf{T}\text{TA}}$, contain a difference in a single nucleotide which can be exploited for DNA fingerprinting.

As of October 2011, sequencing of human genome identified over 52 million SNPs, usually occurring in non-coding regions on DNA. Due to this vast amount of available data in the different SNPs in the genome, one of the biggest task is to select the most appropriate SNPs from the overwhelming numbers that are available. According to the specific application at hand, usually 50-80 highly polymorphic SNPs are selected for most forensic cases. For more details about the detection methods see Goodwin et al. (2007).

Even if it takes about four times more SNPs that STR loci to obtain the same discrimination power, the major advantage of using SNPs is that using current technology SNP analysis can provide results from highly degraded DNA when conventional STR profiling has failed. Furthermore, SNPs show a lower mutation rate than STR loci.

2.3 Elements of population genetics

In the nineteenth century there were several theories of heredity, including *inheritance of acquired characteristics*, asserting that features employed more frequently cause the trait to become more developed in the offspring, and *blending inheritance*, also advocated by Charles Darwin, which states that offspring display characteristics that are an intermediate combinations of those of the parents.

Then another theory established itself in a decisive way.

2.3.1 Mendelian laws and Hardy-Weinberg equilibrium

From 1856 to 1863, the Austrian friar Gregor Mendel carried out experiments with some 29000 pea plants that demonstrated the basics concepts of *particulate inheritance*. His experiments led him to make two generalizations, the Law of Segregation and the Law of Independent Assortment, which later became known as Mendel's Laws of Inheritance.

Law of Segregation Mendel's first law predict independent segregation of alleles at a single locus, stating that every individual possesses a pair of alleles for any particular trait and that each parent passes a randomly selected allele to its offspring.

Law of Independent Assortment Mendel's second law predicts assortment of multiple loci. This means that separate genes for separate traits are passed independently of one another from parents to child.

Although his paper (Mendel (1866)) was criticized at the time and remained almost unnoticed for nearly 35 years, Mendel's results were later recognised and they are now considered revolutionary. Mendel's rediscovered hypothesis of particulate inheritance was also confirmed by a series of experiments and microscope observations of cell division by the independent works of Walter Sutton and Theodore Boveri, pointing out the connection between the rules of inheritance and the behaviour of the chromosomes, known as *Sutton-Boveri Theory* (Sutton (1902), Sutton (1903), and Boveri (1904)).

Mendel's laws established the very foundation of population genetics, i.e. the study of the frequency and interaction of alleles and genes in populations over time, since the concept of particulate inheritance made possible to perform a wide range of prediction about genotype and allele frequencies. Still progress and insight into this new science was gradual. In fact, for several years it was generally believed that rare alleles would disappear from populations over time. Godfrey H. Hardy (1908) and Wilhelm Weinberg (1908) worked independently to show how genetic variation is maintained in a population showing Mendelian inheritance. The *Hardy-Weinberg equation* (abbreviated in HWE) can be used to predict allele frequencies given genotype frequencies or vice-versa. It's formula is:

$$p^2 + 2pq + q^2 = 1, \tag{2.1}$$

where p and q are allele frequencies for a locus with two alleles. If a generic number of alleles, n, are possible in a locus, HWE gives the probabilities of heterozygous and homozygous genotypes:

$$Pr(i,j) = \begin{cases} 2p_i p_j, & \text{if } i \neq j, \\ p_i^2, & \text{if } i = j, \end{cases} \tag{2.2}$$

with $i, j \leq n$.

To hold, (2.1) and (2.2) require some conditions to be met:

- the population is infinitely large;
- random mating occurs within population;
- absence of disturbing influence of selection, mutation, migration and genetic drift (some of this terms will be explained up ahead in this chapter).

If this set of conditions can be reasonably assumed, then (2.2) implies that both allele and genotype frequencies in a population remain constant (i.e. they are in equilibrium) from generation to generation, and that if one of the two is known it is possible to calculate the other one. It is important to understand that one or more of the previously named disturbing influences are always in effect, so the Hardy-Weinberg equilibrium is impossible in nature, but it is an ideal state that provides a baseline against which to measure changes.

2.3.2 Deviation from the equilibrium

The conditions for Hardy-Weinberg equilibrium listed in the previous Section are violated in any realistic human population. This means that there are a number of factors that can change allele proportions as they are defined in (2.2). These are referred to as *disturbing forces* (e.g. Hamilton (2009)). In this paragraph we take a closer look to some of them.

Infinitely large population

Obviously this assumption is violated to greater or lesser extents: this depends on the size of the population of interest, but no population on earth can be infinite. The consequence of finite population size is that the frequency of alleles will change due to a process known as random *genetic drift*, where the frequency of any given allele will increase or decrease through chance events. This is because the alleles in the offspring are a sample of those in the parents, and chance has a role in determining whether a given individual survives and reproduces. The result is that genetic drift may cause gene variants with few allele copies to disappear completely and thereby reduce genetic variation.

The effect of genetic drift is more pronounced in smaller populations, however, most populations are sufficiently large for allele frequencies not to be significantly affected. Even in relatively small and isolated populations, it has been shown that alleles that are present at a frequency of more than 1% are rarely lost.

Random mating

Humans clearly do not mate completely randomly. However, because STR genotypes do not have any physical manifestation, such as height, strength or intelligence, direct selection of an STR through sexual selection is unlikely and has not been demonstrated. Nevertheless it would be wrong to assume from this that random mating is a fair assumption.

In reality a population is often composed of various sub-populations. This could be caused by geographic proximity, or there may be social reasons, where people of different ethnic origins will tend to reproduce within their own ethnic grouping, or linguistic. These sub-populations are not totally isolated from each other, obviously, but still there is a departure from completely random choice of mates, since mating between people belonging to the same sub-population appears more likely. This fact is known as *inbreeding*.

Main consequence of the violation of the random mating assumption is that, within each sub-population, there is a non-negligible probability F of two alleles drawn from the same sub-population being identical by descent (IBD), i.e. arisen from the same allele in an earlier generation. Thus the probabilities of individuals from a sub-population receiving alleles i and j given the population relative frequencies p_i and p_j with a inbreeding coefficient F are

$$Pr(i,j) = \begin{cases} 2p_ip_j(1-F), & \text{if } i \neq j, \\ p_i^2 + p_i(1-p_i)F, & \text{if } i = j, \end{cases} \tag{2.3}$$

so that the presence of co-ancestry increases the probability of observing homozygousity, modifying (2.2).

No migration

Human history is full of migrations and this obviously can lead to changes in the gene pools of populations. The effect of migration on equilibrium depends on the difference in allele frequencies between the donor and recipient populations.

If, by effect of consistent migration, two distinct populations are living in the same geographical area and they have different allele frequencies, each population can be in Hardy-Weinberg equilibrium. If the two different populations are not recognized within the larger population and are not treated as separate populations, deviation from the Hardy-Weinberg equilibrium may be apparent. This is known as the *Wahlund effect*, i.e. the reduction of heterozygosity in a population caused by sub-population structure. If random admixture occurs between the two populations, the admixed population would be in Hardy-Weinberg equilibrium after one generation. In reality, where two populations have differences in language, culture or religion, admixture is normally a much longer process.

Natural selection

Natural selection is a key mechanism of evolution: the gradual, non-random process by which biological traits become either more or less common in a population as a function of the advantaging effect on their bearers. At some loci in the human genome the effect of selective pressures can be detected, for example lactase persistence that is present in populations where milk has been a sustained part of the diet. Mutations that can confer disease resistance can also exhibit strong selection effects. However, the loci that are used for forensic testing are not located within functionally important regions of the genome, since these loci are non coding. In summary, even if some unknown indirect mechanism for selection could non be excluded (for example selection by association with disease loci may possibly affect STR loci), there are enough theoretical reasons to believe that STR loci are selectively neutral or nearly so.

Mutations

The assumption of no mutations is clearly violated, and this is true especially for STR loci, that are natural mutational "hot spots", with mutation rates above much of the coding DNA. Mutation is, in fact, the main source of genetic variation and this is particularly helpful in the field of personal identification, being in fact one of the reasons why STR loci are often very polymorphic: the consequence is that such loci can be fruitfully used as informative markers for forensic purposes.

As a disturbing force for HW equilibrium, the effect of mutation in STR loci on a divided population is that it tends to oppose the effect of genetic drift. If drift tends to remove genetic variation from separated sub-populations, mutation tends to reintroduce it. However, the mutation rates of STRs are still relatively low and do not have a significant effect on the allelic frequencies within a gene pool of different or even mixed populations (Buckleton et al. (2005)).

Anyway, it is important to consider the possibility of a mutation to occur when dealing with kinship identification. To do this it is helpful to construct some simplifying models of the mutation process itself. Mutation models attempt to capture the essence of the genetic changes caused by mutation while at the same time simplifying the process of mutation into a form that permits generalizations about allele frequency changes. There is no single model of the process of mutation, but rather a series of models that serve to encapsulate different features of the mutation process for different classes of loci and different types of alleles. In Section 4.3 we will give a summary of the mutation models adopted in this work.

2.4 Summary

In this Chapter, some very fundamental concepts about genetics have been introduced. The initial part is devoted to describe how DNA is structured and its role as carrier of heritable information through successive generations. The human genome is articulated in 46 chromosomes and a unit of genetic information is represented by a section of it, called gene. Although, only a very small part of the DNA sequence is of value when we want to use DNA as unique, distinctive, characteristic of an individual. Among these non-coding DNA regions, STR loci are the most commonly used type of genes used for personal identification purposes, thanks to the high polymorphism, robustness and absence of dependence between each other (when appropriately selected). Other kinds of genetic evidence can be used, but their adoption is more limited than that of STRs.

Mendelian laws and Hardy-Weinberg equilibrium ensure approximated, although useful, models to describe how heritable traits are transmitted from the parents to the offspring and why the frequency and interaction of alleles and genes in populations does not change over time, under certain assumptions. Of course these very baseline models need to be refined if we want to account for various violations to the mentioned assumptions, for example to consider mutations, which can, in reality, occur with a non-negligible probability.

These models and their refinements will be considered in some of the subsequent chapters, since they play an important role in the individual identification methodologies.

CHAPTER 3

Introduction to Bayesian networks

A Bayesian network is a probabilistic graphical model that represents the joint distribution of a set of random variables and their conditional dependencies via a Directed Acyclic Graph (DAG). More formally a DAG considers a *graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ composed by sets of nodes (or vertices), $\mathcal{V}$, and edges, $\mathcal{E}$, where the number of vertices n_v is finite and the number of the edges is smaller than $n_v \times n_v$. If $X = (X_{v:v\in\mathcal{V}})$ represents the vector of random variables indexed by $\mathcal{V}$, edges linking different nodes must be *directed*: if for example there is a directed edge from node X_1 to X_2, then X_1 is said to be a *parent* of X_2, or similarly X_2 is a *child* of X_1. More generally, when a directed path (of length greater than 2) from a node X_1 to X_3 exists, then X_1 is an *ancestor* of X_3 and X_3 is a *descendant* of X_1. Finally, as suggested by the word *acyclic*, a DAG must not present any cycle, meaning that there is no way to start at some vertex X_v and, by following a sequence of directed edges, to loop back to X_v again. Figure 3.1 gives an example of a simple DAG including five variables.

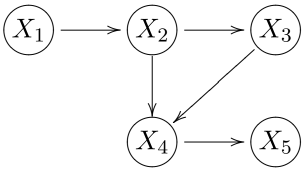

Figure 3.1: Example of a simple DAG with five nodes.

Each node is associated with a *conditional probability table* (CPT) that takes as input a particular set of values for the node's parent variables. For each parental configuration a CPT gives the probability of the variable represented by the incident node. States of each node comprise sets of mutually exclusive and exhaustive values and the probabilities for each node sum to one.

Bayesian networks are thus directed acyclic graphs whose nodes represent random variables: they may be observable quantities, latent variables, unknown parameters or hypotheses. Oriented edges represent (direct) dependencies; nodes which are not connected to each others imply some form of conditional independence. The exploitation of this lack of edges between nodes, and thus of the conditional independence among them, is a fundamental feature of a Bayesian network, making possible to obtain the main result a DAG is devoted to: the full representation of the joint probability distribution of the variables in $\mathcal{V}$, factorizing it by the product of conditional distributions, taking advantage of conditional independence relationships to simplify the probabilistic problem specification and the associated computational complexity.

Not all kinds of probabilistic models can be expressed by means of a Bayesian network (as we saw, the variables' underlying model must determine an "order" among the variables themselves, i.e. some variables hierarchically imply others, and the final network structure must be acyclic), but if this is possible, then a fundamental property of a Bayesian network is the possibility to express its joint probability function as a product of the individual probability functions, conditional on their parent variables solely:

$$Pr(x) = \prod_{v \in \mathcal{V}} Pr\big(x_v \mid x_{\mathrm{pa}(v)}\big) \tag{3.1}$$

where pa(v) is the set of parents of v. For example, the joint probability distribution of the Bayesian network displayed in Figure 3.1 can be expressed as follows:

$$Pr(X_1, \ldots, X_5) = Pr(X_5|X_4)Pr(X_4|X_2, X_3)Pr(X_3|X_2)Pr(X_2|X_1)Pr(X_1).$$

Furthermore, a Bayesian network satisfies the local Markov property, resulting that each variable is conditionally independent of its non-descendants given its parent variables:

$$X_v \perp\!\!\!\perp X_{\mathcal{V} \setminus \mathrm{de}(v)} \mid X_{\mathrm{pa}(v)} \quad \text{for all } v \in \mathcal{V}$$

where de(v) is the set of descendants of v. Again, referring to the DAG in Figure 3.1 we can affirm, for example, that $X_5 \perp\!\!\!\perp \{X_1, X_2, X_3\}|X_4$.

As a final remark, it must be noted that Bayesian networks have no direct reference to Bayesian inference, but the word *Bayesian* is referred to the non-elementary use of Bayes' theorem for the computation of conditional probabilities, to obtain the same result of a naive application of the Bayes' theorem but in a much more efficient way, once the parameters ruling their conditional distribution are known.

Learning

Two kinds of learning, depending on the fact that it is referred to the variables' CPTs parameters or to the edges' structure, can be pursued for the set up of a Bayesian network: they are respectively called *parameter* and *structure* learning.

CPTs' parameters are not always knows when setting up a Bayesian network: if this happens, it is possible to learn them by means of different strategies. Parameters which are unknown must be estimated from data, sometimes using the maximum likelihood approach. Direct maximization of the likelihood is often complex when there are missing data. A classical approach to this problem is the expectation-maximization (EM) algorithm which alternates computing expected values of the unobserved variables conditional on observed data, with maximizing the complete likelihood. Under mild regularity conditions this process converges on maximum likelihood values for parameters.

Automatically learning the graph structure of a Bayesian network is a challenge pursued within machine learning. A popular methodology is based on an algorithm called PC algorithm, introduced by Spirtes et al. (1993), that is based on multiple independence tests on triplets of variables. Exploiting the found conditional independence relationships one can determine the skeleton of the underlying graph for each triplet of variable and, then, orient all arrows whose directionality is dictated by the conditional independences observed. For example, let $\{X_1, X_2, X_3\}$ be a generic triplet in which $X_1 \perp\!\!\!\perp X_2$ and all other pairs are dependent, also conditionally on the third variable (e.g. $X_1 \not\!\perp\!\!\!\perp X_2|X_3$): this set of relationships uniquely defines the graphical structure in Figure 3.2.

Alternative methods of structural learning are based on scoring methods, needing a scoring function (e.g. a common one is posterior probability of the structure

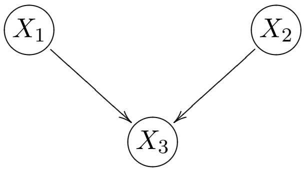

Figure 3.2: Example of structure learning when $X_1 \perp\!\!\!\perp X_2$ and $X_1 \not\perp\!\!\!\perp X_2|X_3$.

given some training data) and a search strategy, aimed to maximize the score (for example search algorithms based on Markov Chain Monte Carlo (MCMC) can be used). Simulation and approximation methods are required since the time for an exhaustive search rapidly increases with the number of variables.

If parameter and/or structural learning are not the issue of the analysis, since CPTs and arrows structure are known (or are assumed known) from established models, the main goal of a BN is to perform probability propagation to obtain the joint and the marginal conditional probabilities on the variables of interest. In this case it is common to refer to a Bayesian network as Probabilistic Expert System (PES). This is, generally, the case of interest for forensic applications and the reason why a large number of contributions in the field refers to PESs instead of Bayesian networks.

Software

The most important advantage in representing a complex probabilistic problem by a graphical model, is to make use of efficient computational algorithms (Pearl (1988), Lauritzen and Spiegelhalter (1988)) to perform calculations that would be otherwise nearly intractable since the high-dimensional nature of the problem.

Several pieces of software have been created to facilitate the network building and to compute the required conditional probability. Cowel et al. (1999) give more insights about Bayesian networks and PESs in a more general framework. These include for example HUGIN and NETICA$^{\text{TM}}$, which are oriented to practitioners since their friendly Graphical User Interface (GUI), but otherwise it is possible to develop Bayesian network algorithms within general purpose statistical programming languages: among others, Grappa is a suite of functions in R for probability propagation in discrete graphical models mainly developed by Peter Green[1], while the Bayes Net Toolbox by Murphy (2001) is implemented in MATLAB®. At an intermediate-level programming language, also C++ libraries for Bayesian networks are available.

In the field of personal identification it is possible to adopt one of the mentioned pieces of software or to use an *ad hoc* one, the package FINEX (Cowel (2003)), specifically designed to handle this kind of problems.

3.1 BNs for forensic analysis

Bayesian networks are able to provide a very useful representation of the identification issues involving the use of DNA evidence. This is of interest for criminal cases, when a DNA trace of unknown origin is compared with that of a suspect, but it is more relevant in case of indirect identification since kinship relationships among relatives are easily conveyed by a Directed Acyclic Graph. There, the inheritable DNA characteristics of an individual probabilistically affect those of their unobservable relatives.

More specifically, the set of kinship relationships among a group of individuals forms a *pedigree*, that smoothly fits into a PES. In Figure 3.3 the pedigree of a family formed of a mother (M) a father (F) and a child (C) is displayed, adopting the usual convention that males are included into squares and females into circles.

[1] http://www.stats.bris.ac.uk/ mapjg/Grappa/

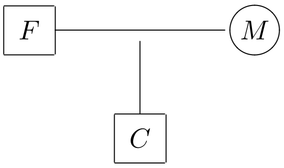

Figure 3.3: A simple pedigree of a family with a mother (M) a father (F) and their child (C).

There are several ways to design a Bayesian network starting from a pedigree and the seminal work of Lauritzen and Sheehan (2003) in this field gives an overview of different methodologies, introducing three kinds of networks.

- The *segregation* network gives the most complete representation of the inheritance relationships in a pedigree. Two nodes for each individual represent the paternal and maternal inherited genes. Then, for each non-founding node (founders are those nodes that have no parents pointing to them), one additional variable represents the segregation indicator, conventionally taking value 1 to denote that the paternal allele has been inherited and 0 to indicate inheritance from maternal gene, according to specified inheritance rules (the mendelian one as in Section 2.3 or others).

- The *allele* network can be convenient since complete information about the segregation mechanism is often unnecessary (or unavailable). This is obtained from the previous one by removing the segregation indicators and associated edges. Convenient modifications to the CPTs ensure the correct representation of the segregation laws relevant for the case at hand.

- The *genotype* network, as in Figure 3.4, is visually the most parsimonious but not the most useful: nodes represent genotypes, thus the state space for each node can be huge with respect to those of nodes representing alleles. For systems with k possible allelic values, the genotype can assume $k(k+1)/2$ different states. This is of substantial importance for computational issues.

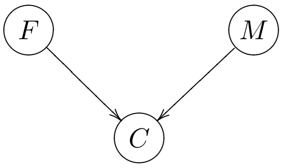

Figure 3.4: Genotype network for the pedigree in Figure 3.3.

Also Dawid et al. (2002) explores the applications of reconstructing relevant pedigrees ad Bayesian networks when approaching to identification problems, according to some simplifying assumptions that will be released in the following Sections. They make use of a slightly modified version of the allele network introduced earlier, enhancing it with nodes expressing the genotype of each individual, deterministically defined by the relevant couple of alleles (i.e. the paternal and maternal inherited alleles fully define the child's genotype). Figure 3.5 is an example of the use of this kind of modified allele network, showing the Bayesian network associated to Figure 3.3. There the maternal and paternal genotypes (respectively mgt and fgt) are given by their paternal and maternal inherited alleles, i.e. the mother maternal and paternal allele (mma and mpa) and the father maternal and paternal allele (fma and fpa). Then the child inherits from the mother and the father the two alleles (cma and cpa) forming his genotype (cgt). For the purposes of this thesis we will mainly make use of this latter kind of network, eventually further modified, for the analysis and the calculations, but we will also rely on some genotype networks for displaying purposes only.

Since the independence between loci, as explained in Section 2.2, it is possible to consider a separated and specific PES for each locus, and then obtain the overall

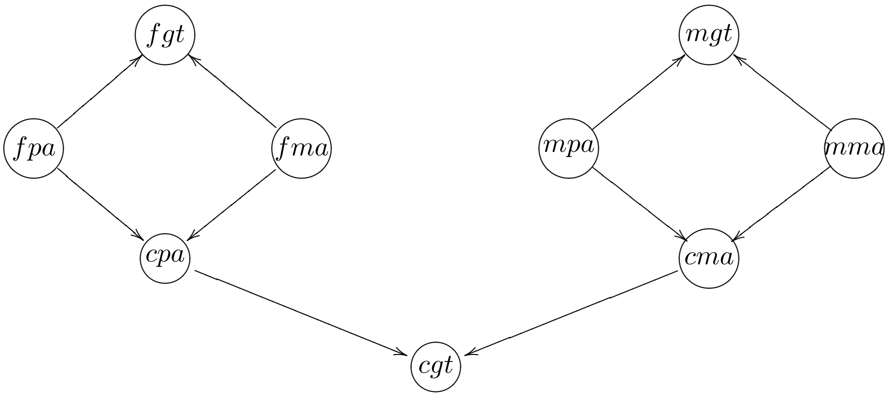

Figure 3.5: Bayesian network representing the pedigree in Figure 3.3.

result as a combination of the results from each network.

Among others, an important feature that helps Bayesian networks to solve problems based on DNA evidence is what is called *allele recoding*, which ensures that if only a subset of the possible allelic types is represented in the set of observations, it is possible to take advantage of the merging of all the unobserved alleles into a residual one labelled, say, *other*. This reduces the state space, and thus the computational complexity of any multi-allelic system without loss of information.

The importance of Bayesian networks goes beyond the analysis of genetic evidence, since they represent a valuable tool for forensic applications based on other kinds of evidences, different from the genetic one, e.g textile fibres or footprints. Bayesian networks have been also proposed for structuring and reasoning about issues of complex cases in judicial contexts (see, for example, Taroni et al. (2006), Section 2.3.2), or to interpret combined different items of evidence (Dawid and Evett (1997)). A further advantage is that Bayesian networks can be constructed for single pieces of evidence; then these networks can be easily combined together to produce a larger network which combines evidence from different sources from a case to allow all information to be considered in a natural, but probabilistically rigorous, way.

3.1.1 Bayesian networks for complex genetic problems

One of the most important advantages given by the use of Bayesian networks to solve forensic problems stands in the flexibility of this methodology. This practically translates in the possibility to increase the complexity of a networks just adding the appropriate nodes and edges, without altering the existing network structure.

Mutations

The first, and perhaps one of the most relevant complicating feature affecting an identification based on DNA data is the possibility of a mutation (or more) to occur. In Section 2.3.2 we introduced the genetic aspects of mutation, now we give some detail about their treatment using a Bayesian network. Dawid et al. (2001) and Vicard and Dawid (2004) give a summary of various models that can be easily translated into a BN framework:

- the *uniform* mutation model simply gives a fixed value for the probability of a mutation towards any other allele;
- for the *proportional* mutation model the probability of a mutation to allele j is proportional to the frequency p_j in the reference population;
- the *one-step* mutation model assumes that any mutation can only be to a neighbouring allele value;
- the *mixed* mutation model is a mixture the one-step and the proportional models with respective weights h and $1 - h$ (for some fixed h between 0 and 1), so combining features of both.

All these models have pros and cons, but the mixed mutation model seems the most biologically plausible. Furthermore it implies as special cases the single-step and the proportional models, which can be obtained just setting respectively to 1 or 0 the parameter h. In Section 4.3 we will present the key features of various models, discuss their mathematical and biological aspects and we will detail the application of the mixed mutation model in our evaluation methodology.

More complex features involving founding nodes

There are a number of complicating features, such as uncertainty in allele frequencies (UAF), coancestry, identity by descent (IBD) and mixed populations, sharing a common characteristic: all of them involve the violation of assumption about founding nodes in a Bayesian network.

In Green and Mortera (2009) these features are considered, some of which are used in our research, and they provide solutions on how to handle them in a Bayesian network context. The issue that is relevant here is the one involving uncertainty in allele frequencies. In fact a realistic model should relax the assumption of known allele probabilities, and considers them uncertain, since they are estimated from a sample of the population of interest. This is done by setting up a Dirichlet process model, that can be expressed as a Pólya urn scheme, which is amenable to representation as a Bayesian network and intuitively means that it is reasonable to update the initial allele frequencies derived from the sample by increasing the probability of the alleles that have been observed to belong the individuals involved in the analysis.

Finally, we conclude this Section accounting for another complication arising when considering sub-populations. Coancestry can be taken into account considering that they had a small number of founders so that the probability F an individual's receiving two copies of the same allele transmitted by the same ancestor is not negligible. The solution of this, as already presented in Section 2.3.2, is to consider (2.3) instead of (2.2), so that the presence of co-ancestry increases the probability of observing homozygousity. This feature can be introduced into a BN just coherently modifying the CPTs of the nodes representing the genotypes of the individuals.

Object Oriented Bayesian Networks

In order to include all the previously introduced refinements in the structure of a PES, one can either augment the Bayesian network adopted for a particular application, or rely on an extension of Bayesian network technology called *object-oriented Bayesian Networks* (OOBN, Koller and Pfeffer (1997) and for their application for personal identification Dawid et al. (2007)), allowing hierarchical definition and construction of a BN, using simple modular building blocks. Additional complexity can be introduced by adding new modules or refining the existing ones. HUGIN is one of the most important reference software which gives the opportunity to easily create and work with object-oriented Bayesian Networks.

Although, in this work, we do not make use of the recursive computation implied by the strict applications of OOBNs, they represent a non-negligible tool to cope with complicating features, and the algorithms we built for our methodology to evaluate an identification system take into account concepts originally developed for the OOBN framework.

3.2 Summary

In this Chapter we have introduced Bayesian networks as valuable tools to cope with and solve complicated probabilistic problems efficiently exploiting conditional independence relationships among the variables involved. We defined a Bayesian network as a set formed of vertexes and edges, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which the very fundamental property is given by equation (3.1).

A large number of publications have been issued addressing the topic of the application of Bayesian networks' methodology to solve forensic problems. We saw that a pedigree easily lend itself to be translated into a Bayesian network, relying on several methods detailing at different levels the biological relations among individuals which ensure computational efficiency and tractability.

Finally we introduced some more refined models involving features, as mutations and coancestry, that are of interest for this work, accounting for some possible ways in which these refinements can be included into a Bayesian network. How this is actually performed in this thesis will be detailed in the following of the thesis.

CHAPTER 4

Evaluation of kinship identification systems

This Chapter presents the core of the researches we conducted in the last years: some proposals to evaluate kinship identification systems. The proposed system evaluation is case-specific, does not require any additional laboratory costs and should be carried out before an identification trial is performed. Under this pre-experimental perspective, we evaluate whether a system fulfils the requirements of the parties involved.

4.1 Kinship identification

The application of DNA profiling to kinship analysis is widespread and aims to provide support to some biological relationships. Since the first DNA-based kinship test in 1985, detailed by Jeffreys et al. (1985), DNA analysis has been applied to an increasing number of kinship tests: paternity testing is, by far, the most common form of kinship testing, with hundreds of thousands of tests being performed worldwide each year, but also more complex relationships are commonly investigated. Kinship investigation features prominently in forensic science within both criminal and civil jurisdictions.

In a criminal context, such testing can be required following sexual assaults to identify the father of a child conceived as a result of an alleged assault. In cases involving concealed births, abandoned children, or infanticide, it may be necessary to prove a genetic relationship to either ensure the rightful return of an infant or to support criminal charges.

During civil trials, indirect identification of an alleged father can be required to substantiate claims for financial support and maintenance of a child. Similarly disputes over inheritances can benefit by the application of genetic testing. Kinship analysis is also now being widely applied by governmental bodies to adjudicate in cases of immigration and naturalization. The identification of bodies for legal purposes can also be effected using familial testing.

4.1.1 Notation and mathematical aspects

In this Section we introduce the notation and the terminology we will use throughout the thesis. In kinship identifications we consider a family asking for the identification, as one of its members, of an individual, the *candidate* (C). Two hypotheses about the actual state of the kinship relationship are usually taken into account: conventionally H_1 is the hypothesis claimed by the family and assumes that C is the person posed in the familial pedigree in a defined position (U); H_0, the alternative, implies C being either another relative (U') or an individual not recently related to the family, i.e. C is a generic member of a population. Let the set $\mathcal{F} = \{\mathcal{F}^+, \mathcal{F}^-, U, U'\}$ contain all the family's members involved in the analysis: $\mathcal{F}^+$ is the set of relatives providing their DNA profiles, while $\mathcal{F}^-$ considers the unobserved relatives required to link the members in $\mathcal{F}^+$ to U or U'.

In Figure 4.1 we represent using genotype networks (cfr. Section 3.1), the pedigrees induced by the hypotheses in a specific kinship problem. There, under hypothesis H_1, S_1 claims C to be his full sibling, occupying the position U in the familial pedigree; alternatively H_0 reckons S_1 as the half sibling of C, i.e. the individual U' in the graph representation. H_0 modifies the familial pedigree originally induced by H_1, adding to $\mathcal{F}^-$ an additional (unobserved) individual, F_2, and establishing different relations. In this example the set $\mathcal{F}^+$ includes only one member of the family under both hypotheses: S_1. For pictorial purposes only, hereafter solid lined nodes indicate observed variables, while dotted lines are for the unobserved ones.

Please note that we maintain distinct variables to probabilistically represent the positions in the pedigrees (U and U') which are occupied by the candidate (C) under either of the two hypotheses, H_1 and H_0. In fact, conditionally on U (or U'), the genotype of C (represented by the variable X_C) is deterministically implied under both hypotheses: the variable X represents the genotype of an individual, hence we define a CPT for $X_C|X_U$ (or $X_C|X_{U'}$) assigning probability 1 to the events that $C \equiv U$ or $C \equiv U'$.

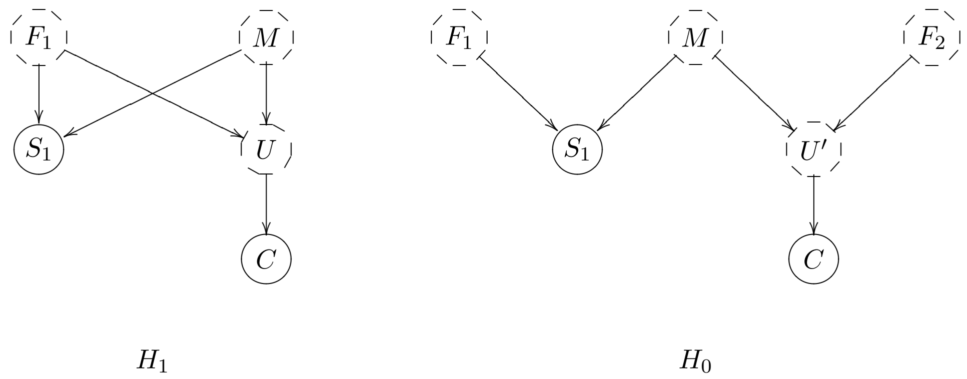

Figure 4.1: Graphical representation of a kinship problem. Solid lines indicates observations, dashed lines are unobserved individuals.

4.2 LR computations based on STR DNA evidence

The solution of a kinship identification problem is achieved by the evaluation of the probability of the observed evidence for people in $\mathcal{F}^+$ and for C conditional on two competing hypotheses, synthesized by the Likelihood Ratio (LR). In terms of a criminal case, involving either direct or indirect identification, it is common to refer to these hypotheses as the *prosecution* hypothesis (H_p) and the *defence* hypothesis (H_d). The likelihood approach is a logical way to interpret and present the DNA profile information, since it considers an alternative scenario.

Since both criminal and civil cases (in these latter two parts are in opposition without prosecution and defence roles) are involved in the field of indirect identification based on DNA data, we decided to generically label the two hypotheses as H_0 and H_1, as already detailed in Section 4.1.1. Considering the relative support to the hypothesis H_1, against H_0, provided by the entire observed genetic evidence, generically indicated by $\mathcal{E}$, we define the likelihood ratio supporting H_1 as

$$LR = \frac{Pr(\mathcal{E}|H_1)}{Pr(\mathcal{E}|H_0)}. \tag{4.1}$$

In a kinship problem, let the random variable X_I represent the probability distribution for the genotype of the generic individual I and let $X_{\mathcal{S}}$ represent the joint distribution for the genotypes of the set of individuals $\mathcal{S}$: using the notation introduced in Section 4.1.1, generic evidence $\mathcal{E}$ is replaced by the available evidence for the case, i.e. x_C and $x_{\mathcal{F}^+}$, and (4.1) becomes

$$LR = \frac{\Pr(x_C, x_{\mathcal{F}^+}|H_1)}{Pr(x_C, x_{\mathcal{F}^+}|H_0)}. \tag{4.2}$$

Once the candidate C and the donors in $\mathcal{F}^+$ have been typed, the required LR in (4.2) can be easily computed since the following assertions of conditional independence hold:

- States of H only affect the probability of observing x_C, i.e., $X_{\mathcal{F}^+} \perp\!\!\!\perp H_z$ with $z \in \{0, 1\}$, so that:

  $$\Pr(x_{\mathcal{F}^+}|H_1) = \Pr(x_{\mathcal{F}^+}|H_0),$$

  meaning that the kinship relationships between people in $\mathcal{F}^+$ are know without uncertainty (or they are, at least, not affected by the hypotheses under debate).

- If H_1 holds, $C \equiv U$, so that:

  $$Pr(x_C|x_{\mathcal{F}^+}, x_U, H_1) = \begin{cases} 1, & \text{if } x_C \equiv x_U, \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

If, otherwise, H_0 is assumed, then $C \equiv U'$, so that:

$$Pr(x_C|x_{\mathcal{F}^+}, x_{U'}, H_0) = \begin{cases} 1, & \text{if } x_C \equiv x_{U'}, \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

Considering formulas (4.3) and (4.4) we have:

$$\begin{aligned} & LR(X_C = x_C) = \\ & \quad = \frac{\Pr(x_C, x_{\mathcal{F}^+}|H_1)}{\Pr(x_C, x_{\mathcal{F}^+}|H_0)} = \frac{\Pr(x_C|x_{\mathcal{F}^+}, H_1)\Pr(x_{\mathcal{F}^+}|H_1)}{\Pr(x_C|x_{\mathcal{F}^+}, H_0)\Pr(x_{\mathcal{F}^+}|H_0)} \\ & \quad = \frac{\displaystyle\sum_{x_{\mathcal{F}^-}, x_U \in \mathcal{X}} \Pr(x_C|x_{\mathcal{F}^+}, x_{\mathcal{F}^-}, x_U, H_1)\Pr(x_U|x_{\mathcal{F}^+}, x_{\mathcal{F}^-}, H_1)\Pr(x_{\mathcal{F}^-}|x_{\mathcal{F}^+}H_1)}{\displaystyle\sum_{x_{\mathcal{F}^-}, x_{U'} \in \mathcal{X}} \Pr(x_C|x_{\mathcal{F}^+}, x_{\mathcal{F}^-}, x_{U'}, H_0)\Pr(x_{U'}|x_{\mathcal{F}^+}, x_{\mathcal{F}^-}, H_0)\Pr(x_{\mathcal{F}^-}|x_{\mathcal{F}^+}H_0)} \\ & \quad = \frac{\displaystyle\sum_{x_U \in \mathcal{X}} \Pr(x_C|x_{\mathcal{F}^+}, x_U, H_1)\Pr(x_U|x_{\mathcal{F}^+}, H_1)}{\displaystyle\sum_{x_{U'} \in \mathcal{X}} \Pr(x_C|x_{\mathcal{F}^+}, x_{U'}, H_0)\Pr(x_{U'}|x_{\mathcal{F}^+}, H_0)} \\ & \quad = \frac{\Pr(x_C \equiv x_U|x_{\mathcal{F}^+}, H_1)}{\Pr(x_C \equiv x_{U'}|x_{\mathcal{F}^+}, H_0)}. \end{aligned} \tag{4.5}$$

If H_0 assumes C as a generic member of the reference population, then $X_C \perp\!\!\!\perp X_{\mathcal{F}}|H_0$ so that the denominator of (4.5) simply becomes $Pr(x_C|H_0)$. As a result, in kinship identifications based on STR loci, the LR is evaluated by assessing the probability of $X_C = x_C$, conditionally on two different states of information.

We have seen that computing the LR does not require the assessment of the prior probabilities for the hypotheses, which simply appear as conditioning circumstances. If, instead, H is considered a random variable, with $\mathcal{H} = \{H_z : z \in \{0, 1\}\}$, and $Pr(H_0)$ and $Pr(H_1)$ are available, we can use Bayes' Theorem to update prior odds to posterior odds using

$$\text{posterior odds} = LR \times \text{prior odds}$$

which here becomes

$$\frac{Pr(H_1|x_C, x_{\mathcal{F}^+})}{Pr(H_0|x_C, x_{\mathcal{F}^+})} = \frac{\mathrm{Pr}(x_C, x_{\mathcal{F}^+}|H_1)}{Pr(x_C, x_{\mathcal{F}^+}|H_0)} \times \frac{Pr(H_1)}{Pr(H_0)}. \tag{4.6}$$

Finally we can use the LR to easily derive interpretable posterior probabilities. In fact starting from (4.6) and considering that $Pr(H_0|x_C, x_{\mathcal{F}^+}) = 1 - Pr(H_1|x_C, x_{\mathcal{F}^+})$, one can write:

$$\begin{aligned} Pr(H_1|x_C, x_{\mathcal{F}^+}) &= LR\frac{Pr(H_1)}{Pr(H_0)}\big(1 - Pr(H_1|x_C, x_{\mathcal{F}^+})\big) \\ &= LR\frac{Pr(H_1)}{Pr(H_0)} - LR\frac{Pr(H_1)}{Pr(H_0)}Pr(H_1|x_C, x_{\mathcal{F}^+}) \\ &= LR\frac{Pr(H_1)}{Pr(H_0)}\Big\{1 + LR\frac{Pr(H_1)}{Pr(H_0)}\Big\}^{-1}. \end{aligned}$$

On the other hand, we can compute the LR required to update a given prior to a specified posterior:

$$LR = \frac{Pr(H_1|\mathcal{E})}{Pr(H_0|\mathcal{E})} \times \frac{Pr(H_0)}{Pr(H_1)}. \tag{4.7}$$

We conclude this section with a remark on the choice of the prior probabilities for the hypotheses, $Pr(H_0)$ and $Pr(H_1)$. Depending on the case, every prior probability on H is acceptable and, usually, prior probabilities equal to 0.5 for both are considered to represent a non-informative state of knowledge. It may be appropriate in some cases, but equally may be totally inappropriate in others. However, for practitioners, it has become customary to assign prior probabilities of 50% to both H_0 and H_1. This assumption is hard to justify at the fundamental level (Good (2001)) and must be seen simply as a pragmatic choice. For example, in a paternity case in which two alleged fathers are questioned to be the real one this assumption could be reasonable, but if also non-genetic evidence were available, prior probabilities should be modified accordingly: this evidence could in fact include statements of the mother as to with whom she had intercourse, or evidence that may suggest that the alleged father was out of the Country or in prison at the time of conception. Such evidence, if relevant and admissible, affects the prior odds.

4.3 Models and data

In this Section some of the topics introduced in Sections 2.3 and 3.1 will be discussed in deeper details. The main focus is on the models we adopt in this work to take into account features, such as mutations or coancestry, relevant for a kinship identification system based on DNA evidence. One fundamental distinction among these models is the following: the probability distribution for the genotype of an individual (X_I) is provided by a *segregation* model if their parents are explicitly included in the pedigree or by a *population* model if the individual I is a founder. Some of these models are considered below.

For the present thesis' purposes, the segregation model has to define if (and, in case, how) mutations must be handled; while the population model serves to refine the standard setting of assuming the allele frequencies as known, considering them uncertain instead, and acknowledging for the existence of sub-populations.

4.3.1 Segregation models

Mendel's law (abbreviated, ML) is the baseline segregation model, stating that each parent passes at random one of their allele to the offspring, not considering any mutation mechanism. A more realistic approach allows for mutations in the segregation process. Various models have been proposed to deal with the phenomenon of mutation and, as already acknowledged, Vicard and Dawid (2004) discuss a number of such models. Here we consider three of them: the *proportional*, the *one-step* and the *mixed* mutation model.

These three models share some common elements that we introduce here, before analysing them singularly. First of all, all the models are based on mutation rates which are, usually, estimated over a large number of cases of complete trios (mother, father and child), for which parental relations have been already ascertained. In this work we will make use of mutation rates estimated by the American Association of Blood Banks in one of the latest annual report (AABB (2008)), correcting them to take account for hidden mutations (Chakraborty et al. (1996),Vicard and Dawid (2004), Brenner (2004)), i.e. the fact that mutations do not always lead to a genetic inconsistency. Details on how this correction is performed are given in Chapter 7.

Secondly, since the independence between loci, we do not need to treat several loci simultaneously, so we will omit the notation referring to a specific locus. Hence we will assume that the allelic ladder $\mathcal{L}$, i.e. the set of possible allelic values, is known for every locus, and we will consider plausible only mutation from and to an already existing allele in the ladder, not allowing for the creation of new off-ladder alleles.

Among the proposed models, only one, the proportional mutation model, is *stationary*, meaning that population allele frequencies are not altered by the mutation process (Dawid et al. (2001)). Since it is not clear whether or not allele frequencies are in reality stationary, and we are proposing a practical tool to handle kinship cases, we believe that this property is not of primary importance here.

The proportional mutation model

The proportional mutation model arises from the assumption that, whenever a mutation takes place, the new allele value is generated at random from the population gene frequency distribution. Under this model, in a locus, every allele can mutate to each of the other possible alleles of the ladder $\mathcal{L}$ with probability proportional to the allele frequency in the reference population. Considering a generic locus, let $i \in \{m, p\}$ indicate the maternal or paternal lineage of the mutation, then μ_i is the associated mutation rate. Let p_a and p_b be the frequency of its alleles a and b in the population, so that $a, b \in \mathcal{L}$. Then $M^i_{a,b}$ is the probability that the parental allele a belonging to lineage i is received as allele b by the child. The model defines M^i as the gender-related mutation matrix with entries $M^i_{a,b}$:

$$M^i_{a,b} = \begin{cases} 1 - \mu_i, & \text{if } a = b, \\ \mu_i p_b, & \text{if } a \neq b. \end{cases} \tag{4.8}$$

Even if the proportional model is not very realistic, since a mutation to an allele far from the original one is rare, it has the advantage of taking care of any possible specific mutation and it is also computationally efficient.

The one-step mutation model

The one-step mutation model assumes that any mutation can only be towards a neighbouring allele value, i.e. only mutations consisting out of exactly one step are possible. Let K be the number of different allele values in the ladder, and, as before, let $M^i_{a,b}$ be the probability that the parental allele a belonging to lineage i is received as allele b by the child, then the elements of the mutation matrix for the one-step mutation model are (for $a \neq b$):

$$M^i_{a,b} = \begin{cases} \mu_i, & \text{if } |a-b| = 1 \ \& \ a = 1 \text{ or } K, \\ \frac{\mu_i}{2}, & \text{if } |a-b| = 1 \ \& \ a \neq 1 \text{ or } K, \\ 0, & \text{otherwise}, \end{cases} \tag{4.9}$$

and finally $M^i_{a,a}$ is such that $\sum_{b \in \mathcal{L}} M^i_{a,b} = 1$. Please note that not all alleles mutate with the same probability: they do so either with probability μ_i, $\mu_i/2$ or zero.

Studies, e.g. Brinkmann et al. (1998), about mutations in STR markers used in forensics suggest that the great majority of mutations on the forensic STR loci involves the addition or deletion of one repeat unit from the parental allele. This makes the one-step mutation model plausible at the biological level, although the absolute ban on mutation by more than one step could be too extreme.

The mixed mutation model

The mixed mutation model (hereafter abbreviated with MMM) is obtained by mixing a single step model and a proportional model with weights h and $1-h$

respectively, with $h \in [0,1]$. For this reason the elements of M^i are

$$M^i_{a,b} = \begin{cases} h\mu_i + (1-h)p_b\mu_i, & \text{if } |a-b| = 1 \ \& \ a = 1 \text{ or } K, \\ h\dfrac{\mu_i}{2} + (1-h)p_b\mu_i, & \text{if } |a-b| = 1 \ \& \ a \neq 1 \text{ or } K, \\ (1-h)p_b\mu_i, & \text{if } |a-b| > 1, \end{cases} \tag{4.10}$$

and again $M^i_{a,a}$ is such that $\sum_{b\in\mathcal{L}} M^i_{a,b} = 1$. The mixing proportions are specified according to the fact that a mutation of more than one repetition in the STR sequence is rather uncommon, thus a reasonably realistic value for h might be 0.9, emphasising single step mutations, but retaining a non-negligible probability for mutation by other amounts.

We decided to implement in our analysis the mixed mutation model since it is the most plausible among the three presented, it comprises the proportional and the one-step models as special cases and it is still quite computationally tractable.

Furthermore, if *null alleles* are not explicitly modelled, the choice of a mutation model indirectly implies the choice of a way to account (or not) for their presence. Null alleles (also called *silent* alleles) are alleles that do not amplify, thus leading to believe to observe homozygous genotypes while in fact the true genotype is heterozygous and involves a null allele. For example, a shared null allele between parent and child may result in different homozygousity, and lead to an incorrect exclusion of paternity. Null alleles can, of course, be simply modelled as a new allele, but there are two drawbacks in doing this: it makes the software computationally more complex and hence slower, and it requires the population frequencies of null alleles per locus, for which some data exist[1], but they are generally not precisely known. We will therefore not consider this as a possibility, but let the mixed mutation model handle null alleles. In fact, since the proportional part of model gives a positive probability to every mutation from an allele to any other allele, if a null allele is transmitted from parent to child this would be treated by the model as a mutation, always assigning a probability greater than zero to that genotype combination and not letting a $LR = 0$ arise.

4.3.2 Population models

The baseline model for the populations is, in this thesis, called HW since it is derived from the equation introduced by Hardy-Weinberg for a population in equilibrium. The genotype probability is calculated from the assumed known probabilities of the alleles in the population, by simply using formula (2.2).

Uncertain Allele Frequencies

A more realistic model relaxes the assumption of known allele probabilities and considers them uncertain. Considering a database of individuals available for forensic inference as a random sample from a reference population, for a locus, the observed alleles' frequencies, $\mathbf{n} = \{n_1, \ldots, n_k\}$, $N = \sum_{i=1}^{k} n_i$, follow a multinomial distribution conditional on the vector of the actual frequencies in the population $\mathbf{p} = \{p_1, \ldots, p_k\}$. The prior probabilities on $\mathbf{p}$ are usually modelled by a Dirichlet distribution, $\mathbf{p} \sim Dir(\boldsymbol{\delta})$ with $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_k\}$, so that the posterior distribution is $\mathbf{p}|\mathbf{n}, \boldsymbol{\delta} \sim Dir(\delta_1 + n_1, \ldots, \delta_k + n_k)$. If, in the pedigree involved in the identification trial, two or more founders' alleles are not observed and their probabilities are uncertain, the alleles become dependent. The Uncertainty in Allele Frequencies

[1] `http://www.cstl.nist.gov/biotech/strbase/NullAlleles.htm`

model (UAF), proposed by Green and Mortera (2009), states that if S founders' alleles are considered, the marginal distribution of the Sth allele's probability's assuming the value j is a mixture formed by the marginal of $\mathbf{p}|\mathbf{n}, \boldsymbol{\delta}$, i.e., a Beta distribution, and the probability mass proportional to the number of js observed on the previous $S-1$ founder alleles, i.e.

$$p_j^{(S)}|\mathbf{n}, \boldsymbol{\delta} \sim \frac{\sum_i^k \delta_i + N}{M} Beta(\delta_j + n_j, \sum_{i \neq j}^{k} \delta_i + n_i) + \frac{1}{M} \sum_{s=1}^{S-1} \mathbb{I}_{\{s\}}(j),$$

so that

$$Pr(p_j^{(S)}|\mathbf{n}, \boldsymbol{\delta}) = \frac{\delta_j + n_j}{M} + \frac{1}{M} \sum_{s=1}^{S-1} \mathbb{I}_{\{s\}}(j), \tag{4.11}$$

where $M = \sum_{i=1}^{k} \delta_i + N + S - 1$ and $N = \sum_{i=1}^{k} n_i$. Including (4.11) into (2.2) as one of the $\mathbf{p}$ produces the required genotype probability.

The UAF model can be also expressed by the Pòlya Urn scheme, and for this reason it is easily amenable to be represented by a Bayesian network.

Coancestry

The existence of sub-populations can be taken into account considering that they had a small number of founders so that the probability F an individual's receiving two copies of the same allele transmitted by the same ancestor is not negligible. In Section 2.3.2 we modified HW equation to take into account this sub-population effect, obtaining formula (2.3).

It is straightforward to include coancestry in the population model of an identification system: it is sufficient to redefine the CPTs according to (2.3). This model is known as the Balding-Nichols model (Balding and Nichols (1995)) and the parameter F may be interpreted as measuring the degree of uncertainty about

p_a as an estimate of the match probability for a single allele. A number of plausible values for F ranging from less than 1% to more than 5% have been proposed in the literature, mainly depending on the ethnicity, and the mentioned paper of Balding and Nichols is a good starting point. In this thesis, when coancestry will be considered in the applied cases of Chapter 6, a value for F equal to 0.02 will be adopted.

4.4 The evaluation of the identification system

In this section we detail our main proposals to evaluate a case-specific kinship identification system. This is a pre-experimental activity for which a key result is getting the LR distributions under the two hypotheses, H_0 and H_1. All these fundamental aspects will be described in details in the next paragraphs.

4.4.1 The pre-experimental phase

We believe that the evaluation activity about an identification system must precede the usual kinship analysis. We maturated this belief when we were called to support familial reunification in collaboration with the IOM, as stated in Section 1.1. An example clarifies this issue.

Suppose some member of an immigrated foreign family ask the permission to bring in an alleged relative still living in the Country of origin: while their genetic evidences are immediately and easily available, that of the candidate relative could not, due to the distance. It may be worthy to assess the potentialities of the identification system before incur costs (in terms of money and time) of getting evidence from the candidate. Another example could involve the need to evaluate the necessity of an exhumation of the body of a deceased person in advance.

We, thus, defined as *pre-experimental* the phase in which the genetic evidence of C is not available yet, either because it is actually difficult to be collected or because it is of interest to mask it to perform the evaluation of an identification system as we propose in this work.

4.4.2 The LR distributions

In the pre-experimental perspective, the analysis considers the DNA evidence belonging to the individuals promoting the identification trial, i.e. those included in the set $\mathcal{F}^+$, but not that of the candidate to the identification whose position in the familial pedigree is uncertain.

Let X_C be the random variable for the genetic profile of the candidate, C. If its value x_C is known, we have seen that it easy to calculate a ratio, the LR (cfr. Section 4.2), able to express the support to an hypothesis against another one. But, since here C's profile is unknown, the likelihood ratio becomes a random variable, and to stress the dependence on the unobserved X_C we named it $LR(X_C)$. More precisely there are two distributions for $LR(X_C)$ conditionally on the states of H, and the first activity to evaluate a system is to derive them; then the analysis of these distributions produces an evaluation of the system.

To derive the LR distributions we can take advantage of the fact that the loci commonly used in forensic identification are located at large genetic distances and therefore are considered independent. The case in which H_0 considers the individual C as a random man from the population is the most common, so now to differentiate with respect to H_1 we use this specific identification hypothesis for simplicity, being clear that the extension to the case in which $C \equiv U'$ under H_0 formally coincides with the treatment of $C \equiv U$ under H_1.

Let i be a generic locus with k_i different allele values, the possible LRs are determined by calculating (4.5) for all $k_i(k_i+1)/2$ genotypes. Considering n different loci (i.e. $i \in \{1, \ldots, n\}$) the number of possible genetic profiles observable for an individual is $\prod_{i=1}^{n} k_i(k_i+1)/2$. Let X_C be the variable representing the genotypes of C for all the n loci jointly, while $X_{C,i}$ is the genotype of the candidate for the locus i (practically meaning that $X_C = (X_{C,1}, \ldots, X_{C,n})$), then the support of $LR(X_C)$ is given by the set:

$$\mathcal{LR} = \Big\{ \prod_{i=1}^{n} LR(x_{C,i}) : x_{C,1} \in \mathcal{X}_{C,1}, \cdots, x_{C,n} \in \mathcal{X}_{C,n} \Big\}, \tag{4.12}$$

where $\mathcal{X}_{C,i}$ is the sample space for the genotype of C in a generic locus i. Using allele recoding introduced in Section 3.4 we will be able to apply substantial reduction to this space. In Chapter 5 we will give more details about state space reduction and approximations.

Once the state space of the LR has been established, the following step is to obtain the LR distributions which, assuming H_0 or H_1 to hold, depend on the probability of the genetic profiles of C conditionally on H_0 and H_1.

If H_0 holds, the probability of a genetic profile is obtained by factorizing the genotypes' probabilities over the loci through the assumed population model conveyed by H_0, so that:

$$Pr\big(LR(X_C)|H_0\big) = Pr\big(LR(x_{C,1}, \ldots, x_{C,n})|H_0\big) = \prod_{i=1}^{n} Pr(x_{C,i}|H_0) \quad \forall x_{C,i} \in \mathcal{X}_{C,i}. \tag{4.13}$$

If H_1 holds, i.e. $C \equiv U$, then the genetic profiles' probabilities are obtained by factorizing the loci's probabilities derived by $X_C|x_{\mathcal{F}+}, H_1$ and

$$\begin{aligned} Pr\big(LR(X_C)|H_1\big) = Pr\big(LR(x_{C,1}, \dots, x_{C,n})|H_1\big) = \\ = \prod_{i=1}^{n} Pr(x_{C,i} \equiv x_{U,i}|x_{\mathcal{F}+,i}, H_1) \quad \forall x_{C,i} \in \mathcal{X}_{C,i}. \end{aligned} \tag{4.14}$$

Hereafter, we consider the likelihood ratio distributions with regard to all the available loci altogether and for simplicity we refer to LR instead of $LR(X_C)$ or $LR(x_{C,l_1}, \dots, x_{C,l_n})$.

Distributions can be directly represented by histograms or, more conveniently, by Tippets plots (Evett and Buckleton (1996), Gill et al. (2008)), which allow of comparing, in the same graph, the LR probability distributions obtained in different conditions. The analysis of these distributions produces an evaluation of the system.

4.4.3 The probabilistic evaluation of the system

Our proposal to assess the value of an identification system is based on the computation of the probabilities the LR does not support the hypotheses when they actually hold, since originated by *misleading* evidence ($\mathcal{E}^M$), i.e. the genetic profiles producing LR values against the hypothesis assumed to hold. Hence we define two probabilities:

$$Pr(\mathcal{E}^M|H_1) = \sum_{x_C \in \mathcal{X}_C : LR<1} Pr(x_C \equiv x_U|x_{\mathcal{F}+}, H_1), \tag{4.15}$$

$$Pr(\mathcal{E}^M|H_0) = \sum_{x_C \in \mathcal{X}_C : LR>1} Pr(x_C|H_0). \tag{4.16}$$

Evidence in $\overline{\mathcal{E}^M}$ is referred to as *faithful* evidence ($\mathcal{E}^F$). Each party involved in the trial can evaluate whether the system matches their requirements, either when

H_0 holds or when H_1 holds.

Typically, people favouring identification, the *pro-id*, believe $C \equiv U$ and are more interested in (4.15); those favouring no-identification, the *con-id*, since for them $C \not\equiv U$, are more worried about (4.16). Obviously the smaller (4.15) and (4.16) are, the better the system is. If prior probabilities on H are introduced by someone who is balanced between the positions, typically the *Judge*, the hypotheses can be marginalized out:

$$Pr(\mathcal{E}^M) = \sum_{x_C:LR<1} Pr(x_C \equiv x_U | x_{\mathcal{F}^+}, H_1) Pr(H_1) + \sum_{x_C:LR>1} Pr(x_C|H_0) Pr(H_0). \tag{4.17}$$

Even if $LR = 1$ is the most natural threshold for making a distinction between faithful and misleading evidence, Royal (2000) specifies two others values, $\tau_0 \ll 1$ and $\tau_1 \gg 1$, to partition the LR space into regions providing *strong* and *weak* support for the hypotheses (Table 4.1). In a judicial setting, these thresholds have an interesting meaning since they specify the values of LR capable of updating some prior probabilities on the hypotheses to some required posteriors. A large distance between prior and posterior probabilities produces τ values far from unity. For instance, if $Pr(H_0) = Pr(H_1) = 0.5$ represents the prior state of knowledge and conclusive posterior probabilities for both hypotheses are specified to be at least equal to 0.9933, then τ_0 and τ_1 can be determined by using (4.7), so that

$$\tau_0 = \frac{0.0067}{0.9933} = 0.0067 \text{ and } \tau_1 = \frac{0.9933}{0.0067} = 148.25.$$

The specification of the posteriors on H implicitly refers to a decision rule which ascertains the case only if a certain posterior probability is reached: this translates numerically the requirement that a decision has to be taken "beyond any

Table 4.1: Evidence classification according to the LR thresholds and the hypotheses.

	LR			
	$[0, \tau_0)$	$[\tau_0, 1)$	$(1, \tau_1]$	(τ_1, ∞)
H_0	strong faithful	weak faithful	weak misleading	strong misleading
H_1	strong misleading	weak misleading	weak faithful	strong faithful

reasonable doubt". The issue is considered by Egeland et al. (2006), who reports a different proposal for the posterior probabilities required for identification and states that the figure 0.9973 corresponds to a threshold introduced by Essen-Möller (1938) to consider a paternity as practically proved. We stress that τ_0 and τ_1 are dependent on the prior probabilities, which are the initial state of uncertainty and the required posteriors representing the final stage of (uncertain) knowledge at which the analysis aims. Both probabilities must be ascertained by the person who is in charge of taking a decision about the identification trial.

For the evaluation of the system, the distinction between *strong* and *weak* evidence is helpful. Evidence producing $LR \in [\tau_0, \tau_1]$ is considered *weak* since it updates the prior probabilities to a level below the judge's requirement for posteriors. On the other hand, evidence producing $LR \in [0, \tau_0) \cup (\tau_1, \infty)$ is classified as *strong* evidence and leads to a decision. A high probability of faithful evidence, especially if it is largely strong faithful, $\mathcal{E}^{SF}$, indicates highly satisfactory performance. Also the distinction between strong and weak misleading evidence, $\mathcal{E}^{SM}$ and $\mathcal{E}^{WM}$, is meaningful. Finally note that the popular exclusion probability (Buckleton et al. (2005)) corresponds to $\Pr(LR = 0|H_0)$ and the evidence producing the exclusion is a part of the strong faithful evidence when H_0 holds.

Finally, please note that it is well known that, if H_0 holds, it is always true that:

$$E(LR(X_C)|H_0) = \sum_{x_C \in \mathcal{X}_C} \frac{\Pr(x_U \equiv x_C | x_{\mathcal{F}^+}, H_1)}{\Pr(x_C|H_0)} \Pr(x_C|H_0) = 1,$$

i.e. the probability to get some LRs values favouring the wrong hypothesis is always positive, except in the trivial case when all the probability mass is concentrated in 1. The fact that misleading evidence is unavoidable motivates our approach which keeps under control the probabilities of its occurrence.

4.5 Alternative approaches for the evaluation

During these years of research activity, also other approach to the identification system evaluation have been explored, among whom the most interesting are those based either on information theory or on decision thoery. In the next two paragraphs these techniques will be briefly revised, even if they will not be employed when we will discuss applied cases in Chapter 6.

4.5.1 Information theoretic approach

Information theory (Shannon (1948)) is a branch of applied mathematics and computer science involving the quantification of information. A key measure of information is known as *entropy* which is a measure of the amount of uncertainty associated with the value of a discrete random variable, X. In other words, entropy is a measure of how much the probability mass is scattered over different states. Thus, let $p(x)$ be the probability mass function of X, then

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

is its entropy. Maximum entropy, $\log(m)$, is therefore achieved when X takes m distinct values each with probability $1/m$, while minimum, 0, is obtained when one single state gets probability equal to 1.

If, instead, two discrete random variables, X and Y, are considered, one can calculate the *joint* entropy, that is

$$H(X,Y) = -\sum_{\mathcal{X},\mathcal{Y}} p(x,y)\log p(x,y),$$

or may be interested in knowing the value of uncertainty when one variable (say Y) becomes known. This is called *conditional* entropy and its formulation is

$$H(X|Y) = -\sum_{\mathcal{X},\mathcal{Y}} p(x,y)\log p(x|y).$$

Starting from the different kinds of entropies introduced above, we are able to derive one of the most useful measure in information theory. *Mutual information*, in fact, measures the amount of information that can be obtained about one random variable by observing another. If variables X and Y are considered, it is indicated with $I(X;Y)$ and its formulation is

$$\begin{aligned} I(X;Y) &= \sum_{\mathcal{X},\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= H(X) - H(X|Y), \end{aligned}$$

i.e. the mutual information criterion is the gain in the amount of information due to the knowledge of Y. An important property of mutual information is symmetry, meaning that $I(X;Y) = I(Y;X)$.

The application of information theory for identification purposes was adopted in a paper by Lauritzen and Mazumder (2008) and in Muzumder's PhD thesis (Mazumder (2010)) to assess the informativeness of genetic markers in kinship identification. A similar approach can be of interest for our purposes. Let $\mathcal{F}$ be the set given in Section 4.1.1 and let $\mathcal{H} = \{H_0, H_1\}$ be the set of the possible alternative hypotheses, then applying the mutual information criterion we can define the mutual information as

$$\begin{aligned} I(H; X_{\mathcal{F}}) &= I(X_{\mathcal{F}}; H) = \\ &= \sum_{H \in \mathcal{H}} \sum_{x_{\mathcal{F}} \in \mathcal{X}_{\mathcal{F}}} Pr(x_{\mathcal{F}}, H) \log \frac{Pr(x_{\mathcal{F}}, H)}{Pr(x_{\mathcal{F}})Pr(H)} = \\ &= \sum_{H \in \mathcal{H}} \sum_{x_{\mathcal{F}} \in \mathcal{X}_{\mathcal{F}}} Pr(x_{\mathcal{F}}, H) \log \frac{Pr(x_{\mathcal{F}}|H)}{Pr(x_{\mathcal{F}})} = \\ &= \sum_{H \in \mathcal{H}} \sum_{x_{\mathcal{F}} \in \mathcal{X}_{\mathcal{F}}} Pr(x_{\mathcal{F}}|H) Pr(H) \log \frac{Pr(x_{\mathcal{F}}|H)}{Pr(x_{\mathcal{F}})} = \\ &= \sum_{H \in \mathcal{H}} Pr(H) \sum_{x_{\mathcal{F}} \in \mathcal{X}_{\mathcal{F}}} Pr(x_{\mathcal{F}}|H) \log \frac{Pr(x_{\mathcal{F}}|H)}{Pr(x_{\mathcal{F}})}, \end{aligned} \tag{4.18}$$

hence $I(H; X_{\mathcal{F}})$ is the measure of the information gain on H supplied by $X_{\mathcal{F}}$.

Alternatively, mutual information can be calculated as the difference between the entropies of H and $H|X_{\mathcal{F}}$,resulting in

$$\begin{aligned} &I(H; X_{\mathcal{F}}) = I(X_{\mathcal{F}}; H) = \\ &\quad \sum_{x_{\mathcal{F}} \in \mathcal{X}_{\mathcal{F}}} Pr(x_{\mathcal{F}}) \sum_{H \in \mathcal{H}} \log(Pr(H|x_{\mathcal{F}}))Pr(H|x_{\mathcal{F}}) - \sum_{H \in \mathcal{H}} \log(Pr(H))Pr(H). \end{aligned}$$

As it appears, this proposed measure is not case-specific since all the family members are considered unobserved so that (4.18) provides a single measure for both hypotheses averaging with respect to the prior probability for H.

For our purposes, i.e. conditionally to the available familial genetic evidence, i.e. $\mathcal{F}^+$, (4.18) becomes:

$$I(H; X_C|x_{\mathcal{F}^+}) = \sum_{H\in\mathcal{H}} Pr(H) \sum_{x_C\in\mathcal{X}_C} \log\Big(\frac{Pr(x_U \equiv x_C|x_{\mathcal{F}^+}, H)}{Pr(x_C)}\Big) Pr(x_U \equiv x_C|x_{\mathcal{F}^+}, H) \quad (4.19)$$

Considered as the expected utility of an experiment, $I(H; X_C|x_{\mathcal{F}^+})$ first computes the divergence between the distribution of X_C for each hypothesis, and the "average" model $Pr(x_C) = Pr(x_C|H_0)Pr(H_0) + Pr(x_C|x_{\mathcal{F}^+}, H_1)Pr(H_1)$; then it averages the results according to the prior probabilities on H.

The proposal is attractive, but the unavoidable dependence on the prior probabilities of H could be problematic since in a trial it could be not possible to reach an agreement among the parties. At the same time this result could not be easily perceived in a Court because what really matters the judges is the loss represented by the probability the identification system provides support against an hypothesis if it is actually true. Furthermore, also the use of the marginal probability on X_C (i.e. $Pr(x_C)$) can be confusing and/or poorly attractive. For these reasons, as in Berger (2000), we propose to use the conditional approach to derive the LRs distributions and to evaluate the system in a frequentist fashion, computing the probability to observe LR values when H_0 and H_1 are assumed to be true.

An historical excursus concludes this section. Before information theory was developed by Claude E. Shannon, Alan M. Turing, in some unpublished works (as accounted by Good (1979)), was the first to consider the expected values of the $\log(LR)$, i.e. the logarithm of the updating factor in (4.7), as the fundamental quantity to measure the value of an experiment devoted to compare two hypotheses.

For kinship identification, for one locus, the expected $\log(LR)$s are:

$$E(\log(LR)|H_1) = \sum_{x_C \in \mathcal{X}_C} \log\Big(\frac{Pr(x_U \equiv x_C|x_{\mathcal{F}+}, H_1)}{Pr(x_C|H_0)}\Big) Pr(x_U \equiv x_C|x_{\mathcal{F}+}, H_1) \tag{4.20}$$

$$E(\log(LR)|H_0) = \sum_{x_C \in \mathcal{X}_C} \log\Big(\frac{Pr(x_U \equiv x_C|x_{\mathcal{F}+}, H_1)}{Pr(x_C|H_0)}\Big) Pr(x_C|H_0), \tag{4.21}$$

which can be easily generalized to more loci by (4.12). Obviously an identification system should provide large ($\gg 0$) and small ($\ll 0$) values for (4.20) and (4.21), respectively.

Elsewhere Good (1985) emphasized the importance of the entire distribution of the $\log(LR)$, which is related to the mutual information criterion (e.g. Cover and Thomas (2006)) and originally applied to the design of an experiment by Lindley (1956).

4.5.2 Decision theoretic approach

Another approach for the evaluation of an identification system relies on statistical decision theory. Decision theory in economics, mathematics, and statistics is concerned with identifying the uncertain variables and other issues relevant in a given decision, its rationality, and the resulting optimal behaviour. Here only the very essential concepts of decision theory will be introduced and employed.

The decision theoretic framework has been applied in the field of forensic DNA analysis by Taroni et al. (2007). To evaluate the system by a decision analysis we need to define the following quantities:

- ***Decisions***. Consider the possibility to choose among n identification systems, differing for some characteristics, and devoted to cope with a specific identification problem. The decision consists in choosing among the alternatives $\mathcal{D} = \{d_1, \ldots, d_n\}$ indicating the system to use.

- ***Outcomes***. The LR distributions, one for each identification hypothesis, are the uncertain outcomes. They vary according to the system employed.
- ***Consequences***. Each possible value of the outcome, $LR = j$ or simply LR_j, jointly with a decision d_i produces a consequence C_{ij}. For instance, different systems may require different laboratory activities and costs, leading to different consequences for the same LR_j. Here costs related to different decisions are considered negligible with respect to the matter implied in an identification. For this reason consequences simply coincide with LRs.
- ***Utility or loss***. Consequences, conditionally to the hypothesis assumed to hold, can be measured by using an utility, $u(LR_j|H_z)$, or a loss function $l(LR_j|H_z)$, $z \in \{0, 1\}$.

In this decision theory framework the final decision is strictly connected with the fundamental aim of the evaluation: to effectively classify among alternative hypotheses.

In the following we describe two aptitudes, relevant in identification and concerning the evaluation of consequences. We define as *Problem-solver* aptitude that of an actor who eminently appreciates systems strongly supporting the identification hypothesis assumed to hold. The same value of utility is attributed to all the LR_js strongly supporting the holding hypothesis. On the opposite no utility is attributed to the other LR_js. The proposed utility functions, also represented in Figure 4.2), are

$$u(LR_j|H_0) = \begin{cases} 1, & \text{if } LR_j \leq \tau_0, \\ 0, & \text{if } LR_j > \tau_0, \end{cases} \tag{4.22}$$

and

$$u(LR_j|H_1) = \begin{cases} 0, & \text{if } LR_j < \tau_1, \\ 1, & \text{if } LR_j \geq \tau_1. \end{cases} \tag{4.23}$$

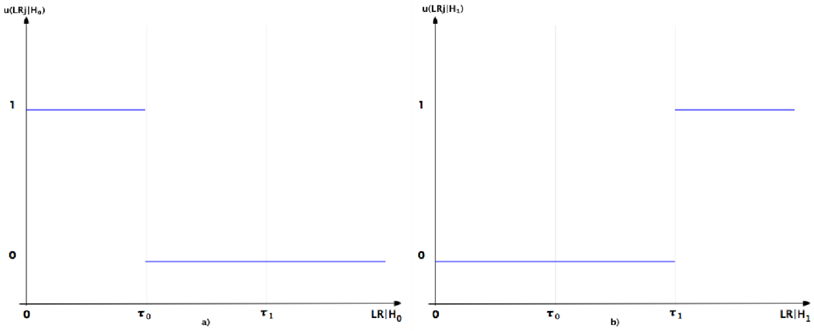

Figure 4.2: Utility functions under H_0 a) and H_1 b), for the problem solver aptitude.

Alternatively, the *Conservative* aptitude is that of individuals mostly alarmed by the possibility the system produces false identifications. Since the pessimistic aptitude, consequences are measured by a loss function. The proposal, also represented in Figure 4.3, is

$$l(LR_j|H_0) = \begin{cases} 0, & \text{if } LR_j < \tau_1, \\ 1, & \text{if } LR_j \geq \tau_1, \end{cases} \tag{4.24}$$

$$l(LR_j|H_1) = \begin{cases} 1, & \text{if } LR_j \leq \tau_0, \\ 0, & \text{if } LR_j > \tau_0. \end{cases} \tag{4.25}$$

To define the thresholds one can refer to those defined in Section 4.4.3: if a single $\tau = 1$, is used (collapsing the previous two in $\tau_0 \equiv \tau_1 \equiv \tau$), this value

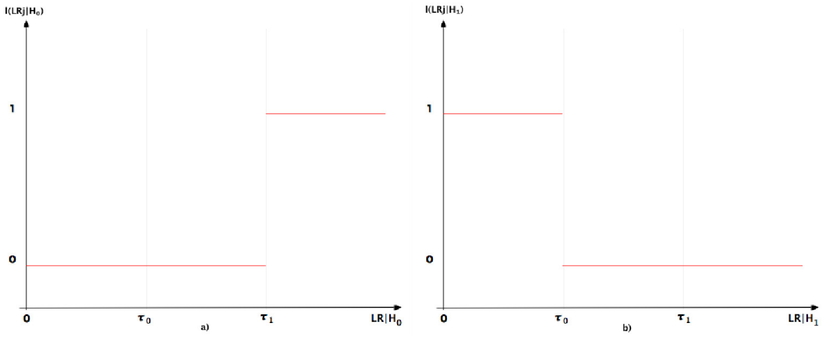

Figure 4.3: Loss functions under H_0 a) and H_1 b), for the conservative aptitude.

splits the LR support in two regions, one favouring H_0 ($\tau < 1$), the other H_1 ($\tau > 1$). Other, perhaps more meaningful, thresholds can be specified considering the behaviour of the actors involved in the system evaluation.

More specifically, the *Judge* is the person called to decide about the identification controversy: their prior probabilities on the identification hypotheses, and the posteriors required to assume a decision about the identification, can be used to define the thresholds. Once these probabilities have been elicited, then the evaluation of τ_0 and τ_1 in (4.22 - 4.25) is reached by using (4.7), having previously specified the requirements for a decision. Finally, two other parties have interest in assessing the value of the system: those favouring identification (*pro-id*) and those against (*con-id*). Their role suggests reasonable prior probabilities on H. Since the pro-ids believe that $C \equiv U$, their prior probabilities could be close to $Pr(H_0) = 0$ and $Pr(H_1) = 1$. The con-ids strongly believe that $C \not\equiv U$, so their prior probabilities could tentatively be $Pr(H_0) = 1$ and $Pr(H_1) = 0$. To easily convey the concept with a simple example, in a paternity case the mother could wish the candidate to be identified as the father of the baby; the alleged father might desire the opposite and a judge wants to provide a fair sentence.

The aim of a decision analysis is to compute the expected utility and/or loss conditionally to each d_i through:

$$E(u|d_i) = \sum_{z\in\{0,1\}} \sum_{j\in\mathcal{LR}} u(LR_j|H_z) \cdot Pr(LR_j|d_i, H_z) \cdot Pr(H_z), \tag{4.26}$$

$$E(l|d_i) = \sum_{z\in\{0,1\}} \sum_{j\in\mathcal{LR}} l(LR_j|H_z) \cdot Pr(LR_j|d_i, H_z) \cdot Pr(H_z), \tag{4.27}$$

and to maximize (4.26) or minimize (4.27) choosing among the available decisions.

The expected loss and utility vary according to the parties since their different prior probabilities on H. The pro-ids actually consider only the case in which the LR distribution is expressed conditionally on H_1 (Figure 4.2b and Figure 4.3b). The con-ids only take account of the distribution of $LR|H_0$, as shown in Figure 4.2a and Figure 4.3a.

In this proposal the expected loss and utility evaluated for the parties are easily interpretable. If we choose $\tau = 1$ and the problem-solver attitude, the expected utility amounts to the probability the system supports H_1 (pro-id), H_0 (con-id) and average of them (Judge) when the hypotheses hold. Conversely, if the conservative attitude is assumed, the expected loss is the probability the System supports the hypotheses when they are not actually true. Alternatively, if τ_0 and τ_1 are specified, the expected utility and loss are, respectively, equal to the probability of false and correct identification according to the decision rule on which τ_0 and τ_1 are chosen.

The proposed loss and utility functions are an extreme version of more realistic and smooth alternatives, but they have the merit to produce expected values of the utility and loss functions interpretable in term of the probability of some meaningful regions of the LR distributions, for example establishing a connection between the obtained losses and the probability of getting misleading results using the system. Attempts to combine them inevitably would obscure important characteristics of the system.

4.6 Summary

In this Chapter we have presented the very fundamental aspects of our research. In the first part we detailed the logic and the main aspects behind kinship analysis, also introducing the relevant mathematical notation; then we focused on the method used to easily obtain the likelihood ratio based on the available evidence.

Then the segregation and population models, already introduced in Sections 2.3 and 3.1, are detailed to cope with model refinements due to mutations, coancestry and uncertainty in allele frequencies, replacing the baseline assumptions of Hardy-Weinberg equilibrium and mendelian segregation.

The evaluation of an identification system is pursued in a pre-experimental phase in which some evidence, the one from the so called candidate to the identification, is unavailable or masked. The LR distributions under the alternative identification hypotheses are thus obtained and, on the basis of them, we gave a probabilistic assessment of the potentialities of a kinship identification system.

Finally we gave two alternative approaches to assess the value of an identification system either using information theory concepts or Bayesian decision theory techniques. Although promising, they both have the drawback to be difficult to be perceived and interpreted in a court.

CHAPTER 5

Computational aspects

For the purposes of our research, the probabilistic evaluation of an identification system, one very fundamental aspect is the ability to get the likelihood ratio distributions conditionally on both the identification hypotheses, for a specific case of interest, as explained in Section 4.4.2. Even if this may appear a simple task, it is not trivial instead. In fact, as fast more detailed models refine the analysis, it becomes more and more difficult to handle the huge state space an individual genetic trace, and hence the LR, can possibly assume.

In this Chapter we therefore provide some original contributions in the field of approximated methods in order to get rid of the computational complexity and finally to obtain the distributions of interest not relying on simulations, but rather on approximations.

Furthermore we will illustrate the software we used to implement the proposed methodology, especially focusing on how to build Bayesian networks to model the genetic information of a pedigree, also handling features of the models detailed in Section 4.3.

5.1 Computational strategies

The main computational issue is to obtain the LR distributions for all the loci jointly considered, when H_0 and H_1 respectively hold. The task can be usefully pursued in two steps.

First of all, for each locus we need to efficiently derive the LR distribution conditionally to each hypothesis. This first step is presented in Section 5.1.1.

The second step consists in deriving the LR distributions for each possible arrangement of the genotypes on different loci and we will detail how to do this in Section 5.1.2.

5.1.1 LR single locus computations

For each locus we need to efficiently derive the X_C distribution conditionally on each hypothesis, following the population and the segregation models appropriate to the case. By the ratio of the probabilities of each possible state of X_C, conditionally on H_0 and H_1, we can derive the possible values assumed by the LR for each locus. The LR distributions are derived assigning to each LR value either the probability given by $Pr(X_C|H_0)$ or that from $Pr(X_C|x_{\mathcal{F}^+}, H_1)$, according to which hypothesis is assumed to hold, which can be obtained from equations (4.13) and (4.14) of Section 4.4.2.

Since, as seen in Section 3.1, a Bayesian network can easily address the probabilistic structure of an identification problem, we settled up BNs appropriate for the cases, taking into account the segregation and population models described in Section 4.3.

5.1.2 LR multi locus computations

The second step is to derive the LR distributions for each possible arrangement of the genotypes on different loci. The most immediate but naive way to obtain the LR distributions is by exact computation. The LR support, for all the loci considered altogether, is derived by applying (4.12), and then their distributions under H_0 and H_1 are obtained using (4.13) and (4.14).

Unfortunately, exact computations produce a number of LR values exponentially increasing according to the number of genotypes for each locus, so that the LR sample space rapidly becomes intractable. For example, if 15 loci in a kit are considered, all of them with allelic ladder of length 10, the resulting size of the genetic profiles state space, and thus that of the LR, is $|\mathcal{LR}| = \Big(\frac{10(10+1)}{2}\Big)^{15} = 1.2748 \times 10^{41}$.

LR equivalence classes

Not all the LRs induced by every possible genetic trace assumed by the Candidate are different. Consider for instance the case provided in Appendix A, where a DNA donor posed on the direct lineage n generations far from U, is attempting the identification of a candidate. In this case, by (A.1), the number of different LR values in a locus is three or six, depending on whether the donor in $\mathcal{F}^+$ is homozygous or heterozygous, so it does not depend neither on the number of possible alleles in the locus nor on n. This implies that we are not required to consider as many LRs as the number of possible different candidate's profiles, but the LR distributions can be obtained summing up the probabilities of the profiles producing the same LR. For this reason, these profiles constitute an equivalence class. For identification cases differing from the one detailed in Appendix A, the reduction in complexity can be obtained numerically by aggregating the identical LRs obtained for each locus.

LR quasi-equivalence classes

The strategy outlined above does not introduce approximations of the LR distributions, but it could not downsize $|\mathcal{LR}|$ to a tractable dimension. A possibility for going further is to use a form of approximation treating profiles producing very close LR values as belonging to the same equivalence class. Here we consider the issue of obtaining a reduced LR size along with the evaluation of the LR distributions' approximation accuracy.

We decided to adopt the following rule, ordaining that the jth and $j+1$th ordered LRs are collapsed if

$$\Delta_{(j)} = \frac{LR_{(j+1)} - LR_{(j)}}{LR_{(j)}} < \varepsilon$$

where ε is related to the amount of downsizing and to the approximation accuracy. To produce the LR distributions, calculations are performed by considering the loci iteratively. At every iteration a new locus is added, thus increasing the state space. By applying this rule at every step we can contrast the induced size growth.

To control the approximation accuracy issue, consider that if two LRs are merged, their values are substituted by a unique LR^* obtained by specifying $f(\cdot,\cdot)$ in

$$LR^*_{(j)} = f\big(LR_{(j+1)}, LR_{(j)}\big). \qquad (5.1)$$

The LR^* distributions are obtained by summing up all the probabilities of the profiles included in the quasi-equivalence classes evaluated according to each hypothesis.

Furthermore, to achieve a given desired accuracy, consider that if

$$f(\cdot,\cdot) = max\big(LR(l_i)_{(j+1)}, LR(l_i)_{(j)}\big),$$

then

$$Pr(LR^*_{\mathtt{max}} > \tau) \geq Pr(LR > \tau) \quad \forall\tau.$$

If $f(\cdot,\cdot) = min\big(LR(l_i)_{(j+1)}, LR(l_i)_{(j)}\big)$, then:

$$Pr(LR^*_{\mathtt{min}} < \tau) \geq Pr(LR < \tau) \quad \forall\tau.$$

This implies that:

$$Pr(LR^*_{\mathtt{max}} < \tau) \leq Pr(LR < \tau) \leq Pr(LR^*_{\mathtt{min}} < \tau)$$

and

$$Pr(LR^*_{\mathtt{min}} > \tau) \leq Pr(LR > \tau) \leq Pr(LR^*_{\mathtt{max}} > \tau),$$

i.e., the $LR^*_{\mathtt{min}}$ and $LR^*_{\mathtt{max}}$ distributions are the upper and lower bounds determining an interval surrounding the probability of the required subset of LR values. The approximation accuracy can be measured by the relative difference between the two probability approximations and the reduction in complexity is the relative difference between the size of LR and of LR^*.

5.2 Software

We decided to implement all the algorithms in a general purpose programming language and we chose MATLAB®, which has shown excellent performance in terms of flexibility and computational times. There are many useful toolboxes that can be added to the MATLAB® core to expand its features. We used one of these toolbox, the Bayes Net Toolbox (BNT) for MATLAB® developed by Murphy (2001), to built appropriate Bayesian networks directly into the MATLAB® environment.

The Bayes Net Toolbox is open-source, free of charge, and supports many different models and inference algorithms. For this latter aim, conditional on the parameters, inference of the variables can be performed exactly using the junction tree algorithm.

For our purposes, the main advantage of the BNT is that it is able to easily return the complete conditional probability distribution on any variable of interest with just one propagation step for every locus considered. This greatly benefits the feasibility of the analysis.

Other computational resources

There are many computer programs for computing the likelihood ratio in kinship analysis. Drábek (2009), for example, reviews and tests Familias and DNA-View™. Familias derives the probability of the observed evidence marginalizing the unobserved individuals in the pedigrees by means of the peeling algorithm (Lauritzen and Sheehan (2003)); DNA-View™ obtains the same result algebraically as a function of the parameters required by the population and segregation models which are used. DNA-View™ is also able to simulate the genetic traces for a specific individual in a pedigree. Unfortunately, the software does not provide any tool with which to assess the convergence of the simulated distributions, nor provide any control on the accuracy.

A version of Familias, called OpenFamilias, written as an R package, allows to use, within an R shell, all the facilities of Familias, manipulating the intermediate results. To use OpenFamilias for our computations a possible strategy would be to call repetitively the *getProbabilities* function, devoted to obtaining the probability of the evidence, considering at each run a different genotype for the candidate. This procedure is feasible, but it is not efficient and fails to exploit the Bayesian networks' ability to derive the distribution of the candidate genotypes according to each hypothesis by a single propagation step.

5.3 Bayesian networks

In this Section we will shortly present and display the relevant Bayesian networks we built to cope with the models described in Section 4.3. We refer to the pedigree represented as genotype network in Figure 3.4. Each of the networks that will be given describes the kinship structure for a single locus: distinct loci require distinct networks, but these will differ only in the details of the alleles' population frequencies, which can easily be changed.

Note that here the distinction between the solid and dashed nodes is purely for presentational purposes and has no practical effect on the analysis. This serves only to distinguish between nodes representing genotypes, which are in principle observable (displayed with solid lines), and nodes for other variables (depicted using dashed lines).

Baseline models

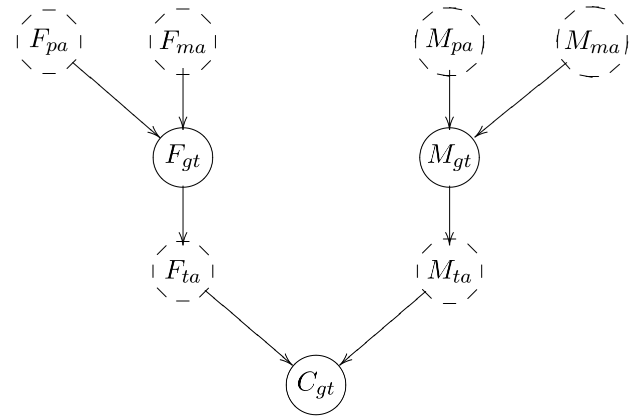

Figure 5.1: Bayesian network for the trio in Figure 3.3.

In Section 4.3 we called baseline models those representing the simplest biological mechanisms behind allele frequencies in the population and segregation processes. Both mendelian law of segregation and Hardy-Weinberg equilibrium can be handled by the Bayesian network in Figure 5.1. Subscripts ma and pa respectively stand for maternal allele and paternal allele, i.e. the two alleles making the genotype of an individual from the population, this latter indicated by the subscript gt.

Conditionally on the alleles, the distributions of which depend on their frequency in the population, the genotype is obtained deterministically. Then, with equal probability, one of the two allele of the parental genotype is selected to be transmitted (subscript ta) and it is passed to the child with no mutation process.

Coancestry

To handle the sub-population structure induced by coancestry it is sufficient to modify the Hardy-Weinberg equation, (2.2), including a new parameter F as in equation (2.3) of Section 2.3.2. Recalling from Section 4.3.2, F is the probability that an individual receives two copies of the same allele transmitted by the same ancestor. If an extra node pointing to the genotype node of an individual (as the node F in Figure 5.2 that is a parent node for the genotype of a generic person I) is added to our Bayesian network, then it is possible either to consider different values for it, each one associated with a certain prior probability, or to consider a degenerate distribution assigning the entire probability to a value of interest. As stated in Section 4.3.2, in our applied cases we will adopt this latter strategy and consider $F = 0.02$.

The structure depicted in Figure 5.2 must obviously be replied for every person involved in the network.

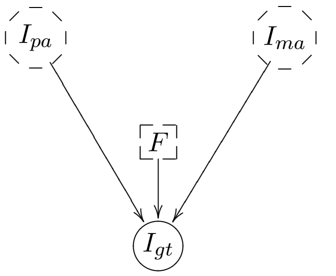

Figure 5.2: Bayesian network for coancestry.

Mixed mutation model

In Figure 5.3 the Bayesian network for the mutation process induces by the mixed mutation model of Section 4.3.1, for the paternal inheritance lineage, is shown. Since the mutation mechanism, there the child paternal allele (C_{pa}) can be in turn a mutated version (F_{mut}) of the father's transmitted allele (F_{ta}) or a copy of it. The node labelled with *mut?* rules these options, assigning probability μ_p to the earlier case and $1 - \mu_p$ the the latter. Finally the value of F_{mut} is obtained either as a result of a one step mutation (*1step* node) or as the outcome of a proportional mutation (handled by node *prop*, the probability distribution of which is that of the associated allele in the population), mixed with weight given by the h node.

An analogous network can be settled up for the segregation process of the maternal lineage.

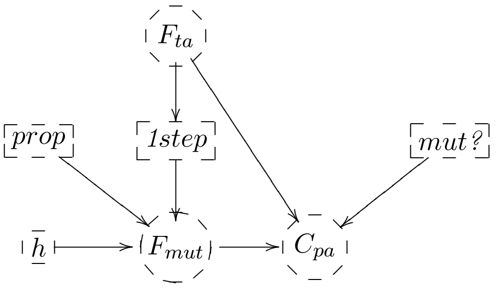

Figure 5.3: Bayesian network for the mixed mutation model.

Uncertain allele frequencies

The Bayesian network structure to model uncertainty in allele frequency is depicted in Figure 5.4. There, node M represents the parameter of the same name (representing the size of the allele database once the familial evidence has been considered) given in Section 4.3.2 while talking about the UAF model. Then the four paternal and maternal genes of the individuals in the pedigree (F_{ma}, F_{pa}, M_{ma} and M_{pa}) are either a random draw from the population (the *pool* nodes) or a copy from the previous ones. Binary d and *temp* nodes have only an instrumental role, avoiding gene nodes' space state growth. Finally, nodes representing founding genes are then connected to those for the genotypes and the segregation mechanism given in the previous paragraphs.

More details and the pseudo-code for this Pòlya Urn scheme can be found in

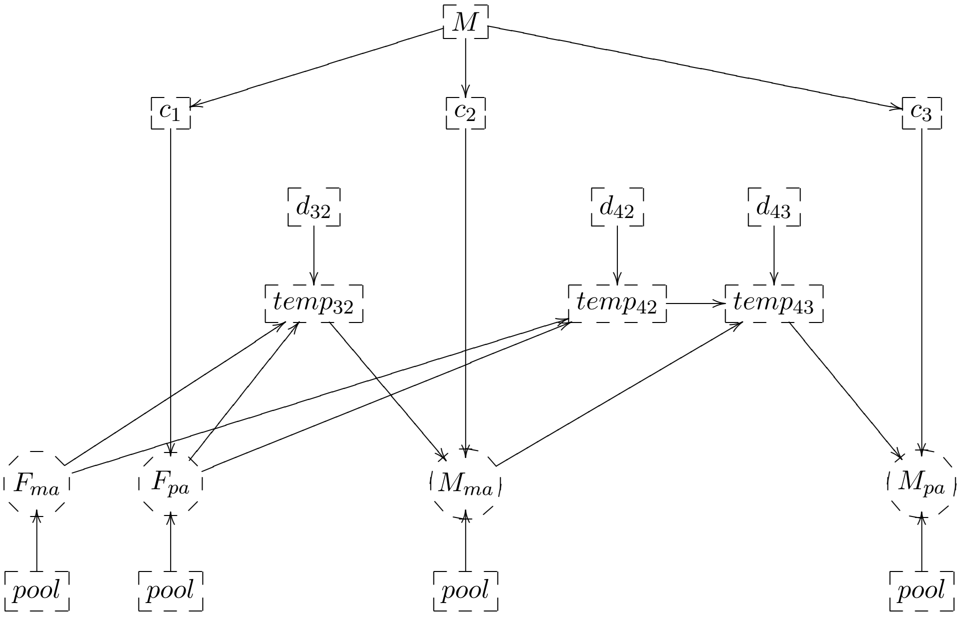

Figure 5.4: Bayesian network for the uncertain allele frequency model.

the original paper by Green and Mortera (2009).

5.4 Summary

In the first part of this Chapter we presented some relevant findings of our research in terms of computational strategies to cope with the high dimensional space of the LR distributions. We showed how to build approximation methods to shrink the state space of the genetic trace of an individual, thus obtaining the whole LR distributions without the need to use simulation methods.

In the, brief, second part of the Chapter we made clear what pieces of software we used for this work, also mentioning some alternative computational resources and the reasons why we did not make use of them.

Finally we showed the Bayesian networks built to treat models introduced in

Section 4.3. We stressed the role of the more relevant nodes appearing in the figures and gave a practical demonstration on how an identification problem can be easily described by exploiting the conditional independence relationships of a Bayesian network.

CHAPTER 6

Applied cases

In this Chapter we first describe how the LR distributions can vary within and between some archetypical identification issues by reviewing 71 real cases we have come across in the last ten years, since the start of our collaboration with the IOM. We experimented a great variety of systems' behaviours, depending on the kind of kinship analysis, on the kind and number of evidences considered and on the alternative hypotheses investigated.

After that, we will give further details about four interesting cases, also showing the sensitivity of the results to the model assumptions, and whether the cases need to be improved by including new familial evidence and/or more loci.

6.1 Cases review

Here we re-examine 71 of the cases we treated in our collaboration with the IOM. For each case, we performed a system evaluation from the pre-experimental perspective under the probabilistic approach we advocated in Section 4.4. We will see that this activity could have revealed in advance the low effectiveness of some identification systems in achieving decisive findings.

The 71 cases, classified with respect to the specific identification issue, are summarized in Table 6.1, which gives the ranges of probabilities of faithful evidence conditionally on both hypotheses for all the cases included in each class of kinship. The 71 cases were treated without altering the systems used at the time they have been originally considered, and the results were obtained assuming HW equilibrium and Mendel's law to hold. The classification on the identification issues is structured in the following manner:

- **Full Paternity:** a mother and her child try to recognize a man as the father, the alternative being that this man is a generic person from the reference population;
- **Motherless Paternity** As before, but the mother's genetic data are not available;
- **Full Sibling I:** a person tries to identify a candidate as a full sibling, the alternative being that the candidate is an unrelated individual;
- **Full Sibling II:** as above, but the alternative hypothesis claims that the two individuals are half siblings;
- **Half Sibling:** the candidate is supposed to be the half sibling of the sponsor, or, alternatively, to be unrelated;

Table 6.1: Ranges of observed $Pr(\mathcal{E}^F|H_0)$ and $Pr(\mathcal{E}^F|H_1)$ in 71 kinship cases.

Class of kinship	n	$Pr(\mathcal{E}^F\|H_0)$	$Pr(\mathcal{E}^F\|H_1)$
Full Paternity	25	(0.99997 - 1)	(0.99999 - 1]
M.less Paternity	15	(0.9997 - 0.9999)	(0.99999 - 1]
Full Siblings I	6	(0.9571 - 0.9713)	(0.9605 - 0.9902)
Full Siblings II	7	(0.8702 - 0.9195)	(0.8521 - 0.8889)
Half Siblings	9	(0.8460 - 0.9368)	(0.8398 - 0.9474)
Uncle-Nephew	6	(0.8359 - 0.9015)	(0.8308 - 0.8838)
G.parent-G.child	3	(0.8323 - 0.8606)	(0.8503 - 0.8694)

- **Uncle - Nephew:** an uncle tries to identify a person as his nephew, being alternatively an unrelated individual;

- **Grandparent - Grandchild:** the candidate can be either the grandchild of the sponsor or an unrelated individual.

Table 6.1 shows the clear dependence of the system performance on the class of identification. Paternity cases originate almost perfect classifiers when a mother and a child try to identify an alleged father and also when the mother's genetic data are not available. This is reassuring since these are the most common kinds of identifications. Kinship analyses implying siblings show more variation. For the issue considered by Full Sibling I, the range of probability of observing faithful evidence declines compared to paternity cases, and if the hypotheses become closer (Full Sibling II), the system performance further decreases appreciably. Results for the Half Sibling class show that these cases produce much less satisfactory performance. Finally as expected, for Uncle–Nephew and Grandparent–Grandchild cases, the results belong to the lowest class of performance and support our claim that the less is the degree of relatedness between the family member(s) who require the identification and provide their DNA, and the hypothesized position in the pedigree of the sought for individual, the less reliable is a strategy of identification based on standard procedures.

6.2 Some cases detailed

In this section we give more details about the LR distributions of some cases, and also verify the effect of including new evidence and assess how sensitive the results are to modifications of the model assumptions.

All the cases involve Italian individuals, thus allelic frequencies referring to the Italian population are from Brisighelli et al. (2009). Mutation rates comes from AABB (2008), but they are further corrected for "hidden mutations" in the way we will present in Chapter 7. The results are displayed considering the apportionments of the LR support introduced in Section 4.4.3 and showing some relevant Tippet plots when the case is judged to benefit of this representation. The values of τ_0 and τ_1 are derived keeping in mind the following scheme: $Pr(H_0) = Pr(H_1) = 0.5$ are the prior probabilities and the posterior probabilities are equal to 0.9933. As a consequence, by equation (4.7), $\tau_0 = 0.00675$ and $\tau_1 = 148.254$. The approximation accuracy has always been achieved at a value of $\varepsilon \leq 0.0005$. When coancestry is considered, the parameter $F = 0.02$ is adopted in Equation (2.3). Case specific genetic data are included in Appendix B of this thesis.

For cases representation we used genotype networks (cfr. Section 3.1): observed variables are represented with solid-lined nodes while unobserved variables with broken-lined nodes.

A Motherless Paternity case

A man, B, would like to assess his father's identity. He has serious reasons to believe he is the son of AF (the Alleged Father) who died some years ago. B has no information about his mother, consequently her genetic evidence is not available. Before exhuming AF, to give some indications about the reliability of the system and using data on B, we obtained the LR distributions evaluated according to the hypotheses specified below, also graphically shown by means of genotype networks in Figure 6.1.

- H_0: AF and B are not recently related;
- H_1: AF is B's father.

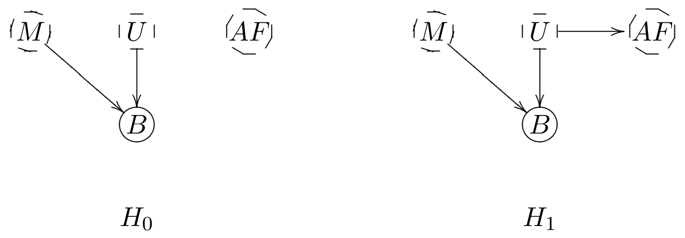

Figure 6.1: Motherless Paternity case - genotype networks under H_0 and H_1.

Looking at Table 6.2, it is apparent that the LR distributions concentrate almost all of the probability on LR values strongly supporting the hypothesis supposed to hold. Moreover, the results are almost insensitive to model's variation by either introducing uncertainty in the allele frequencies, or the mixed mutation model, or allowing for coancestry. Hence any system used for this identification has satisfactory characteristics and can be accepted without any modifications. The same conclusions can be derived looking at the whole LR distributions obtained for the system allowing for the largest number of sources of uncertainty ($COA\&MMM$) and shown in Figure 6.2: it is impressive to note that under both hypotheses, there is no appreciable probability outside the strong faithful regions.

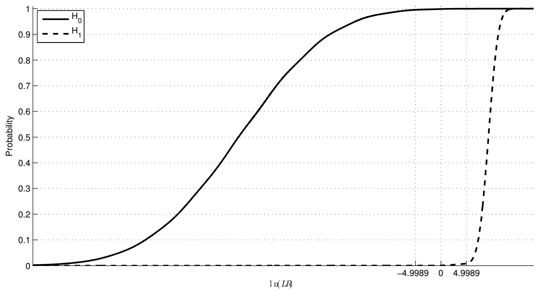

Figure 6.2: Motherless Paternity case - Tippett plot for $\ln(LR)$ distributions allowing for coancestry (COA) and Mixed Mutation model (MMM).

A Full Sibling II case

For reasons concerning transplant compatibility, a man, F_1, would like to evaluate whether one of his brothers, F_X, sharing the mother with him and with two other brothers, F_2 and F_3, also has the same father. The issue arose because in the past there had been some rumours about this matter and the parents died some years ago. Before confessing his doubts to the brothers, F_1 wants to verify the effectiveness of a system using his genetic evidence only. The relevant hypotheses are specified below and depicted in Figure 6.3.

- H_0: F_1 and F_X are half brothers;
- H_1: F_1 and F_X are full brothers.

In Table 6.3, the first two rows answer the question posed by F_1, introducing only his genetic evidence. As it is apparent, there is a large difference from the results shown in Table 6.2: now, for instance, conditionally on H_0 the probability of getting misleading evidence is around 10% and the probabilities of strongly support the correct hypotheses are not very large, being about 10%, if F_X and F_1 were half brothers, and 30%, if they were full brothers.

Table 6.2: Motherless Paternity case - LR distributions obtained using combinations of models.

Models		LR $[0, 0.0067)$	 $[0.0067, 1)$	 $(1, 148.25]$	 $(148.25, \infty)$
HW & ML	H_0	0.9997	0	0.00002	0.00029
	H_1	0	0	0.0019	0.9981
HW & MMM	H_0	0.9951	0.0033	0.0011	0.00029
	H_1	0.00001	0.0006	0.0085	0.9909
UAF & ML	H_0	0.9997	0	0.00004	0.00031
	H_1	0	0	0.0032	0.9968
COA & ML	H_0	0.9997	0	0.00003	0.00028
	H_1	0	0	0.0025	0.9975
COA & MMM	H_0	0.9955	0.0032	0.0011	0.00026
	H_1	0	0.0005	0.0079	0.9916

To realize an enhanced system, F_1 asked F_2 to provide his genetic evidence. The results are in rows 3–4 of Table 6.3: there is a large increase in $Pr(\mathcal{E}^{SF})$ and the probability of obtaining misleading evidence is reduced to 2% for both hypotheses, and is negligible for being strongly misleading. Later, since the story filtered through the family, F_3 provided his genetic contribution to the case: now the system not only contemplates faithful LR values almost with certainty, but it also contributes with high probability to reach high posterior probabilities for both hypotheses, whichever of them holds. Looking at the Tippett plots for $\ln(LR)$ distributions under both hypotheses in Figure 6.4, we have a full display of how the identification system reacts to the addition of evidence. Under H_1, the probability of observing misleading evidence becomes smaller and smaller as long as new evidence is introduced. Under H_0, if two or three brothers are considered, the probability of faithful evidence under H_0 is largely concentrated on very small

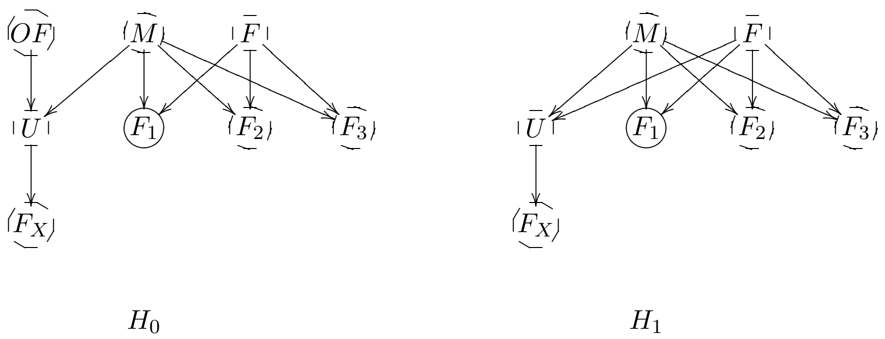

Figure 6.3: Full Sibling II case - genotype networks under H_0 and H_1 when only F_1's evidence is available.

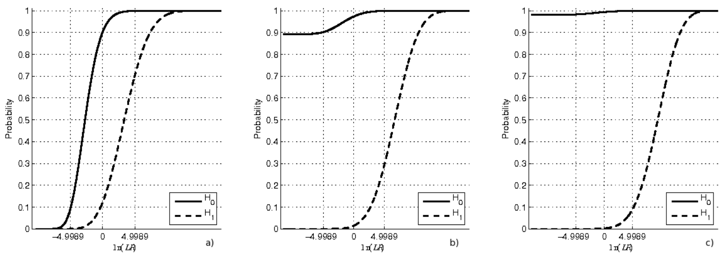

Figure 6.4: Full Sibling II case - Tippett plots for ln(LR) distributions for different amounts of genetic data: a) $\mathcal{F}^+ = \{F_1\}$; b) $\mathcal{F}^+ = \{F_1, F_2\}$; c) $\mathcal{F}^+ = \{F_1, F_2, F_3\}$.

values of the LR distribution, which it is not possible to display in detail even in the log scale. For the system including F_1, F_2 and F_3 we also verified the limited influence on the results due to the introduction of the UAF, the mixed mutation model, and coancestry (lines 7–14 of Table 6.3).

A Two Cousins case

Two people, S and A, allege that they are cousins (in particular, that their fathers are brothers). If the fact were supported by their genetic evidence, the immigration of the applicant, A, to the country where the sponsor S already is a citizen, would be made easier. Because of the alleged weak kinship relationship

Table 6.3: Full Sibling II case - LR distributions and some different familial evidence and models.

			LR			
Evidence	Models		$[0, 0.0067)$	$[0.0067, 1)$	$(1, 148.25]$	$(148.25, \infty)$
F_1	HW & ML	H_0	0.0924	0.8094	0.0976	0.0006
		H_1	0.0003	0.1216	0.5818	0.2962
F_1, F_2	HW & ML	H_0	0.9024	0.0713	0.0255	0.0008
		H_1	0.00003	0.0145	0.2756	0.7099
F_1, F_2, F_3	HW & ML	H_0	0.9829	0.0108	0.0058	0.0004
		H_1	0.000006	0.0023	0.0861	0.9116
F_1, F_2, F_3	COA & ML	H_0	0.9832	0.0110	0.0053	0.0005
		H_1	0.000007	0.0023	0.0780	0.9197
F_1, F_2, F_3	HW & MMM	H_0	0.9648	0.0281	0.0067	0.0004
		H_1	0.00008	0.0038	0.0923	0.9039
F_1, F_2, F_3	COA & MMM	H_0	0.9667	0.0268	0.0062	0.0003
		H_1	0.00008	0.0036	0.0841	0.9123
F_1, F_2, F_3	UAF & ML	H_0	0.9832	0.0108	0.0056	0.0004
		H_1	0.000006	0.0023	0.0831	0.9146

between S and A, some doubts arose to the immigration authority about the ability of the system to treat the case with respect to the following hypotheses (graphical representation of which is given in Figure 6.5):

- H_0: S and A are not recently related;

- H_1: S and A are cousins.

At first, the sponsor's genetic evidence was typed on 13 loci since for two of them the typing had some laboratory problems. Later, the analysis was replicated to obtain all 15 loci included in the kit, and, finally, since it was not possible to get additional familial evidence, another kit was employed to include two more loci. The results are in Table 6.4. At first glance, the LR distributions appear

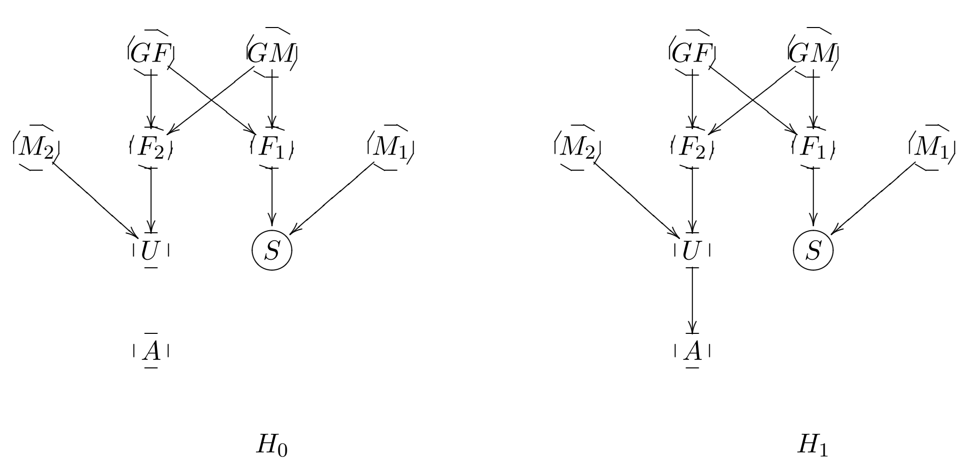

Figure 6.5: Two Cousins case - genotype networks under H_0 and H_1.

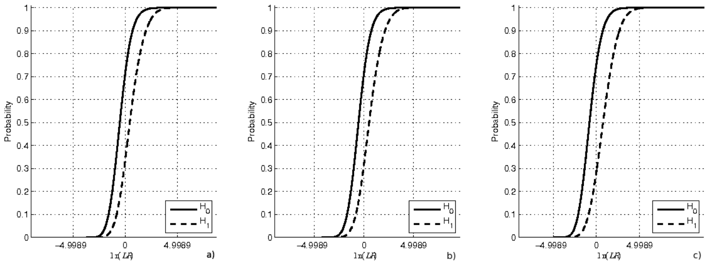

Figure 6.6: Two Cousins case - Tippett plots for $\ln(LR)$ distributions for different amounts of genetic data: a) 13 loci; b) 15 loci; c) 17 loci.

almost totally concentrated in the interval $[0.00676, 148.25]$, i.e. it is very rare to get strong support to both hypotheses when they actually hold. Moreover, there is a probability around 25% of getting weakly misleading results if H_0 holds, and around 30% if H_1 holds. The probabilities of these unpleasant events only slightly decrease if more loci are included, but they still persist at a level which seems too high. Pictorially, the result can be appreciated by looking at Figure 6.6, where, as the number of the loci employed increases, the distance between the curves becomes slightly larger, testifying to a small increase in the performances obtained.

Table 6.4: Two Cousins case - LR distributions apportioned for 13, 15, 17 loci. Hardy–Weinberg (HW) and Mendelian Segregation (ML) models are considered.

			LR			
Evidence	n. Loci		$[0, 0.0067)$	$[0.0067, 1)$	$(1, 148.25]$	$(148.25, \infty)$
S	13	H_0	0	0.7389	0.2611	0.00005
		H_1	0	0.3481	0.6409	0.011
S	15	H_0	0	0.7493	0.2507	0.00007
		H_1	0	0.3231	0.6607	0.0162
S	17	H_0	0	0.7761	0.2238	0.0001
		H_1	0	0.2911	0.6800	0.0289

A Stepwise case

A man, B (the candidate), would like to discover his father's identity. He has serious reasons to believe he is the son of AF (the Alleged Father), who died some years ago. To assess B's paternity consider that AF had a daughter (S) with his wife M (who is not the mother of B). To avoid the exhumation of AF, S requested that her own genetic profile should be used. Initially, since M did not provide her DNA profile, the case was addressed by evaluating the LR according to the hypotheses specified below:

- H_0: B and S do not share recent relatives;
- H_1: B is S's half brother.

Stated in this way the identification procedure clearly deals with one person, S, who wants to identify her half brother (U). This circumstance only indirectly implies they share the father. In other words, if the case were merely labelled as a "paternity test", the very indirect nature of the identification would be obscured. In this case only the baseline models, i.e. Hardy-Weinberg (HW) as population model and the Mendelian Laws (ML) as segregation models, are employed, not

Table 6.5: Stepwise case, first attempt - LR distributions apportioned under Hardy-Weinberg (HW) and Mendelian Segregation (ML) models.

		LR			
Evidence		$[0, 0.0067)$	$[0.0067, 1)$	$(1, 148.25]$	$(148.25, \infty)$
S	H_0	0.1647	0.7243	0.1102	0.0008
	H_1	0.0005	0.1146	0.6830	0.2019

allowing for mutation, coancestry or uncertainty in allele frequencies.

The first attempt (an Half Sibling case) To asses the system's potential, we derived the posterior distribution for B's genotypes according to H_0 and H_1. The graphical representations of the BNs able to derive the X_B distributions under H_0 and H_1 are in Figure 6.7.

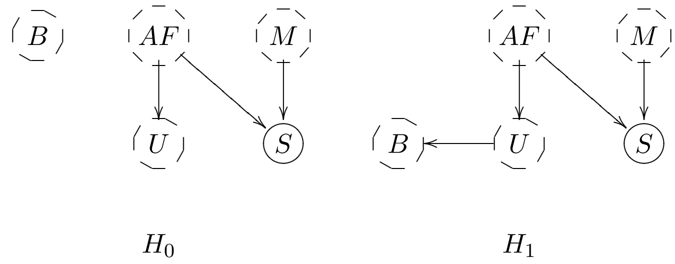

Figure 6.7: Stepwise case, first attempt - genotype networks according under H_0 and H_1 (only S's evidence available).

System's expected performance are quite unsatisfactory. In fact, looking at Table 6.5, whatever hypothesis is assumed, the system rarely can achieve a definite result, since the probability of weak evidence is about 0.80 (more precisely, $0.7243 + 0.1102 = 0.8345$). Moreover, if we consider the partition of the LR support into the sets $[0, 1)$ and $(1, +\infty)$, an even more embarrassing result arises. Now the probabilities to observe evidence not supporting H_1 or H_0, when they are actually true turn out to be greater than 11%.

Table 6.6: Stepwise case, second attempt - LR distributions apportioned under Hardy-Weinberg (HW) and Mendelian Segregation (ML) models.

Evidence		LR $[0, 0.0067)$	 $[0.0067, 1)$	 $(1, 148.25]$	 $(148.25, \infty)$
$S\&M$	H_0	0.3079	0.6183	0.7280	0.0010
	H_1	0.0008	0.0796	0.5165	0.4031

The second attempt (an enhanced Half Sibling case) Later, the system evaluation was replicated since M was convinced to provide her genetic profile. In fact, the BNs in Figure 6.8 only differ from those in Figure 6.7 for the observed node M. The results, provided in Table 6.6, make clear the benefits of such additional evidence: the probability to observe weak evidence now is reduced to 0.70 if H_0 holds, or to less than 0.60 if the identification hypothesis is assumed. Nevertheless, Table 6.6 shows that the probabilities to observe evidence *against* the a hypothesis, when it is in fact true, are still high, being around 8%.

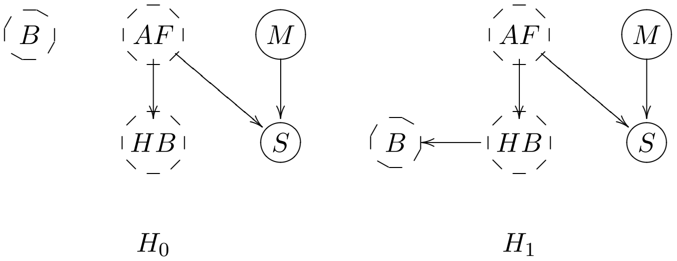

Figure 6.8: Stepwise case, second attempt - genotype networks under H_0 and H_1 (S and M evidences available).

After these attempts it seems sensible to suggest two ways to cope with the case.

- Make use of more extensive evidence including, if available, the genetic profiles of some more family members such as, for instance, the mother of B

Table 6.7: Stepwise case, third attempt - LR distributions apportioned under Hardy-Weinberg (HW) and Mendelian Segregation (ML) models.

		LR			
Evidence		$[0, 0.0067)$	$[0.0067, 1)$	$(1, 148.25]$	$(148.25, \infty)$
B	H_0	0.9997	0.000295	0.000001	0.000004
	H_1	0.000005	0.000001	0	0.999994

and/or extend the analysis to more loci, like in the Two Cousins case.

- If the previous suggestions cannot be undertaken, another possibility is to give up the ambition to evaluate the hypothesis of a common father for B and S without exhuming AF and plan for this latter possibility. Also in this case it should be useful to know the system expected performances in advance.

The third attempt (a Motherless Paternity case) Since it was not possible to follow the first route and since B made his genetic profile available, we considered the expected performance of the motherless identification system where AF assumes the role of the *candidate* to the identification of the father (U) of B, as is shown in Figure 6.9.

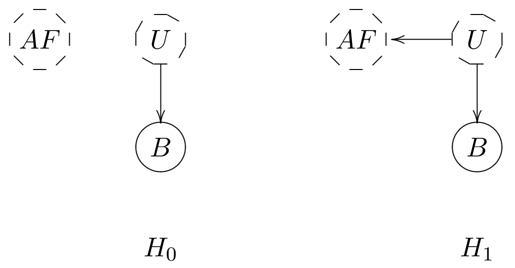

Figure 6.9: Stepwise case, third attempt - genotype networks under H_0 and H_1

The performance of this identification system is summarized in Table 6.7. It clearly shows that, now, B has the opportunity to carry on with the identification of his father quite safely. The only drawback is that, unfortunately, AF must be exhumed, but the analysis itself helps in motivating this choice. The results obtained in this last attempt are, in our opinion, an example of satisfactory performances.

6.3 Summary

The main contribution of this Chapter is to recognise the large variety of kinship identification systems related to specific cases. This implies the need to use different amounts of information to reach satisfactory standards of performance whose specification must be made clear to all the parties involved. Looking at the cases reviewed in Section 6.1, it seems reasonable to suggest that systems devoted to solve paternity cases, whether including maternal evidence or not, generally have a good level of performance, while other more indirect identification systems can be fruitfully evaluated with our proposed methodology.

In the second part of the Chapter we have presented a number of cases, analysing the performance of their identification systems: in some cases, to improve a system, additional genetic profiles from family members and/or an increase in the number of typed loci have been adopted. Furthermore, in some selected cases, we managed a sensitivity analysis on the models employed to evaluate the dependence of the results on such choices. Finally, results are mainly given by means of the probabilities that the LR belongs to a certain subset of values, but also graphical representations of the entire distributions using Tippett plots are given in some cases.

CHAPTER 7

Mutation rates estimation

This Chapter is one of the results of the period (April-July 2012) I spent working at the Netherlands Forensic Institute (NFI) in The Hague, The Netherlands. There I collaborated with people belonging to the Human Biologische Sporen (Human Biological Traces) team, especially with Klaas Slooten, in some projects on mutation rates and mutation models, with special attention to the application of these issues to Disaster Victim Identification (DVI) cases.

The results of this collaboration are a mid-term ongoing project on the DVI topic and a paper (Slooten and Ricciardi (2012)) on mutation rates estimation, the main aspects of which will be described here.

7.1 Introduction

We want to investigate the estimation of mutation probabilities for autosomal STR markers focusing on those currently employed in paternity testing and forensic DNA kinship analysis. It is well known that the estimation of these probabilities is made difficult by the fact that mutations do not always lead to a genetic inconsistency (i.e., the mutation is not visible from the observed genotypes) and since the parental origin is not always deducible by looking at the genotypes: it may become clear that a mutation must have happened in order to allow parenthood, but not from which parent.

Here we focus on a generalization of the first of the above phenomena, the hidden (sometimes also called covert) mutations. The fact that mutations may not lead to genetic inconsistencies is well-known: in Chakraborty et al. (1996) the probability to underestimate the true mutation rate due to this fact is derived algebraically. The authors show that the expected bias is a decreasing function of the number of alleles at a locus. In other words, the underestimation is smaller for the loci where mutation rate is higher and vice-versa. In a more recent work, Vicard and Dawid (2004) discerned and corrected a subtle error in the previous analysis. The probability to observe an incompatible triplet of genotypes for mother, father and child is thus:

$$pr(I) = \sum_{i=1}^{K} \mu_i (1 - p_i)^2 + \sum_{i \neq j} \mu_{i[j]} p_i p_j (1 - p_i - p_j)^2$$

where K is the length of the allelic ladder for the considered locus and $\mu_{i[j]}$ is the conditional probability of a (paternal) gene mutating into allele i, given that it is initially neither i nor j. In Appendix C we will give a different, and shorter, derivation of the latter formula. Brenner (2004) described the results of a simulation study of the same phenomenon, naming a mutation that does not lead to an inconsistency a *covert* mutation.

Mutations on the STR markers consider (almost) always the gain or the loss of a certain number of repeat units. If d such units have been lost or gained, we call the mutation a d-step mutation and d is the distance of the mutation. It is also acknowledged that the, by far, most common mutations are 1-step mutations, and that mutations become less likely as the mutational distance increases. However, if a d-step mutation has occurred from (say) father to child, then by looking at the genotypes of father, mother and child, it may be possible to explain these genotypes by a k-step mutation with $k < d$ (if $k = 0$, then the mutation is hidden). For example, suppose a confirmed father's genotype is (11,12) whereas the child has (9,10), then there are four mutational events that could had happened: a one-step mutation $11 \to 10$, a two-step mutation $11 \to 9$ or $12 \to 10$, and a three-step mutation $12 \to 9$. Consistently classifying mutations as the shortest possible one (in this case $11 \to 10$) leads to a bias towards shorter mutations and our goal is to investigate the magnitude of this bias.

We should point out here that these computations should serve, in our opinion, to improve estimates of mutation probabilities per locus and distance, prior to plugging-in these estimates into a mutation model that is intended to fully model the mutation process, as is done for example in Vicard and Dawid (2004) and Vicard et al. (2008). Therefore the results obtained here can be viewed as complementary to those papers.

7.2 Mutation models

First, we introduce some notation that we will use throughout. Since we do not need to deal with several loci simultaneously, we will omit notation that refers to a specific locus and assume that we have chosen one, for which an allelic ladder $\mathcal{L}$ is known. Furthermore, we will distinguish between calculations in which we do not, and do allow off-ladder alleles to be created, but we will only consider mutations that consist out of a loss or gain of an integer number of repeat units. We call a mutation model where alleles mutate only inside $\mathcal{L}$ a *restricted* model, and a model in which alleles can mutate to alleles outside of $\mathcal{L}$ an *unrestricted* model and we will consider these two types of models for our analysis. Note that these are auxiliary models to enable the study of mutations of a fixed k-step distance, they do not represent the actual mutational process if taken by itself.

7.2.1 Restricted k-step model

According to this model, only mutations consisting out of exactly k steps are possible. Since we are defining a restricted model, not all alleles can mutate: they need to have a k-step neighbour in $\mathcal{L}$. Here by $\mu_{i,j}$ we denote the probability that allele i mutates into allele j. Let $0 \leq \mu \leq 1$. We let, for $\epsilon \in \{-1, 1\}$,

$$\mu_{i,i+\epsilon k} = \begin{cases} \mu/2 & \text{if } \{i-k, i+k\} \subset \mathcal{L}, \\ \mu & \text{if } i+\epsilon k \in \mathcal{L}, i-\epsilon k \notin \mathcal{L}, \\ 0 & \text{if } \mathcal{L} \cap \{i-k, i+k\} = \emptyset, \end{cases}$$

all other $\mu_{i,j} = 0$ for $i \neq j$, and $\mu_{i,i} = 1 - \sum_{j \neq i} \mu_{i,j}$. In particular, if $\mathcal{L} \cap \{i-k, i+k\} = \emptyset$ then $\mu_{i,i} = 1$, meaning that allele i cannot mutate.

7.2.2 Unrestricted k-step models

To allow for mutations of distance k, we let

$$\tilde{\mathcal{L}}^{(\pm k)} = \mathcal{L} \cup \{\{a+k, a-k\} \mid a \in \mathcal{L}\}$$

be the extension of $\mathcal{L}$ with all alleles at distance k from the alleles in $\mathcal{L}$. For example, for the locus D2S441 we have used as allelic ladder

$$\mathcal{L} = \{8, 9, 10, 11, 11.3, 12, 12.3, 13, 13.3, 14, 14.1, 15, 16\},$$

so the extended ladders are, for $k = \{1, 2\}$

$$\tilde{\mathcal{L}}^{(\pm 1)} = \mathcal{L} \cup \{7, 10.3, 13.1, 14.3, 15, 15.1, 17\},$$

and

$$\tilde{\mathcal{L}}^{(\pm 2)} = \mathcal{L} \cup \{6, 7, 9.3, 10.3, 12.1, 14.3, 15.3, 16.1, 17, 18\}.$$

According to this model,

$$\mu_{i,j} = \begin{cases} \mu/2 & \text{if } |i-j| = k, \\ 0 & \text{otherwise,} \end{cases}$$

for $i \in \mathcal{L}, j \in \tilde{\mathcal{L}}^{(\pm k)}$. Note that these unrestricted k-step models can only serve to model mutation from one generation to the next one. In case more generations are involved, the ladder would have to be expanded further.

7.2.3 Note on the mutation rate

The mutation rate on the locus is the probability that a randomly selected population allele on that locus will mutate. It is given by

$$\sum_{i \in L} \sum_{j \neq i} p_i \mu_{i,j} = \sum_i p_i (1 - \mu_{i,i}),$$

where the sum runs over all possible alleles on the locus under consideration, and p_i is the population frequency of allele i.

Thus, the mutation rate of the unrestricted k-step models is equal to μ, but the mutation rate of the restricted k-step models with the same parameter is smaller if there are alleles that cannot mutate because they have no k-step neighbour on the ladder.

7.3 Mutation rates per mutational distance

We are interested in the difference between a mutation's apparent distance and its actual distance. Suppose that a d-step mutation has occurred from one of the parents to the child (we do not, in view of mutation probabilities being very small, consider the possibility of several mutations on the same locus), then the smallest k such that there exists a k-step mutation from (at least) one of the parents to the child that explains the genotypes is what we call the apparent distance. We do not specify yet whether we are looking at parent-child duo's[1] or father-mother-child trio's[2]: in both cases the apparent length of any mutation is well-defined.

Example 7.3.1. *Suppose that the mother's genotype is (13,15), the father's genotype is (14,15) and the child's genotype is (12,15). Then, the genotypes can be ex-*

[1] only one parent and the child are observed

[2] father, mother and child's genetic evidences are available

plained either by a maternal mutation $13 \to 12$, *or by a paternal mutation* $14 \to 12$. *Thus, the apparent distance of this mutation is one.*

If the father's genotype would have been (11, 15), then either a mutation $13 \to 12$ *or* $11 \to 12$ *has happened. This mutation also has apparent distance one, but now both the paternal and the maternal possible mutation of distance one are able of explaining the child genotype.*

In view of the different mutation rates between men and women, our goal is to estimate mutation rates per mutational distance and per gender. First, in Section 7.4, we will show how to take into account that the apparent mutational distance is possibly shorter than the actual mutational distance. Then, in Section 7.5, we will discuss how to use this result to determine the required mutation probabilities.

Suppose that μ_k^* is the observed frequency of mutations of apparent distance k. We let $A_{k,l}$ be the probability that a k-step mutation has apparent distance l. If the actual probability of a k-step mutation is μ_k, the relation between the μ_k and the μ_k^* can be conveniently summarized by the matrix equation

$$\begin{pmatrix} \mu_0^* \\ \mu_1^* \\ \mu_2^* \\ \vdots \\ \mu_k^* \end{pmatrix} = \begin{pmatrix} 1 & A_{1,0} & A_{2,0} & \dots & A_{k,0} \\ 0 & A_{1,1} & A_{2,1} & \dots & A_{k,1} \\ 0 & 0 & A_{2,2} & \dots & A_{k,2} \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & A_{k,k} \end{pmatrix} \begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}$$

which we may write more compactly as

$$\vec{\mu}^* = A \cdot \vec{\mu},$$

or equivalently (since A is invertible) as

$$\vec{\mu} = A^{-1}\vec{\mu}^*. \tag{7.1}$$

If $\mu_k^* = 0$ for all $k \geq n_0$, then so are the corresponding μ_i. Thus, knowledge of all $A_{k,l}$ for $k < n_0$ is sufficient.

7.4 Computation of entries $A_{k,l}$

The entries $A_{k,0}$ expressing the probability that a k-step mutation does not lead to an inconsistent genotype, can be easily computed algebraically. As stated in Section 7.1, this is done by Chakraborty et al. (1996) and Vicard and Dawid (2004) and in Appendix C we shortly derive it in a different, but more intuitive, way. The formula for parent-child duo's is obtained along similar lines, and it is not detailed. For $l > 0$, we estimated the $A_{k,l}$ by computer simulations.

7.4.1 Method

For each actual mutational distance $k \in \{1, 2, 3\}$ we selected 100,000 times a father's genotype and a mother's genotype, drawn at random according to the allele frequencies observed in the NFI reference database (containing the genotypes of 2085 individuals). For each couple we created a child that inherited its maternal allele without mutation, but applied a k-step mutation to the paternally inherited allele. We then calculated the apparent distance of the mutation both for the father-child and for the father-mother-child pedigree. This simulation was carried out separately for the restricted and the unrestricted k-step models.

Table 7.1: Probabilities $A_{k,l}$ of apparent mutational distance l for a k-step mutation, unrestricted k-step model (for $k = 1, 2, 3$), for parent-child duo's.

Locus	$A_{1,0}$	$A_{1,1}$	$A_{2,0}$	$A_{2,1}$	$A_{2,2}$	$A_{3,0}$	$A_{3,1}$	$A_{3,2}$	$A_{3,3}$
D1S1656	0.25	0.75	0.23	0.31	0.46	0.22	0.28	0.19	0.31
TPOX	0.61	0.39	0.6	0.21	0.19	0.61	0.17	0.09	0.14
D2S441	0.48	0.52	0.44	0.3	0.25	0.46	0.25	0.09	0.2
D2S1338	0.29	0.71	0.28	0.34	0.38	0.27	0.31	0.19	0.22
D3S1358	0.47	0.53	0.44	0.4	0.15	0.41	0.4	0.15	0.04
FGA	0.34	0.66	0.32	0.39	0.29	0.3	0.37	0.19	0.13
D5S818	0.58	0.42	0.55	0.37	0.08	0.52	0.39	0.08	0.02
CSF1PO	0.56	0.44	0.52	0.38	0.1	0.49	0.39	0.1	0.02
SE33	0.14	0.86	0.14	0.23	0.63	0.13	0.22	0.17	0.48
D7S820	0.42	0.58	0.41	0.4	0.19	0.38	0.39	0.17	0.06
D8S1179	0.42	0.58	0.4	0.39	0.21	0.38	0.38	0.16	0.08
D10S1248	0.52	0.48	0.49	0.39	0.12	0.46	0.39	0.12	0.03
TH01	0.45	0.55	0.43	0.24	0.33	0.42	0.22	0.13	0.23
VWA	0.43	0.57	0.41	0.4	0.19	0.38	0.39	0.17	0.06
D12S391	0.27	0.73	0.26	0.35	0.39	0.25	0.32	0.2	0.22
D13S317	0.45	0.55	0.43	0.38	0.19	0.41	0.36	0.14	0.08
PENTA E	0.26	0.74	0.26	0.26	0.48	0.25	0.25	0.2	0.29
D16S539	0.48	0.52	0.47	0.37	0.17	0.44	0.37	0.15	0.05
D18S51	0.32	0.68	0.3	0.38	0.32	0.29	0.36	0.2	0.15
D19S433	0.47	0.53	0.43	0.38	0.19	0.4	0.37	0.13	0.09
PENTA D	0.41	0.59	0.39	0.38	0.23	0.37	0.35	0.17	0.11
D21S11	0.37	0.63	0.34	0.37	0.29	0.31	0.36	0.18	0.16
D22S1045	0.52	0.48	0.49	0.35	0.16	0.47	0.34	0.09	0.1

7.4.2 Results

Now we present the obtained entries $A_{k,l}$ (with $1 \leq k \leq 3$) for the various models that we have considered. Results show significative differences according to different amount of data considered, thus we classified them following this scheme.

- **Duo's** For the unrestricted and restricted k-step models, the results are summarized in Table 7.1 and Table 7.2 respectively.

Table 7.2: Probabilities $A_{k,l}$ of apparent mutational distance l for a k-step mutation, restricted k-step model (for $k = 1, 2, 3$), for parent-child duo's.

Locus	$A_{1,0}$	$A_{1,1}$	$A_{2,0}$	$A_{2,1}$	$A_{2,2}$	$A_{3,0}$	$A_{3,1}$	$A_{3,2}$	$A_{3,3}$
D1S1656	0.25	0.75	0.24	0.31	0.45	0.23	0.3	0.2	0.27
TPOX	0.61	0.39	0.61	0.21	0.18	0.63	0.14	0.08	0.14
D2S441	0.48	0.52	0.45	0.31	0.24	0.49	0.24	0.1	0.17
D2S1338	0.29	0.71	0.28	0.34	0.38	0.28	0.31	0.2	0.22
D3S1358	0.47	0.53	0.45	0.41	0.14	0.43	0.38	0.14	0.04
FGA	0.33	0.67	0.32	0.4	0.28	0.31	0.38	0.19	0.12
D5S818	0.58	0.42	0.55	0.37	0.08	0.52	0.38	0.08	0.01
CSF1PO	0.55	0.45	0.52	0.37	0.1	0.49	0.39	0.1	0.02
SE33	0.14	0.86	0.14	0.23	0.63	0.13	0.22	0.17	0.48
D7S820	0.43	0.57	0.42	0.4	0.18	0.4	0.38	0.16	0.06
D8S1179	0.42	0.58	0.41	0.39	0.2	0.4	0.37	0.15	0.07
D10S1248	0.52	0.48	0.49	0.39	0.12	0.46	0.39	0.12	0.03
TH01	0.45	0.55	0.44	0.32	0.24	0.43	0.27	0.16	0.14
VWA	0.43	0.57	0.41	0.4	0.19	0.4	0.39	0.16	0.05
D12S391	0.27	0.73	0.26	0.35	0.39	0.26	0.33	0.2	0.21
D13S317	0.45	0.55	0.43	0.39	0.18	0.44	0.35	0.13	0.07
PENTA E	0.26	0.74	0.26	0.26	0.48	0.25	0.26	0.21	0.28
D16S539	0.48	0.52	0.46	0.37	0.17	0.45	0.37	0.14	0.04
D18S51	0.31	0.69	0.31	0.38	0.31	0.29	0.35	0.2	0.15
D19S433	0.46	0.54	0.43	0.38	0.19	0.41	0.38	0.13	0.08
PENTA D	0.4	0.6	0.38	0.38	0.23	0.37	0.36	0.17	0.11
D21S11	0.37	0.63	0.34	0.37	0.28	0.33	0.36	0.18	0.14
D22S1045	0.52	0.48	0.49	0.36	0.15	0.48	0.34	0.08	0.09

Table 7.3: Probabilities $A_{k,l}$ of apparent mutational distance l for a k-step mutation, unrestricted k-step model (for $k = 1, 2, 3$), for trio's.

Locus	$A_{1,0}$	$A_{1,1}$	$A_{2,0}$	$A_{2,1}$	$A_{2,2}$	$A_{3,0}$	$A_{3,1}$	$A_{3,2}$	$A_{3,3}$
D1S1656	0.09	0.91	0.07	0.14	0.79	0.05	0.1	0.11	0.74
TPOX	0.12	0.88	0.08	0.27	0.65	0.18	0.1	0.1	0.62
D2S441	0.14	0.86	0.05	0.27	0.69	0.14	0.13	0.13	0.61
D2S1338	0.1	0.9	0.08	0.19	0.73	0.08	0.14	0.15	0.63
D3S1358	0.22	0.78	0.16	0.29	0.55	0.09	0.18	0.19	0.54
FGA	0.15	0.85	0.13	0.24	0.63	0.1	0.19	0.16	0.55
D5S818	0.28	0.72	0.13	0.29	0.58	0.03	0.14	0.24	0.59
CSF1PO	0.27	0.73	0.15	0.28	0.57	0.04	0.16	0.23	0.58
SE33	0.05	0.95	0.05	0.1	0.86	0.04	0.08	0.08	0.8
D7S820	0.2	0.8	0.16	0.28	0.57	0.09	0.19	0.18	0.54
D8S1179	0.19	0.81	0.14	0.27	0.6	0.09	0.18	0.18	0.55
D10S1248	0.25	0.75	0.15	0.28	0.57	0.07	0.16	0.21	0.56
TH01	0.1	0.9	0.07	0.14	0.8	0.04	0.07	0.09	0.8
VWA	0.21	0.79	0.15	0.28	0.57	0.09	0.19	0.18	0.54
D12S391	0.11	0.89	0.1	0.2	0.7	0.08	0.17	0.14	0.61
D13S317	0.2	0.8	0.12	0.28	0.6	0.09	0.18	0.18	0.55
PENTA E	0.08	0.92	0.08	0.14	0.78	0.06	0.14	0.15	0.65
D16S539	0.21	0.79	0.16	0.27	0.57	0.08	0.19	0.18	0.55
D18S51	0.13	0.87	0.12	0.23	0.65	0.09	0.19	0.16	0.56
D19S433	0.22	0.78	0.11	0.23	0.65	0.04	0.12	0.19	0.66
PENTA D	0.17	0.83	0.13	0.27	0.6	0.11	0.19	0.16	0.54
D21S11	0.16	0.84	0.1	0.19	0.71	0.04	0.11	0.14	0.71
D22S1045	0.22	0.78	0.07	0.24	0.69	0.03	0.14	0.25	0.59

- **Trio's** For the unrestricted and restricted k-step models, the results are summarized in Table 7.3 and Table 7.4 respectively.

7.4.3 Discussion on the obtained $A_{k,l}$

The tables indicate that, especially for large mutational distances, the apparent distance is very often smaller than the actual distance, especially for duo's but also for trio's. Within each class (duo's or trio's) the effect is most notable for the restricted model. This is not surprising, since in the unrestricted model alleles can

Table 7.4: Probabilities $A_{k,l}$ of apparent mutational distance l for a k-step mutation, restricted k-step model (for $k = 1, 2, 3$), for trio's.

Locus	$A_{1,0}$	$A_{1,1}$	$A_{2,0}$	$A_{2,1}$	$A_{2,2}$	$A_{3,0}$	$A_{3,1}$	$A_{3,2}$	$A_{3,3}$
D1S1656	0.09	0.91	0.08	0.16	0.76	0.07	0.15	0.13	0.65
TPOX	0.12	0.88	0.11	0.41	0.48	0.37	0.19	0.1	0.33
D2S441	0.14	0.86	0.05	0.28	0.67	0.21	0.23	0.18	0.38
D2S1338	0.1	0.9	0.08	0.18	0.73	0.08	0.15	0.16	0.61
D3S1358	0.23	0.77	0.17	0.32	0.51	0.14	0.24	0.21	0.41
FGA	0.15	0.85	0.12	0.25	0.63	0.11	0.22	0.17	0.5
D5S818	0.28	0.72	0.13	0.29	0.58	0.04	0.15	0.24	0.57
CSF1PO	0.27	0.73	0.16	0.26	0.58	0.05	0.16	0.23	0.56
SE33	0.05	0.95	0.05	0.1	0.85	0.04	0.08	0.08	0.8
D7S820	0.2	0.8	0.18	0.31	0.5	0.13	0.27	0.21	0.39
D8S1179	0.19	0.81	0.14	0.28	0.58	0.12	0.24	0.2	0.44
D10S1248	0.25	0.75	0.15	0.29	0.56	0.08	0.17	0.22	0.53
TH01	0.1	0.9	0.15	0.31	0.54	0.12	0.21	0.18	0.49
VWA	0.21	0.79	0.15	0.28	0.56	0.12	0.23	0.2	0.44
D12S391	0.11	0.89	0.1	0.22	0.69	0.1	0.18	0.15	0.57
D13S317	0.2	0.8	0.13	0.31	0.56	0.14	0.24	0.22	0.41
PENTA E	0.08	0.92	0.09	0.14	0.77	0.07	0.15	0.17	0.61
D16S539	0.21	0.79	0.16	0.27	0.57	0.11	0.23	0.19	0.47
D18S51	0.13	0.87	0.12	0.22	0.65	0.1	0.19	0.16	0.54
D19S433	0.22	0.78	0.11	0.25	0.64	0.05	0.15	0.2	0.6
PENTA D	0.17	0.83	0.12	0.27	0.61	0.1	0.19	0.17	0.54
D21S11	0.16	0.84	0.11	0.2	0.69	0.06	0.14	0.16	0.64
D22S1045	0.22	0.78	0.07	0.25	0.68	0.04	0.18	0.29	0.49

mutate into a new, off-ladder, allele that is smaller (or larger) than all of the alleles on the ladder. Such mutations are more likely to have their apparent distance to be equal to the actual distance than mutations that are forced to stay on the ladder. The biological truth will, in our opinion, be in between the two models. Indeed, on the one hand one may argue that alleles do not have knowledge of the ladder as we know it, and that therefore the unrestricted model is appropriate. On the other hand, the fact that alleles have not been observed in a large reference sample (recall that we have taken allele frequencies from a reference database containing 2085 genotypes), means that alleles not included in the ladder are rare, and there may be a biological mechanism that prevents their existence.

Note also that the differences between loci are quite substantial: considering for example the upper left corner of Table 7.2, we see that $A_{1,0} = 0.25$ for locus D1S1656 and $A_{1,0} = 0.61$ for locus TPOX. This is explained by the greater polymorphism of the former locus. In general, largely polymorphic loci often reveal the actual length as the apparent length (as can also be seen from the Tables) since the alleles that have not mutated in reality have a greater probability of being more distant from the mutated allele.

Sampling uncertainty

Each of the entries $A_{k,l}$ (for $l > 0$) has a sampling uncertainty that we can estimate using the normal approximation. In this case, for $n = 100,000$ trials we obtain that the 95% confidence interval has half-width equal to up to 0.003 for the probabilities close to 0.5, and down to about 0.001 for the probabilities down to 0.05.

7.5 Corrected mutation rates

We will now use the obtained estimates of the $A_{k,l}$ to evaluate the effect of the correction for apparent mutational distance.

7.5.1 Duo's

If mutation rates are estimated from parent-child duo's, it is clear from Tables 7.1 and 7.2 that the μ_i^* are much smaller than the μ_i for $i \geq 1$. We do not have a data set at our disposal to determine the μ_i^* from, therefore we restrict ourselves to the exposition of one of the more dramatic examples, namely the locus D5S818 (using results from the unrestricted models). In that case, when mutational distances up to 3 are considered, the matrix inversion (7.1) yields $(\mu_0, \mu_1, \mu_2, \mu_3) =$

$$(\mu_0^* - 1.37\mu_1^* - 0.56\mu_2^* + 3.73\mu_3^*, 2.37\mu_1^* - 10.02\mu_2^* - 3.86\mu_3^*, 11.57\mu_2^* - 63.01\mu_3^*, 64.14\mu_3^*)$$

from which we observe again that high-distance mutations are much more frequent than they appear to be. Since $\mu_k^* << \mu_{k-1}^*$, we see for example that 2-step mutations are in the order of ten times as frequent as they appear to be in parent-child duo's.

7.5.2 Trio's

The situation is more complex now, compared to duo's, since the apparent length of a mutation may be smaller than the actual length, but it may also be ascribed to the other parent.

Referring to Example 7.3.1, suppose that the mother has genotype $(13, 15)$, the father has genotype $(14, 15)$ and the child has genotype $(12, 15)$. Then either a maternal mutation 13 → 12 or a paternal mutation 14 → 12 could had occurred (recall that we only consider a single mutation per locus as possibilities). If indeed the mutation that has occurred was the paternal 14 → 12, then it is counted as a mutation of apparent distance 1, but its apparent distance is obtained from a possible mutation that is in reality from the other parent than the one that did have the mutation.

This shows that when observed frequencies of apparent mutational distances are recorded in trio's, they do not immediately allow us to determine gender-specific mutation rates per distance. In order to be able to do so, we make the following assumption. Let $\mu_{1,k}$ be the mutation frequency for paternal k-step mutations and $\mu_{2,k}$ that for women, then we assume that

$$\frac{\mu_{1,k}}{\mu_{2,k}} = m, \tag{7.2}$$

independently of k for all $k \geq 1$.

Now we can count apparent distances of all mutations in trio's, and this will give us μ_k^*, the apparent mutation frequency of k-step mutations, from either parental lineage. From the μ_k^* we can obtain $\mu_k = \mu_{1,k} + \mu_{2,k} = (m+1)\mu_{1,k}$ by applying (7.1). Finally, the factor m can be estimated by only considering the non-indeterminate mutations: the ones where the parental origin is clear. This conclusion has been drawn already in Vicard and Dawid (2004), but we present a simple way to arrive at this result for completeness.

In view of (7.2), the probability with which a paternal mutation is indeterminate is the same as the one for a maternal mutation, and we denote it by α_I. If we do not assume (7.2) then this need not be true. Now, let $\mu^i = 1 - \mu_{i,0} = \mu_{i,1} + \mu_{i,2} + \ldots$ be the mutation rate for gender i, and let $\mu^{i,NI}$ be the apparent mutation frequency for gender i of non-indeterminate mutations (i.e., where a mutation is visible and only a mutation of the parent of gender i can explain the genotypes). Since $\mu^{i,NI} = \mu^i(1-\alpha_I)\gamma$, where γ is the (again, gender-independent) probability that a non-indeterminate mutation has positive apparent distance, we have

$$\frac{\mu^{1,NI}}{\mu^{2,NI}} = \frac{\mu^1}{\mu^2}.$$

This implies that we can estimate m by

$$m = \frac{\mu^{1,NI}}{\mu^{2,NI}}.$$

7.5.3 Accounting for apparent versus actual distance based on AABB data

The AABB regularly publishes data on mutations in their Annual Report Summaries for Testing (e.g., AABB (2003), AABB (2008)). These reports contain counts of apparent mutations as well as counts of indeterminate mutations. It is however not clear to us what the definition of an indeterminate mutation is in these reports, nor if mutations are counted according to apparent (in our definition) length. For the purpose of illustrating our methodology, we have therefore decided to combine the following data:

- Data in the 2003 report (Appendix 3 of AABB (2003)) on the distance of the obligatory allele, which we have interpreted to contain apparent mutational distances. These data distinguish between mutations of distance +1, -1, +2, -2 and other. We have merged the +1 and -1 data and similarly for distance 2. Moreover for simplicity, we have considered all the other mutations (if observed) as 3-step mutations. We consider the Appendix to contain the

$\mu^*_{1,k}/(1-\mu^*_{1,0})$ (observed paternal mutation rate for apparent distance k, given that there is mutation with positive apparent distance) and $\mu^*_{2,k}/(1-\mu^*_{2,0})$ (similarly for maternal mutations). We have only used the loci for which more than 50 mutations had been counted (thereby excluding TPOX, TH01, PENTA D, PENTA E) and for which not only one-step mutations had been recorded (thereby excluding D19S433);

- Appendix 1 in the 2008 report (AABB (2008)) on the apparent paternal and maternal mutation rates, which we have interpreted as $1-\mu^*_{1,0}$ (paternal) and $1-\mu^*_{2,0}$ (maternal).

We have combined these data to define the vectors $(\mu^*_{i,0}, \mu^*_{i,1}, \mu^*_{i,2}, \mu^*_{i,3})$ for the loci that we have allele frequencies for. The results, for paternal mutations, are summarized in Tables 7.5 and 7.6 below. To ease notation we have omitted in these tables the subscripts that refer to the paternal gender, e.g., we write μ_k instead of $\mu_{1,k}$. In all calculations we have used the entries $A_{k,l}$ from Table 7.4, i.e., from the restricted model for trio's.

7.5.4 Discussion

From Table 7.5, we see first of all that apparent mutation rates underestimate the actual mutation rates substantially, as was already well known: this is due to mutations being hidden. For one-step mutations, we see that their frequency is underestimated with about the same factor as the mutation rate. For mutations of distance two and three, we see that their frequency is underestimated by more than this factor: by about a factor two for 3-steps mutations. From Table 7.6 we see that, as a result of this, the ratio of 1-step versus 2-steps mutations is in general lower in reality than it is when apparent mutational distances are considered. In

Table 7.5: Effect of correcting for apparent mutational distance for paternal mutation rates, per mutational distance

Locus	μ/μ^*	$\frac{\mu_1}{\mu_1^*}$	$\frac{\mu_2}{\mu_2^*}$	$\frac{\mu_3}{\mu_3^*}$
D2S1338	1.11	1.11	1.36	N/A
D3S1358	1.29	1.28	1.98	N/A
FGA	1.18	1.16	1.14	1.98
D5S818	1.4	1.39	1.72	N/A
CSF1PO	1.37	1.37	1.75	N/A
D7S820	1.26	1.25	1.99	N/A
D8S1179	1.24	1.23	1.36	2.27
VWA	1.26	1.25	1.67	2.29
D13S317	1.25	1.24	1.82	N/A
D16S539	1.27	1.27	1.77	N/A
D18S51	1.16	1.15	1.17	1.83
D21S11	1.18	1.18	0.75	1.59

Table 7.6: Effect of correcting for apparent mutational distance for paternal mutation rates: relative frequency of smaller mutations versus longer mutations

Locus	$\frac{\mu_1/\mu_2}{\mu_1^*/\mu_2^*}$	$\frac{\mu_1/\mu_3}{\mu_1^*/\mu_3^*}$	$\frac{\mu_2/\mu_3}{\mu_2^*/\mu_3^*}$
D2S1338	0.82	N/A	N/A
D3S1358	0.65	N/A	N/A
FGA	1.02	0.59	0.57
D5S818	0.81	N/A	N/A
CSF1PO	0.78	N/A	N/A
D7S820	0.63	N/A	N/A
D8S1179	0.91	0.54	0.6
VWA	0.75	0.55	0.73
D13S317	0.68	N/A	N/A
D16S539	1.07	0.6	0.56
D18S51	0.98	0.63	0.64
D21S11	1.58	0.74	0.47

other words, 2-steps mutations are relatively more frequent with respect to 1-step mutations than the apparent distances suggest; and the same is true for other k versus l-steps mutations where $k < l$. Note also that for the locus D21S11, the opposite effect seems to occur for 1-versus 2-steps mutations, we believe that this may be due to the fact that $\mu_3^* > \mu_2^*$ for this locus, which suggests that it would be better to refine the analysis extending it to more than 3-steps mutations. Indeed, our assumption that all mutations that are neither one nor 2-steps mutations are 3-steps mutations, may not be correct. As previously remarked, in the absence of more precise data the results here are included only for the purpose of illustrating the methodology.

Remark also that the AABB data, for many loci, do not contain any mutations of apparent distance greater than two. A larger data set would be required in order to be able to estimate μ_3 by means of (7.1); alternatively, the available data could be fitted into a model that makes predictions on the rates of mutations of larger distances.

7.5.5 Sampling uncertainty

The data presented in Tables 7.5 and 7.6 are subjected to uncertainty that can be divided into three sources: uncertainty regarding the entries $A_{k,l}$, regarding $\mu_{i,0}^*$ (i.e., the mutation frequency) and uncertainty regarding the $\mu_{i,k}^*$ (i.e., regarding the specific mutation probabilities for distance k and gender i). As previously noted, the sampling uncertainty for the $A_{k,l}$ is very small and we will hence treat it as negligible. The apparent mutation frequencies $1 - \mu_{i,0}^*$ are based on large numbers of meioses (typically, more than 100,000). They may be subjected to a bias which is greater than their sampling uncertainty, but without information on this possibility we will simply treat the $\mu_{i,0}^*$ as known. The distance-specific $\mu_{i,k}^*$ for $k > 0$ however, are what we are interested in this work and hence we will focus on them. They are based on much smaller samples (ranging, for the loci in Tables 7.1 - 7.4, from 81 for D2S1338 to 663 for FGA for paternal mutations in AABB (2003)) and we have investigated the uncertainty that this implies on the estimates for the μ_i, by bootstrapping. If, in (AABB, 2003, Appendix 3), n_l mutations had been observed with $n_{l,i}$ of these of apparent distance i, we sampled 10.000 times a bootstrapped sample of mutations (i.e., we sampled n_l mutations, with replacement, from the reported set). We then recalculated μ/μ^* and μ_1/μ_1^*. For longer mutations, it is essential to know whether or not all of the mutations that are neither one nor two-steps mutations were indeed 3-steps mutations. In the absence of such knowledge, the figures μ_2 and μ_3 are purely illustrative and we have therefore concentrated on the most reliable inferences. In view of the limited available amount of data, it is hard to assess the sampling uncertainty on μ_2/μ_2^* without making additional assumptions on the mutation process.

The results are displayed in Table 7.7. We can observe that estimates μ/μ^* and μ_1/μ_1^* seem to be quite insensitive to sampling uncertainty in the μ_i^*.

Table 7.7: Mean and confidence interval of actual versus apparent mutation rates, obtained by bootstrapping data from (AABB, 2003, Appendix 3)

Locus	$\frac{\mu}{\mu^*}$	95% CI	$\frac{\mu_1}{\mu_1^*}$	95% CI
D2S1338	1.115	(1.113,1.116)	1.108	(1.097,1.116)
D3S1358	1.288	(1.286,1.290)	1.278	(1.262,1.290)
FGA	1.176	(1.175,1.177)	1.163	(1.157,1.169)
D5S818	1.396	(1.388,1.402)	1.390	(1.374,1.402)
CSF1PO	1.372	(1.367,1.375)	1.368	(1.357,1.375)
D7S820	1.257	(1.256,1.257)	1.252	(1.243,1.257)
D8S1179	1.238	(1.236,1.24)	1.231	(1.221,1.238)
VWA	1.257	(1.256,1.258)	1.250	(1.242,1.256)
D13S317	1.246	(1.244,1.248)	1.241	(1.232,1.248)
D16S539	1.274	(1.271,1.275)	1.267	(1.254,1.275)
D18S51	1.155	(1.154,1.156)	1.148	(1.143,1.153)
D21S11	1.184	(1.180,1.186)	1.179	(1.172,1.184)

7.6 Discussion

As we have seen, mutations of greater mutational distance are more common in reality than those estimated from apparent mutational distances. In trio's, mutations consisting of two or three steps appear to be about twice as likely as compared to apparent mutational distances in trio's. In estimation of mutation rates per distance from parent-child duo's, the effect is much more dramatic.

When a likelihood ratio needs to be calculated for a locus that involves an apparent mutation, it can be underestimated when using apparent mutational distance frequencies, especially when 1-step mutations cannot explain the genotypes. This is of course an undesirable situation, and we believe that when precise likelihood calculations are required, the corrections proposed in this article should be taken into account.

From section 7.5 it is clear that it is conceptually easier to estimate gender-specific and distance-specific mutation rates in duo's than in trio's. Indeed, in that case there is no need to make an assumption such as (7.2), since no mutations from the other parent are considered. However, in duo's many more mutations are hidden (i.e. have apparent distance zero) or have a shorter apparent than actual length compared to those in trio's. This means that more data are needed in order to obtain reliable estimates of the μ_k^* for $k > 1$ when using data from duo's than when using data from trio's.

Thus, as it is completely logical, genotypes of trio's are more informative about mutation frequencies than genotypes of duo's. The drawback is that mutations cannot directly be attributed to a parental lineage. We have made assumption (7.2) in order to facilitate computations, and one important consequence is that mutations are indeterminate with the same probability in either parental lineage.

This last fact is well known. In Vicard and Dawid (2004) it is also assumed that men and women mutate according to the same model with different mutation rates. Assumption (7.2) is frequently made in the literature if a non-zero female mutation rate is considered.

Therefore, the difference between male and female mutation rates can be reliably estimated by the non-indeterminate mutations. For distance and gender specific rates however, a precise analysis becomes more complex, requires a large data set, and we have refrained from carrying it out.

Finally, we have chosen not to distinguish between mutations that gain or lose the same number of repeat units. This may not be entirely realistic, and one may consider undesirable to make that distinction. The adaptation of the method is straightforward: one needs to choose an ordering on the possible distances, e.g., $0 \prec -1 \prec 1 \prec -2 \prec 2 \prec \ldots$ and then define the apparent distance of a mutation as the lowest possible according to this ordering. This will lead to a similar matrix $A_{k,l}$ as the one that we have considered. However, implementation of this refined method also requires more data as we now distinguish between more types of mutations. For this reason, we have chosen not to make the distinction.

7.7 Summary

In this Chapter there is a brief account on a part of my research activities while visiting the Netherlands Forensic Institute (NFI), in The Hague. There I focused on mutation rates and mutation models related topics. Aim on this part of the thesis is to handle some complicating features commonly affecting the mutation rates estimation, in particular to study a generalization of the hidden mutation phenomena, considering how largely the observed mutation rate per mutational distance differ from the actual rate.

To pursue this aim, in the first part of the Chapter we built appropriate auxiliary mutation models, differentiating among restricted and unrestricted ones. After that, considering both paternity trio's and duo's cases, we obtained the probabilities to underestimate the length of an occurred mutation by means of simulations.

Finally, relying on data issued by the America Association of Blood Banks, we proposed a way to correct the published data on mutation frequencies per distance, also giving some insights about the uncertainty derived from the adoption of simulation methods.

CHAPTER 8

Discussion

In this thesis we proposed a novel methodology to evaluate kinship identification systems. The goal of the analysis is to provide probabilistic information on the misleading results possibly deriving from the analysis before the traditional identification process is undertaken.

We described the theoretical and analytical aspects of our proposal in Chapter 4. The analysis considers the DNA evidence belonging to the individuals promoting the identification trial, but not that of the candidate to the identification, whose position in the familial pedigree is questioned. Thus the procedure precedes in time the usual kinship analysis and is performed without any additional costs and laboratory work since it uses only a subset of the data required for the post-experimental phase of the assessment of the hypotheses under debate, during which the LR is obtained in the way described in Section 4.2. In other words, this implies to perform the LR computation only after one has verified that the identification system has achieved the required expected performance for the case of interest. In our opinion, it should be compulsory to employ identification systems which had been proven to achieve a certain standard of quality, and this should represent a crucial aspect for any identification trial.

It is important to point out that, in our opinion, the final judgement on the goodness of an identification system in the pre-experimental phase is neither upon the shoulders of the statistician nor on those of the forensic scientist, but must be evaluated by the decision maker, i.e. the person or the group of people (a judge, an authority, a committee and so on) called to decide upon a case on the basis of some evidence. For this reason they should be trained in order to be able to take these kind of decisions based on probabilistic results given to them by the scientist.

This methodology shares one of the goals of the topic of the Design of Experiments (DOE): the purpose of improving the statistical inference by appropriately selecting the conditions under which a crucial unobserved random variable has certain desirable characteristics. DOE usually aims to evaluate and control the variability of the random variables the experiment is going to observe by means of treatments under which the experiment could be carried out, followed by an optimization step. In this paper, the focus is on the probabilities the candidate's genetic characteristics can assume, conditional on the alternative hypotheses. The different conditions under which the candidate's genetic traces could arise are related to some realistic scenarios arising by sequentially including, up to the point of having reached satisfactory characteristics, different loci and familial donors.

The adopted technique could be extended to other classification problems, notwithstanding that the formulation of the required models might be, in other circumstances, more demanding, questionable, and not so directly driven by the hypotheses. Here, the hypotheses completely define the variables to be included and the structure of dependence between them. If the stochastic system is represented by a Bayesian network, the only further ingredients are the population and the segregation models which specify the conditional probability of each node with respect to its parents, which are mainly determined and estimated outside the classification at hand. Moreover, in the application section, the realized sensitivity analysis suggests that there is only a slight dependence of the results on such choices.

We believe that the use of probabilities of faithful and misleading evidence for each of the hypotheses under debate is the most natural way for people in the forensic field to appreciate the goodness of a system: other approaches based on utility functions (Taroni et al. (2007)) or information theoretical measures (Lauritzen and Mazumder (2008)) are possible but, in our opinion, not easy to be perceived in a court. Other alternatives could also be taken into account: for example the probabilities of exclusion as detailed in Buckleton et al. (2005) are just (partial) subsets of the LR distribution, i.e. probabilities associated to the event the LR is equal to 0.

Another relevant result reached by this thesis is the development of an efficient computational strategy to obtain the LR distributions required for the analysis, for a specific, well defined kinship case, as shown in Chapter 5. Furthermore the adoption of Bayesian networks allows us to easily consider complicating issues involving more realistic biological models for DNA evidence.

In our opinion, the whole matter presented in this thesis is relevant since, up to now, the capabilities of a proposed system have not been revealed to the parties, including those called to make the final judgement on the identification trial. In order to advocate the effectiveness of this methodology we presented a number of applied cases in Chapter 6.

Finally, the proposal represents a way to take into account the principles expressed in the *Daubert v. Merree Dow Pharmaceutical Inc.* sentence we referred to at the beginning of the thesis, in Section 1.1.

APPENDIX A

LR equivalence classes

Here we consider a case in which an individual, the DNA donor, is trying to identify a candidate as the family member U posited to be a distance of n generations on the direct lineage. If $n = 1$, this is a motherless paternity case; if $n = 2$, it is the case of a grandparent trying to identify a candidate as the grandson; and so on. We illustrate how the number of different LRs arising in this circumstance is not equal to the number of possible genotypes the candidate can assume, $k(k+1)/2$, but is a number independent of k and n, where k is the number of alleles in the locus.

Let $X^0 = (r, s)$ be the genotype of the donor and assume the population alleles' probabilities $\boldsymbol{p}$ are known. For the sake of simplicity, we make use of the HW and the ML models.

On the donor lineage, consider the probability distribution of the transmitted allele. At the first generation, $n = 1$, it can assume only two values, r and s, with probability 0.5. For $n > 1$, the probability of observing r or s is 0.5^n plus the probability of coming from the non-donor lineage.

Let A^n be the distribution of the allele n generations after the donor had provided $X^0 = (r, s)$, then:

$$Pr\Big(A^n = i|X^0 = (r,s)\Big) = \begin{cases} (0.5)^n + (1-(0.5)^{n-1})p_i, & \text{if } i \in \{r,s\}, \\ (1-(0.5)^{n-1})p_i, & \text{if } i \notin \{r,s\}, \end{cases}$$

for $n > 1$.

Since the allele coming from the non-donor lineage still has a probability governed by the population parameters, the genotype probability along the generations, X^n, $Pr\Big(X^n = (i,j)|X^0 = (r,s)\Big)$ is

$$\begin{cases} (0.5)^n(p_r + p_s) + (1-(0.5)^{n-1})2p_rp_s, & \text{if } i = r, j = s, \\ (0.5)^n(p_j) + (1-(0.5)^{n-1})2p_rp_j, & \text{if } i = r, j \neq s, \\ (0.5)^n(p_i) + (1-(0.5)^{n-1})2p_sp_i, & \text{if } i \neq r, j = s, \\ (0.5)^n(p_r) + (1-(0.5)^{n-1})p_r^2, & \text{if } i = r, j = r, \\ (0.5)^n(p_s) + (1-(0.5)^{n-1})p_s^2, & \text{if } i = s, j = s, \\ (1-(0.5)^{n-1})2p_ip_j, & \text{if } i \neq r, j \neq s. \end{cases}$$

For this reason the $LR = \dfrac{Pr\big(X^n = (i,j)|X^0 = (r,s)\big)}{Pr\big(X^n = (i,j)|p\big)}$ is:

$$\begin{cases} (0.5)^{n+1}\dfrac{(p_r + p_s)}{p_r p_s} + (1 - (0.5)^{n-1}), & \text{if } i = r, j = s, \\ (0.5)^{n+1}p_r^{-1} + (1 - (0.5)^{n-1}), & \text{if } i = r, j \neq s, \\ (0.5)^{n+1}p_s^{-1} + (1 - (0.5)^{n-1}), & \text{if } i \neq r, j = s, \\ (0.5)^{n}p_r^{-1} + (1 - (0.5)^{n-1}), & \text{if } i = r, j = r, \\ (0.5)^{n}p_s^{-1} + (1 - (0.5)^{n-1}), & \text{if } i = s, j = s, \\ 1 - (0.5)^{n-1}, & \text{if } i \neq r, j \neq s. \end{cases} \tag{A.1}$$

The last line shows that for the descendant's genotypes with alleles different from r and s, the LR always assumes the value of $1 - (0.5)^{n-1}$. This fact reduces the LR sample space to six or three possible states, depending on whether the donor is heterozygous or homozygous, respectively.

APPENDIX B

Genetic data for the applied cases

In this Appendix we give details about the genetic data employed in the applied cases of Chapter 6. In the next few tables, also genetic evidence of the candidate to the identification will be given, even if it remains clear that we did not make use of such pieces of evidence.

A Motherless Paternity Case

Table B.1: Genetic evidence for the Motherless Paternity case.

Locus	B	AF
D8S1179	13-14	8-14
D21S11	30-32.2	28-30
D7S820	8-11	8-12
CSF1PO	10-10	9-10
D3S1358	14-17	16-17
TH01	7-9	8-9
D13S317	11-13	11-13
D16S539	9-9	9-12
vWa	16-17	16-17
TPOX	9-11	9-12
D18S51	13-15	15-17
D5S818	12-12	12-12
FGA	22-25	22-25

A Full Sibling II Case

Table B.2: Genetic evidence for the Full Siblings II case.

Locus	F_X	F_1	F_2	F_3
D8S1179	11-12	11-13	11-13	11-13
D21S11	28-29	28-28	28-29	28-29
D7S820	10-10	9-11	10-10	9-10
CSF1PO	11-12	11-11	11-12	11-11
D3S1358	17-18	17-18	17-18	17-18
TH01	9-9	8-10	8-9	8-9
D13S317	11-12	11-12	11-12	11-12
D16S539	11-12	11-11	11-12	11-11
vWa	15-17	17-18	16-18	15-17
TPOX	9-11	9-11	8-11	9-11
D18S51	12-19	19-19	16-19	16-19
D5S818	11-12	12-13	11-12	11-12
FGA	20-20	20-20	20-20	20-25
D2S1338	17-20	19-21	20-21	17-19
D19S433	12-15	13-15	12-15	13-15

A Two Cousins Case

Table B.3: Genetic evidence for the Two Cousins case for different sets of employed loci.

	13 loci		15 loci		17 loci	
Locus	*S*	*A*	*S*	*A*	*S*	*A*
D8S1179	13-15	11-13	13-15	11-13	13-15	11-13
D21S11	27-28	28-29	27-28	28-29	27-28	28-29
D7S820	11-12	8-10	11-12	8-10	11-12	8-10
CSF1PO	10-12	10-10	10-12	10-10	10-12	10-10
D3S1358	13-17	17-18	13-17	17-18	13-17	17-18
TH01	6-8	6-9	6-8	6-9	6-8	6-9
D13S317	11-11	8-11	11-11	8-11	11-11	8-11
D16S539	12-13	10-11	12-13	10-11	12-13	10-11
vWa	16-17	14-17	16-17	14-17	16-17	14-17
TPOX	8-11	8-8	8-11	8-8	8-11	8-8
D18S51	13-13	14-15	13-13	14-15	13-13	14-15
D5S818	11-12	11-12	11-12	11-12	11-12	11-12
FGA	21-23	21-25	21-23	21-25	21-23	21-25
D2S1338			17-20	23-26	17-20	23-26
D19S433			15-15.2	14-15	15-15.2	14-15
PENTAD					8-14	14-14
PENTAE					12-17	5-14

A Stepwise Case

Table B.4: Genetic evidence for the Stepwise case.

Loci	S	M	B
D8S1179	14-14	13-14	13-13
D21S11	31.2-32.2	30-32.2	28-28
D7S820	10-10	10-10	8-11
CSF1PO	12-12	12-12	11-12
D3S1358	18-18	15-18	15-18
TH01	6-9	9-9.3	6-7
D13S317	12-12	10-12	8-8
D16S539	10-10	9-10	11-11
VWA	17-17	17-19	16-16
TPOX	8-9	8-9	8-11
D18S51	15-17	16-17	15-15
D5S818	10-12	9-12	10-13
FGA	20-22	22-23	23-25
PENTAD	13-13	12-13	9-10
PENTAE	12-14	12-17	17-17

APPENDIX C

Hidden mutations, algebraically

In this Appendix we algebraically determine the probabilities $A_{k,0}$ with which a mutation goes unnoticed in a trio, i.e. no mutations are needed to explain the observed genotypes of father, mother and child.

We consider a general mutation model with mutation probabilities $\mu_{i,j}$. We consider the observed genotypes to be those of a father-mother-child trio, and suppose that a mutation from (say, paternally inherited) allele i to j has happened. We can now distinguish between three possibilities, assuming that no maternal mutation takes place:

- The father has genotype (i, i). In that case, the mutation goes unnoticed if and only if the mother has genotype (i, j) and has transmitted allele i to the child. This happens with probability

$$\sum_i \sum_{j \neq i} p_i^3 \mu_{i,j} p_j. \tag{C.1}$$

- The father has genotype (i, j). In that case, the mutation always goes unnoticed. This event has probability

$$\sum_i \sum_{j \neq i} p_i \mu_{i,j} p_j. \tag{C.2}$$

- The father has genotype (i, k) for $k \notin \{i, j\}$. In this case, the mutation goes unnoticed if and only if the mother has genotype (j, k) and has passed on allele k, or if she has genotype (i, j) and has passed on allele i. This event has probability

$$\sum_i \sum_{j \neq i} \sum_{k \notin \{i,j\}} p_i \mu_{i,j} p_j (p_k^2 + p_i p_k). \tag{C.3}$$

Clearly, these are the only three ways for a mutation not to lead to an inconsistency in the genotypes, and since they are mutually exclusive, the probability that in a full trio a mutation has happened which has gone unnoticed, is equal to

$$\sum_i \sum_{j \neq i} p_i \mu_{i,j} p_j \left(1 + p_i - p_i p_j + \sum_{k \notin \{i,j\}} p_k^2. \right) \tag{C.4}$$

To compute the $A_{k,0}$ it now suffices to use (C.4) with a k-step mutation model. This gives the same results as those obtained from our computer simulation.

Acknowledgements

I spent three amazing years doing research in an exciting and stimulating field and for this I want to thank my supervisor Professor Fabio Corradi, who introduced me to this topic, allowed me to travel a lot to attend to seminars, conferences and courses and, most of all, spent a great amount of time committed to this project, showing me, despite all the difficulties, the real beauty of this work.

I am particularly grateful to Klaas Slooten. He did not know me until we met in person, but still he trusted me and let me spend four fantastic months at the Netherlands Forensic Institute, in Den Haag, The Netherlands. He forced me to look at things from a different point of view in the field of my research, but also demonstrated me that it is not strange to drink milk at lunch.

I want also say thank you to all the people that gave me valuable advices during these years: professors and researchers from my Department (a great place where to study) who shared some little pieces of their knowledges with me; all the people I met around the world and also all the anonymous referees that rejected my papers, pushing me to study harder and harder to achieve an objective.

Francesco e Alberto have been my academic "family" for these years, and I shared with them moments of fun even before working experiences. This journey, this period of my life, would not be the same without them. They are great guys and I both wish all the best for their future.

My friends, those of a lifetime, encouraged and kept me down-to-earth at the same time, by constantly repeating that I was born to make pies, even if they refer to those one can eat.

I will never forget the continuous material and spiritual support I have received from family. In particular I would like to thank my mum who taught me how little details are important, my dad who showed me every day the importance to work hard to get what you want and my grandpa Gino who taught me how vital it is to always pursue what one loves to do. Hard work, attention to details and love for what I do is what I've learned from them, and what I wish to myself for my future.

And last but not least, I want to thank Stefania from the bottom of my heart. She, from the first moment we met, with her intense love, helped me to face all the difficulties I encountered, much more than she believes. She is my soul mate, and I know that, with her, the future does not scare me at all.

Bibliography

AABB. Annual report summary for testing in 2003. Technical report, American Association of Blood Banks, 2003.

AABB. Annual report summary for testing in 2008. Technical report, American Association of Blood Banks, 2008.

D.J. Balding and R.A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96:3 – 12, 1995.

J.O. Berger. Bayesian Analysis: A Look at Today and Thoughts of Tomorrow. *Journal of the American Statistical Association*, 95(452):1269–1276, December 2000.

T.H. Boveri. *Ergebnisse über die Konstitution der chromatischen Substanz des Zelkerns.* Fischer, Jena., 1904.

C. Brenner. Multiple mutations, covert mutations and false exclusions in paternity casework. In C. Doutremèpuich and D. Morling, editors, *Progress in Forensic Genetics*, volume 10, pages 112 – 114. Elsevier BV, 2004.

C. Brenner and R.W Staub. Can DNA solve this? In *Poster presented at the 14-th International Symposium on Human Identification. Phoenix, AZ, September 30 - October 2*, 2003.

B. Brinkmann, M. Klintschar, F. Neuhuber, J. Huhne, and B. Rolf. Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat. *Am. J. Hum. Genet.*, 62:1408 – 1415, 1998.

F. Brisighelli, C. Capelli, I. Boschi, P. Garagnani, M.V. Lareu, V.L. Pascali, and A. Carracedo. Allele frequencies of fifteen strs in a representative sample of the italian population. *Forensic Science International: Genetics*, 3(2):e29–e30., 2009.

J. Buckleton, C.M. Triggs, and S.J. Walsh. *Forensic DNA Evidence Interpretation.* CRC Press, 2005.

J.M. Butler. Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of Forensic Sciences*, 51(2):253–265, 2006.

R. Chakraborty, D.N. Stivers, and Y. Zhong. Estimation of mutation rates from parentage exclusion data: applications to STR and VTNR data. *Mutation Research*, 354:41 – 48, 1996.

R. Cook, I.W. Evett, G. Jackson, and P.J. Jones. A model for case assessment and interpretation. *Science & Justice*, 38(3):151–156, 1998.

T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Wiley, 2006.

R.G. Cowel. Finex: a probabilistic expert system for forensic identification. *Forensic Science International*, 134:196 – 206, 2003.

R.G. Cowel, A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. *Probabilistic Expert Systems and Bayesian Networks.* Springerr-Verlag, New York, 1999.

A. P. Dawid and I. W. Evett. Using a graphical method to assist the evaluation of complicated patterns of evidence. *Journal of Forensic Sciences*, 42(2):226 – 231, 1997.

A.P. Dawid, J. Mortera, and V.L. Pascali. Non-fatherhood or mutation? A probabilistic approach to parental exclusion in paternity testing. *Forensic Science International*, 124:55 – 61, 2001.

A.P. Dawid, J. Mortera, V.L. Pascali, and D.W. van Boxel. Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics*, 29:577–595, 2002.

A.P. Dawid, J. Mortera, and P. Vicard. Object-Oriented Bayesian Networks for Complex Forensic DNA Profiling Problems. *Forensic Science International*, 169: 195–205, 2007.

J. Drábek. Validation of software for calculating the likelihood ratio for parentage and kinship. *Forensic Science International: Genetics*, 3:112–118, 2009.

T. Egeland, B. Kulle, and R. Andreassen. Essen-Möller and Identification Based on DNA. *Chance*, 19:27 – 31, 2006.

T. Essen-Möller. Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis. Theoretische Grundlagen. *Mitteilungen d Anthrop Ges (Wien)*, 68:2 – 53, 1938.

I.W. Evett and J.S. Buckleton. Statistical Analysis of STR Data. In *Advances in Forensic Haemogenetics*. Springer-Verlag Heilderberg, 1996.

P. Gill, J. Curran, and C. Neumann. Interpretation of complex DNA profiles using Tippett plots. *Forensic Science International: Genetics Supplement Series*, 1: 646 – 648, 2008.

I.J. Good. Studies in the History of Probability and Statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika*, 66(2):393–396, 1979.

I.J. Good. Weight of Evidence: A Brief Survey. In D.V. Lindley J.M. Bernardo, M.H. De Groot and A.F.M. Smith, editors, *Bayesian Statistics 2*, pages 249–270. Elsevier Science Publishers, 1985.

I.P. Good. *Applying Statistics in the Courtroom.* Chapman&Hall CRC - London, 2001.

W. Goodwin, A. Linacre, and S. Hadi. *An introduction to forensic genetics.* John Wiley and Sons, 2007.

P.J. Green and J. Mortera. Sensitivity of inferences in forensic genetics to assumptions about founding genes. *The Annals of Applied Statistics*, 3:731–763, 2009.

M.B. Hamilton. *Population Genetics.* Wiley-Blackwell, 2009.

G.H. Hardy. Mendelian proportions in a mixed population. *Science*, 28:49–50, 1908.

A.J. Jeffreys, V. Wilson, and S.W. Thein. Hypervariable 'minisatellite' regions in human dna. *Nature*, 314:67–73, 1984.

A.J. Jeffreys, J.F.Y. Brookfield, and R. Semenoff. Positive identification of an immigration test-case using human DNA fingerprints. *Nature*, 317:818 – 819, 1985.

D. Koller and A. Pfeffer. Object-oriented bayesian networks. In P. Shenoy (Eds.) D. Geiger, editor, *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence*, pages 302 – 313. Morgan Kaufmann Publishers, San Francisco, 1997.

S. Lauritzen and A. Mazumder. Informativeness of genetic markers for forensic inference-an information theoretic approach. *Forensic Science International: Genetics supplement Series*, 1:652–653, 2008.

S. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50:157 – 224, 1988.

S.L. Lauritzen and N.A. Sheehan. Graphical Models for Genetic Analyses. *Statistical Science*, 18:489–514, 2003.

D. Lindley. On the Measure of Information provided by an Experiment. *Annals of Mathematical Statistics*, 27:986–1005, 1956.

L.R. Mayor and D.J. Balding. Discrimination of half-siblings when maternal genotypes are known. *Forensic Science International*, 159:141 – 147, 2006.

A. Mazumder. *Planning in Forensic DNA Identification Using Probabilistic Expert Systems.* PhD thesis, University of Oxford, 2010.

J.G. Mendel. Versuche über pflanzenhybriden (experiments on plant hybridization). In *Verhandlungen des naturforschenden Vereins Brünn*, pages 3 – 47, 1866.

K. Murphy. The bayes net toolbox for matlab. *Computing Science and Statistics*, 33:1024–1034, 2001.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann Publishers, New York., 1988.

R. Royal. *Statistical Evidence.* Chapman&Hall London - New York, 2000.

C.E. Shannon. A Mathemathical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 1948.

K. Slooten and F. Ricciardi. Estimation of mutation probabilities for autosomal STR markers. *Forensic Science International: Genetics*, to appear:accepted for publication, 2012.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* Springer-Verlag, New York/Berlin, 1st ed. edition, 1993.

W.S. Sutton. On the morphology of the chromosome group in brachystola magna. *Biological Bulletin*, 4:24–39, 1902.

W.S. Sutton. The chromosomes in heredity. *Biological Bulletin*, 4:231–251, 1903.

F. Taroni, C. Aitken, P. Garbolino, and A. Biedermann. *Bayesian Networks and Probabilistic Inference in Forensic Science.* John Wiley and Sons, Chichester, UK, 2006.

F. Taroni, S. Bozza, M. Bernard, and C. Champod. Value of DNA Tests: A Decision Perpective. *Journal of Forensic Sciences*, 52,1:31–39, 2007.

P. Vicard and A.P. Dawid. A statistical treatment of biases affecting the estimation of mutation rates. *Mutation Research*, 547:19–33, 2004.

P. Vicard, A.P. Dawid, J. Mortera, and S.L. Lauritzen. Estimating mutation rates from pa- ternity casework,. *Forensic Science International: Genetics*, 2:9 – 18, 2008.

J.D. Watson and F.H.C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171:737–738, 1953.

W. Weinberg. Über den nachweis der vererbung beim menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg.*, 64:368–382, 1908.