Current challenges in statistical DNA evidence evaluation

Cereda, G.

Part I

Background and summary

Chapter 1

Preliminary concepts

1.1 Forensic DNA analysis

The unicity of each person's entire DNA sequence makes of DNA traces one of the most useful type of scientific findings for forensic identification. This is why, from its first use in the UK in 1986 (*R vs Colin Pitchfork*, Wambaugh (1989)), its use in forensic applications has become widespread (Walsh et al., 2004).

The contents of the forthcoming subsections are mainly based on Buckleton et al. (2005), Butler (2005, chap. 5) and Coquoz and Taroni (2006). They describe, from a biological and technical point of view, what DNA is and how it is used in the forensic context.

1.1.1 DNA as identification tool

DNA is the molecule that encodes the genetic instructions for the development and functioning of all known living organisms. It has a double-strand structure, where each strand is made up of a sequence of four nucleobases: Adenine, Cytosine, Guanine and Thymine (usually denoted by four letters, A, C, G and T). Each base is attached to a sugar molecule and a phosphate molecule to form *nucleotides*, arranged in the two long strands to form a spiral, with the shape of a double helix. Bases of one strand pair up with bases of the other strand – A with T, and C with G – to form units called *base pairs*. The instructions encoded in DNA are stored as a code made up of a double sequence of about 3 billions base pairs where the order of the bases in the sequence determines the information available for building and maintaining an organism. Since the bases are always paired A-T and C-G, to refer to a particular portion of the DNA sequence it is sufficient to consider one strand, such as AATTGCCTTTTAAAAA.

A distinct portion of DNA which codes instructions for a particular body's need (mostly the creation of proteins) is called *gene*. The 32,000 genes present in the human DNA form the so-called *genome*. All nucleotides are not aligned on a single chain: they are organised in thread-like structures called *chromosomes*. Humans have 46 chromosomes which form 23 couples of *homologous chromosomes*, identical to one another in shape and size, one inherited

from the mother and one inherited from the father. One of these pairs is composed by the *sex chromosomes* which, among other functions, determine the sex of the individual. The other 22 pairs of chromosomes are called *autosomal chromosomes* and determine the rest of the body makeup and functions. The DNA code, or *genetic code*, is passed to the offspring through the paternal sperm and the maternal egg: the mother passes 23 chromosomes through her egg, the father passes 23 chromosomes through his sperm.

The entire sequence of the DNA is unique to each individual. The reason for this variability is due to *recombination*, the process by the two chromosomes of each parent exchange some portions of the DNA sequence each other, shown in Figure 1.1. The resulting chromosomes, built of parts from the two chromosomes of one parent, are different from each of the two original chromosomes of that parent. *Meiosis* (or gamete cell production) is the process during

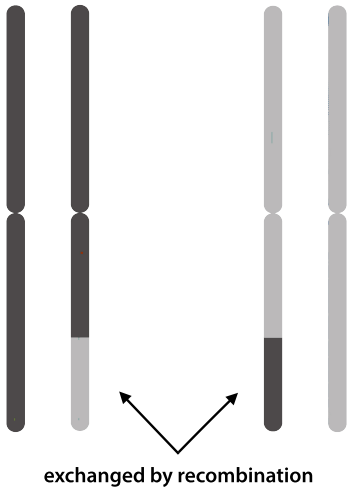

Figure 1.1: Genetic recombination between the chromosomes of each parent.

which each reproductive cell receives randomly one recombined chromosome. Other sources of variability among individuals are *mutations*, which are changes in the nucleotide sequence of the genome, due to deletions, insertions or metamorphosis of some nucleotides.

Although the entire *genome* is unique to each person, the greatest part of it is similar in all humans and only 0.1% characterizes the different individuals: the variations present in this minute portion of DNA are called *polymorphisms*. For forensic purposes, only those portions of the DNA sequence which are known to display a polymorphism are analyzed: these zones are called *genetic markers* (or *loci*) and the alternative possible variants which the DNA sequence displays in these zones are called *alleles*. An individual can have at most two alleles for the same polymorphism: the one carried by the paternal chromosome and the one carried by the maternal chromosome. If these two alleles are equal, the individual is said to be *homozygous* at the specific marker, otherwise he is said to be *heterozygous*. The couple of alleles present at an individual's genetic marker is called *genotype* and a combination of alleles at adjacent locations on a chromosome, inherited together, is called *haplotype*. A *DNA profile* is the combination of genotypes of multiple markers. While the entire DNA sequence is unique to each individual, there is the possibility that a DNA profile can be shared by two unrelated persons, with a (usually tiny) probability that decreases when the number

of loci are analysed. Moreover, a father and a son have the same DNA sequence in their Y-chromosome. Hence, DNA profiles obtained from this portion of DNA are shared by many people in the same population.

There are two different kinds of polymorphism, which are commonly analyzed:

- The *length polymorphisms* are located in particular portions of the DNA molecule where a particular sequence of nucleotides repeats itself many times. Each allele is represented by the number of such repetitions.
- The *sequences polymorphisms* are located in regions of the DNA strands where the type of one or more nucleotides varies among individuals. The different alleles are distinguished by a difference in one ore more nucleotides.

In this project *STR polymorphisms* (belonging to the first group), and *Deletion Insertion polymorphism* (belonging to the second group) will be used.

STR, Short Tandem Repeat Polymorphisms

A short tandem repeat (STR) polymorphism is a length polymorphism made of a pattern of two or more bases, which are repeated directly adjacent to each other. These patterns, called *words*, are typically repeated between 3 and 51 times: the polymorphism is represented by the difference in the number of repetitions of the same word, between the different individuals. *STR markers* are specific regions of the DNA in which an STR polymorphism is known to exist. Each STR allele is a number corresponding to the number of repetitions of the same sequence.

Below, the readers can see what the DNA sequence of an individual looks like, at the same locus of two homologous chromosomes, when an STR polymorphism is present.

Chromosome A_1
..CGGGT$\underbrace{\text{ATTGATTGATTGATTGATTGATTGATTGATTG}}_{8\times\text{ATTG}}$GAAAGGT..
Chromosome A_2
..CGGGT$\underbrace{\text{ATTGATTGATTGATTGATTGATTG}}_{6\times\text{ATTG}}$GAAAGGT..

The repeating pattern is the word ATTG. The genotype of this person at this locus is (6,8). Repeat numbers could be integers or decimals: for instance, if the repeat number is 9.3, this means that there are 9 repetitions and an incomplete repeat consisting of 3 more letters.

Insertion-deletion polymorphisms

Insertion-deletion (INDELs or DIPs) are length polymorphisms created by the presence or the absence of short (typically 1 to 50 base pairs) sequences of nucleotides in the human genome (Pereira et al., 2009a). DIPs are diallelic, the two possible alleles being L (for *long*) in case the specific combination of nucleotides is present, or S (for *short*) in case it is absent.

Allele L
..ATGCGT AATT TAGGGCTGGATC...

Allele S
..ATGCGTTAGGGCTGGATC.....

In the example, the sequence of nucleotides AATT is present in the first sequence, while it is absent in the second one.

In the last years, with the identification of 2000 human diallelic INDELs (Weber et al., 2002), this kind of polymorphisms has received major attention. Since then, a large literature has been developed, about genetic structure of human population (Yang et al., 2005; Rosenberg et al., 2005) and their use as genetic markers in natural population (Vali et al., 2008). A huge map of insertion-deletion variation in the human genome, which contains more than 415,000 distinct polymorphisms is published by Mills et al. (2006). DIPs are currently used for forensic purpose (da Costa Francez et al., 2012), especially for complex pedigree kinships (Pereira et al., 2009b) and for identification studies involving highly degraded DNA (Weber et al., 2002), a field in which the use of DIPs is very promising and more reliable than the use of STRs.

1.1.2 Technical steps of DNA genotyping

DNA can be found on different biological materials, such as blood, sperm, saliva, and hairs, but also on objects which have been in contact with human cells. After a trace is collected, its cells are broken down with chemical reagents, in order to reveal the DNA inside them. After this, the DNA is purified, with the aim of separating it from other molecules that can potentially interfere or inhibit the process of analysis.

In order to obtain a genetic profile from a DNA trace, it is necessary to amplify it across several orders of magnitude. This is done through a biochemical technology, called *Polymerase Chain Reaction* or *PCR* (Reynolds et al., 1991), which generates thousands of millions of copies of a selected portion of the DNA sequence. The method relies on about 30 cycles of repeated heating and cooling. The target DNA sequence to be amplified is pinpointed through the use of *primers*, which are short DNA fragments containing sequences of nucleotides complementary (according to the rule A-T, C-G) to the target region, together with an enzyme, the DNA polymerase, that adds the building blocks in the proper order based on the template DNA sequence.

The genetic markers to be amplified can be of different type. For this project STR markers, Y-STR markers and DIP-STR markers will be considered, as described in the forthcoming Sections 1.1.3 and 1.1.5.

A separation step is then required to pull the different targeted fragments apart, and to allow to distinguish the different alleles. This separation process is performed through *electrophoresis*, a technique which is used to separate molecules on the basis of their weight. The process is based on the migration undertaken by charged molecules, when immersed in a liquid and exposed to the electrical field generated by a couple of electrodes of opposite charge: negatively charged molecules move to the positive electrode, and vice versa. This movement has

a different speed, on the basis of the size of the molecules: light molecules will move faster than heavier ones. This causes the alleles to sort themselves according to weight. If applied to the amplified DNA fragments, the process consists of injecting the amplified sample into a gel, then to pass an electrical current through the gel, causing the alleles, which are all negatively charged, to move towards the positive pole. Since DNA is not visible in natural light, coloured dyes are used to monitor the progress of the electrophoresis. The output from the process is a graph, called *peak profile* or *electropherogram*, in which the horizontal axis gives the base pair measurement, and the vertical axis the light intensity. Each peak indicates the presence of an allele, where the height is a measure of the amount of the allele in the amplified sample. In Figure 1.2, an example of electropherogram obtained with STR markers is shown.

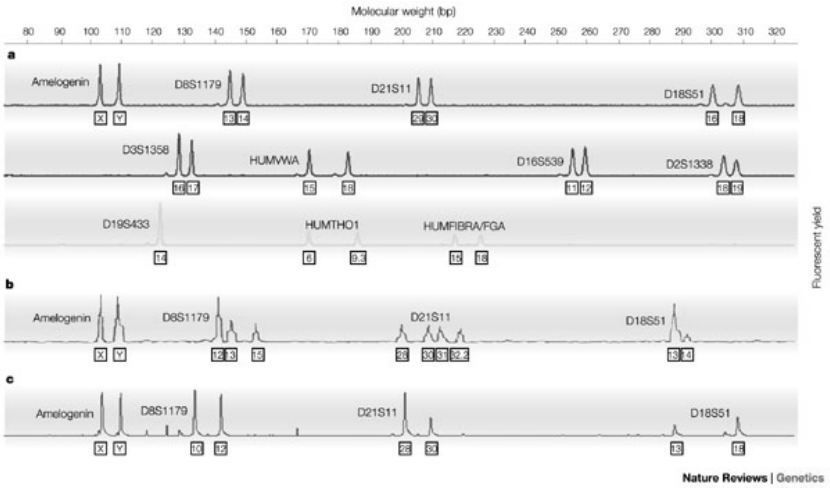

Figure 1.2: Electropherogram (Jobling and Gill, 2004) showing peaks at different *loci*. The numbers below each peak denote the corresponding allelic repeat numbers, while the name of the analysed *locus* is presented on the top of the peak(s) graph (i.e., D3S1358, vWA, etc.).

1.1.3 Two classical techniques of DNA genotyping

In forensic laboratories, two widely used marker systems are STR and Y-STR. This section provides an overview, highlighting advantages and drawbacks of the two methods.

STR marker system

An STR marker system is composed of a kit of analysis, designed to target some STR polymorphisms in a DNA trace of interest. These kits typically allow experts to target between 9 and 15 STR markers (plus Amelogenin, which is also used to assign the sex of the donor).

STR polymorphisms have been the predominant type of polymorphism used in human genetic studies since about 1990, and can be considered a standard approach to help addressing most

problems about forensic inference regarding identity of individuals (Chakraborty et al., 1999; Butler, 2011). There are several benefits in using STRs: they are multi-allelic and highly discriminating between unrelated and even closely related individuals, and they are relatively easy and not expensive to analyse (Vali et al., 2008).

Y-STR marker system

Y-STR markers are STR markers situated on the Y-chromosome (Butler, 2005). They are used in forensic casework (e.g. Roewer et al., 1992; Roewer, 2009), especially for their capacity to reveal male-specific Y-STR alleles in female/male DNA mixtures, even if extremely unbalanced (when classical STR markers are not performing adequately). These markers are thus very useful, in particular for cases of extremely unbalanced mixtures in which the major contributor is female and the minor one is male. There are however some limitations in the use of Y-STR markers. One the one hand, they are usable only for mixtures with a specific gender combination (Y-STRs only detect male component's DNA in a female background), reducing dramatically the number of suitable cases. On the other hand, a Y-STR profile can be quite common in a population (Vermeulen et al., 2009) and patrilineal relatives of a suspect cannot be excluded as being the contributors to the stain, if no mutations occur (Roewer, 2009). Recently, a panel of 13 rapidly mutating (RM) Y-STR markers has been identified (Ballantyne et al., 2012), which successfully differentiate between closely and distantly related males. However, they have the same gender restrictions of classical Y-STR markers.

It is also important to mention that, due to the lack of recombination, the different Y-STR markers form a single haplotype, formed by alleles that are not independent one another.

1.1.4 DNA mixtures

DNA mixtures are stains that contain genetic material from more than one person. This can be due to a contact between the different individuals' DNA material, anytime before the trace is collected.[1] Mixtures are commonly found, after sex rapes, in vaginal swabs obtained from the victim, as the traces usually contains material coming from the victim and the perpetrator (but also from other consensual partner(s)). A single contributor having at most two distinct alleles per locus, the main factor identifying a mixture is the presence of three or more alleles at at least one locus.

The criticality of mixtures is represented by the difficulty of the so-called "deconvolution", that is discerning the particular genotype of each contributor: this happens every time the different contributors share some alleles at some locus, as shown in the example of Table 1.1.

A way to separate the different contributors' allele, is to use information about the height (or the area) of the peaks. Some laboratories do so in order to distinguish the different contribu-

[1] Note also that a mixture can be generated after the collection of the trace, if contamination occurs.

Alleles detected in locus 1	**Alleles of the possible donors**		
11,12,13	(11,11) (12,13)	(11,12) (12,13)	(11,11) (11,12) (11,13)

Table 1.1: Three among the different combinations of the possible contributors' alleles, when (11, 12, 13) is observed at a specific locus. For the same detected alleles there may be a combination of homozygous-heterozygous contributors, two heterozygous contributors, or three contributors.

tors, based on their percentage of share in the whole mixture. This is called the *quantitative aspect* of the data, opposed to the *qualitative* one, which is retained for this research, and uses only information about the detection or not of the alleles at each locus.

1.1.5 Extremely unbalanced DNA mixture

One of the limitations of using STR markers is that this method does not work successfully if the proportion between the DNA quantities of the two contributors is more extreme than 1:10 (Clayton and Buckleton, 2005; Sutherland et al., 2009). Mixtures with these characteristics are referred to in this thesis as 'extremely unbalanced mixtures'. In the process of amplifying the DNA, the primers fail to pinpoint the alleles of the minor contributor, which are masked by those of the major one. Here, the threshold of 10% is retained as limit of detection of the minor DNA for blood:blood mixtures. This value varies depending on the type of biological fluids which constitute the mixture and on the specific combination of genotypes present in the mixture (Applied Biosystems, 2012), which should be assessed in the validation procedure (Butler, 2011).

Extremely unbalanced mixtures are quite common in forensic contexts, such as in cases of sexual assaults when the victim's DNA is largely predominant. Moreover, several fields of medical genetics are concerned, for example with the phenomenon of microchimerism during pregnancy, which is caused by the circulation of minute quantities (from 3% to 6%) of foetal DNA in the maternal blood (Lo et al., 1998; Tjoa et al., 2006), but also after organ transplant, when traces of the donor's DNA are present in the blood fluid of the transplanted patient (Gadi et al., 2006; Pujal and Gallardo, 2008). In forensic contexts, to address the constraint of these kind of mixtures, Y-STR markers are usually adopted, with the limitations described in Section 1.1.3.

As pointed out in Oldoni et al. (2015), it is difficult to estimate the precise incidence of unbalanced mixtures. Thus, it is undoubtely that most of the extremely unbalanced mixtures recovered so far have been evaluated under the assumption that they were single stains, losing interesting information about further contributors which they may have potentially provided. The solution to this problem is the use of the DIP-STR marker system, which allows the experts to obtain (at least part of) the minor contributor's genotype, in many cases, as explained in the next section.

DIP-STR marker system

As explained in Section 1.1.3, STR and Y-STR marker systems have some limitations: the first one is generally not working properly for extremely unbalanced mixtures, the second one requires a good gender combination between the two contributors to the stain and does not allow to discriminate between patrilinear relatives. Both the constraints of these methods can be overcome by the use of DIP-STR markers, which have recently been proposed as a novel type of genetic markers (Castella et al., 2013; Oldoni et al., 2015).

In fact, the problem of the masking of the minor genotype can be addressed with the use of primers that are allele-specific to assure that, each time the two contributors have different genotypes at some marker, the primers will anneal to different alleles. DIP markers have this characteristic: the primer specific for the allele S is different from the primer specific for the allele L.

However, the use of DIP markers alone, has the limitation of a low discriminating power. The novelty, here, consists in pairing a DIP polymorphism with a standard STR polymorphism, to increase the discriminating power and form a superlocus where the two component loci are not independent (less than 500bp apart).[2]

DIP-STR genotyping allows the selected amplification of the minor contributor's genotype as long as it has a DIP allele which is not in the major contributor's DIP alleles, in at least one marker. At each marker, the best scenario is when the DNA of the major and of the minor contributors are homozygous for different DIP alleles (i.e., one S-S and the other L-L). In this case, the possible results can show either two DIP-STR alleles of the minor contributor or one, depending on the STR-homozygosis or heterozygosis of the minor contributor. On the other hand, when the major contributor is DIP-homozygous and the minor contributor is DIP-heterozygous, only one haplotype of the minor DNA can be recovered (i.e., the one with the DIP allele different to the DIP allele of the major contributor's DNA). Table 1.2 summarizes the possible outcomes.

A limitation of this method is that, when the predominant DNA is DIP-heterozygous or both contributors are DIP-homozygous of the same type, it is not possible to obtain any result from the mixture: this happens because both the DIP primers (S and L), if used, anneal to the major contributor's DNA.

However, it is important to notice that, if the major contributor is DIP homozygous, some information about the minor contributor can be inferred by the absence of results (first and fifth row of Table 1.2): in fact, this absence indicates that the minor has the same DIP-homozygosity of the major (they are both S-S or L-L).

A first panel of 9 DIP-STR markers was first presented in Castella et al. (2013), while more recently 9 additional DIP-STR alleles have recently been made available (Oldoni et al., 2015). While within the same DIP-STR marker the DIP locus and the STR locus are chosen close enough so as not to recombine, independence can be assumed between the different allelic configurations of DIP-STR markers in the proposed panel.

[2]The two component loci are not independent because they are so close on the chromosomes that they cannot recombine.

DIP genotype of major contributor	DIP genotype of minor contributor	DIP-STR results using the DIP primer opposite to that of the major contributor
S-S	S-S	No results
	L-L	Complete genotype of the minor contributor
	S-L	Only the L DIP-STR allele
L-L	S-S	Complete genotype of the minor contributor
	L-L	No results
	S-L	Only the S DIP-STR allele
S-L	S-S L-L S-L	No results in each situation

Table 1.2: Informativeness of the different genotypic DIP-STR configurations.

Since DIP-STR alleles of the minor were successfully detected at ratios up to 1:1000 (Castella et al., 2013), the method reveals advantages for extremely unbalanced mixtures, for example in cases of sexual assaults when the victim's DNA is largely predominant, or cases of microchimerism during pregnancy. In Castella et al. (2013), the authors propose to test the use of DIP-STR markers also in early stages of pregnancy to perform kinship analyses, largely demanded for cases of pregnancy after rape (Guo et al., 2012).

1.2 Evaluation of DNA evidence

One of the main aims of forensic statistics is to evaluate to what degree some piece of evidence supports one or the other of exclusive hypotheses of interest. When the piece of evidence is a DNA trace which is found at the crime scene and whose profile corresponds to a known suspect's DNA profile, the hypotheses of interest are (unfortunately (Taroni et al., 2013)) often of the source level kind: 'the crime stain came from the suspect' (h_p) and 'the crime stain came from an unknown donor, unrelated to the suspect' (h_d).[3] When dealing with results from DNA mixtures of two contributors, the situation becomes more complicated, depending on the number of alleles observed at each marker. If the genotype of one of the two contributors (for instance, the victim of a sexual aggression) is known, and the suspect is compatible as contributor to the stain, the two hypotheses of interest become: 'the DNA in the mixed stain belongs to the victim and the suspect' (h_p), versus 'the DNA in the mixed stain belongs to the victim and to an unknown person, unrelated to the suspect' (h_d). The (unknown) true hypothesis h is the parameter of interest of our model. Usually the model involves also nuisance parameters, denoted as θ, unknown quantities necessary to perform the evaluation. To deal with the uncertainty over the nuisance parameters, additional data,

[3]This type of hypotheses are said 'source-level hypotheses' For a discussion on the use (misuse) of source level hypotheses, please refer to Champod et al. (2016). Note that this aspect is not considered in the current research.

which we will refer to as 'background', is usually given to the forensic statistician. This is partially different from the 'background information' I as defined in Aitken and Taroni (2004) and Taroni et al. (2014), but in many cases background data can be thought of as part of the background information. For instance, when dealing with DNA evidence, the nuisance parameter is often the list of the allelic frequencies in the population of interest, and the background data used to deal with it, is a sample of DNA profiles from the population in the form of a database. The reader is invited to notice the difference between θ and h: one is the parameter which we 'test' through the likelihood ratio (h), the other (θ) is a nuisance parameter involved in its calculation. Data to evaluate is made of evidence and background. The extent to which data is helpful to discriminate between the competing hypotheses of interest is called *probative value*.

1.2.1 Bayesian inference and likelihood ratio

Bayesian inference allows one to update subjective beliefs on propositions, when new information is gathered through Bayes' theorem. This is important for forensic statisticians that want to quantify how the available data modifies prior beliefs on hypotheses of interest (Lindley, 1977b; Taroni et al., 2010). Bayesian methods are divided into parametric (when parameters of the model are finite dimensional) or nonparametric (in presence of infinite dimensional parameters). The common ground is that a prior distribution is given to all the unknown quantities of the model. This prior should be based on the subjective belief of the Bayesian statistician performing the inference. This does not mean that these priors are arbitrary, rather they are based on the experience and set of knowledge of the individual, at a given time (Lindley, 1978).

The largely accepted method to quantify the probative value of given forensic findings (data, observation, measurement, ...) is the calculation of the *likelihood ratio*, a statistic that expresses the relative plausibility of the observations under the (generally) two hypotheses of interest (Robertson and Vignaux, 1995; Evett and Weir, 1998; Aitken and Taroni, 2004; Balding, 2005; Steele and Balding, 2014).

After a couple of hypotheses is given, the Bayesian likelihood ratio is defined as

$$\mathrm{LR} = \frac{\Pr(D = d \mid H = h_p)}{\Pr(D = d \mid H = h_d)}, \tag{1.1}$$

where Pr is the Bayesian probability, reflecting the expert's belief on the joint distribution of the random variables of the model, namely D (representing the data), H (representing the hypotheses), and Θ (the nuisance parameter(s)).

The likelihood ratio is used to quantify the way in which new information can change the belief, or 'odds', that a particular hypothesis is true. *Prior odds* and *posterior odds* are the odds before and after introducing information, such as a new piece of evidence. As part of Bayes' theorem, the likelihood ratio connects prior odds to posterior odds in the following way:

$$\underbrace{\frac{\Pr(H = h_p \mid D = d)}{\Pr(H = h_d \mid D = d)}}_{\textit{Posterior odds}} = \underbrace{\frac{\Pr(D = d \mid H = h_p)}{\Pr(D = d \mid H = h_d)}}_{\textit{Likelihood ratio}} \times \underbrace{\frac{\Pr(H = h_p)}{\Pr(H = h_d)}}_{\textit{Prior odds}}. \tag{1.2}$$

This formula clarifies that the likelihood ratio, which is a measure of the probative value of the findings D with respect to two alternative hypotheses, is to be distinguished from the conditional degree of belief on the same hypotheses (represented by posterior odds). The likelihood ratio is the ratio between posterior odds and prior odds. In a full Bayesian approach to the interpretation of evidence, the prior beliefs should be assigned by the commissioner (court, police) and should then be updated using the likelihood ratio, which is domain of the forensic laboratory, with the ultimate aim of obtaining the posterior beliefs. For a review on the importance of the likelihood ratio in the legal context, see Lindley (1991) and Aitken and Taroni (2004).

1.2.2 Frequentist likelihood ratio

On the other hand, in a frequentist context, the nuisance parameter θ and the hypotheses h are taken as fixed (unknown) quantities. The frequentist probability (here denoted as $\mathcal{Pr}$) can be expressed in terms of the Bayesian Pr, in the following way: $\mathcal{Pr}_\theta(\cdot \mid h) := \Pr(\cdot \mid \Theta = \theta, H = h)$, $\forall h$. The name is due to the fact that the probability of an event can be interpreted as the limit of its relative frequency in a large number of experiments. The frequentist likelihood ratio can be thus expressed as

$$\mathcal{LR}_\theta = \frac{\mathcal{Pr}_\theta(D = d \mid h_p)}{\mathcal{Pr}_\theta(D = d \mid h_d)}. \tag{1.3}$$

The difference between Bayesian and frequentist methods consists in how they treat the parameters θ and h. A Bayesian models the uncertainty about their value by random variables Θ and H, which are given prior distributions $p(\theta)$ and $p(h)$. Frequentists consider them as fixed (i.e., without distribution) unknown quantities.

One of the aims of this thesis is to carefully differentiate between the frequentist and the Bayesian approach, in order to provide guidelines for consistent solutions on both sides. A precise and concise account on problems and pitfalls in the LR definition and assessment can be found in Dawid (2016).

1.2.3 Rare type match problem

The evaluation of a match between the profile of a particular piece of evidence and a suspect's profile depends on the proportion of individuals with that profile in the population of potential perpetrators. Indeed, it is intuitive that the rarer the matching profile, the more the evidence is pointing against the suspect. Problems arise when the observed frequency of the profile in a sample from the population of interest (i.e., in a reference database) is 0. Such characteristic is likely to be rare, but it is challenging to quantify how rare it is. This problem is so

substantial that it has been defined "the fundamental problem of forensic mathematics" (Brenner, 2010).

The rare type match problem is particularly important in case a new kind of forensic evidence, such as results from DIP-STR markers is involved, and for which the available database size is still limited. The same happens when Y-chromosome (or mitochondrial) DNA profiles are used: because of the lack of recombination involved when offspring DNA is generated from the DNA of the parents, the haplotype must be treated as a unit and the set of possible haplotypes is extremely large. As a consequence, most of the Y-STR haplotypes are not represented in the database. This research will start by focussing on solving the rare type match problem for Y-STR markers. The final aim is to apply the studied solutions to DIP-STR markers.

1.2.4 Available solutions for the rare type match problem

The *empirical frequency estimator*, also called *naive estimator* or *maximum likelihood estimator* (MLE), that uses the frequency of the characteristic in the database, puts unit probability mass on the set of already observed characteristics, and it is thus unprepared for the observation of a new type. A solution could be the *add-constant* estimators (in particular the well-known *add-one* estimator, due to Laplace (1814), and the *add-half* estimator of Krichevsky and Trofimov (1981)), which add a constant to the count of each type, included the unseen ones. However, this method requires to know the number of possible unseen types, and it is also not very well-performing when this number is large compared to the sample size (see Gale and Church (1994) for an additional discussion).

Alternatively, Louis (1981) proposes the so-called 'rule of three', that states that if n is the size of the database, $3/n$ is a good approximation of the 95% upper bound for the frequency. This is also proposed in a Bayesian framework, by Jovanovic and Levy (1997); Winkler et al. (2002); Chen and McGee (2008).

Of interest for this research is the nonparametric *Good-Turing estimator* of Good (1953), based on an intuition on A. M. Turing. It is an estimator for the total unobserved probability mass, based on the proportion of singletons in the sample. If compared to the maximum likelihood estimator, or to the add one estimator, it has the advantage of being usable for the unobserved species, without additional constraints (Orlitsky et al., 2003). However, it does not allow to separate the frequencies of the unseen species, nor to estimate their number. For a comparison between *add one* and *Good-Turing* estimator, see Orlitsky et al. (2003).

As stated in Anevski et al. (2013), the *naive estimator* and the *Good Turing estimator* are in some sense complementary: the first gives a good estimate for the observed types, and the second for the probability mass of the unobserved ones. Lastly, the *high profile estimator*, introduced by Orlitsky et al. (2004), extends the tail of the *naive estimator* to the region of unobserved types. This estimator has been improved by Anevski et al. (2013) that also give the consistency proof.

Literature provides some examples of approaches to evaluate the likelihood ratio for the rare type match problem for Y-STR haplotypes: Egeland and Salas (2008), the κ method

Brenner (2010, 2014), the coalescent theory method (Andersen et al., 2013a), the haplotype surveying method (Roewer et al., 2000; Krawczak, 2001; Willuweit et al., 2011), and the discrete Laplace method (Andersen et al., 2013b). The latter was not proposed directly for the rare haplotype match case but is usable for that purpose. For more details about the discrete Laplace method, see Section 1.2.5.

Bayesian nonparametric estimators for the probability of observing a new type have been proposed by Tiwari and Tripathi (1989) using Dirichlet process, by Lijoi et al. (2007) using general Gibbs prior, and by Favaro et al. (2009) with specific interest to the two-parameter Poisson Dirichlet prior. However, for the LR assessment one has to obtain not only the probability of observing a new species but also the probability of observing this same species twice (according to the defence, the profile of the crime stain and the profile of the suspect are two independent observations).

1.2.5 The discrete Laplace method

The diversity of STR alleles in the population is the result of mutations. The *stepwise mutation model* (Kimura and Ohta, 1978) assumes that each mutation can, with equal probability, increase or decrease an STR repeat number by at most one, in one generation. In Caliebe et al. (2010), a Markov chain description of the stepwise mutation model is proposed, which shows nice convergence properties. The proposed process is called *normalized allele process* and models the difference between the STR alleles of the each individual and a fixed individual, in each generation. This process is proved to be a positive recurrent irreducible Markov chain on $\mathbb{Z}^{N-1}$. As such, it converges exponentially fast to the unique invariant distribution, which is unimodal if the mutation rate $\mu \leq 0.8$.

In Andersen et al. (2013b), the discrete Laplace distribution (Inusah and Kozubowski, 2006) is suggested (and empirically validated) as an approximation to the invariant distribution of the normalized allele process. A discrete random variable D is said to follow the discrete Laplace distribution $\mathrm{DL}(p)$, with $0 < p < 1$ if

$$\Pr(D = d) = f(d) = \Big(\frac{1-p}{1+p}\Big)p^{|d|}, \qquad \forall d \in \mathbb{Z}.$$

This result is used to model the distribution of single locus alleles, where the allele of reference which normalizes the process is estimated through the median of all the alleles at that locus.

According to this choice, the distribution of each single locus alleles X_i has density function

$$f(|d-m|) = \big(\frac{1-p}{1+p}\big)p^{|d-m|},$$

where m and p can be estimated using MLE from a sample $\{d_i\}_{i=1}^N$, as

$$\hat{m} = \mathrm{median}\{d_i\}_{i=1}^n,$$
$$\hat{p} = \hat{\mu}^{-1}\Big(\sqrt{\hat{\mu}^2+1} - 1\Big),$$

where

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} |d_i - \hat{m}|.$$

Let $\mathbf{X} = (X_1, X_2, ..., X_r)$ denote the random variable which describes an r-loci haplotype configuration. Moreover, there may be c different subpopulations to take into consideration. By making the strong assumption of independence between loci, within the same subpopulation, the following density is used to describe the probability that $\mathbf{X} = \mathbf{x}$:

$$f(\mathbf{x} \mid \{\mathbf{y_j}\}_j, \{\mathbf{p_j}\}_j) = \sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f(x_k \mid y_{jk}, p_{jk}).$$

For each j, τ_j is the probability a priori of generating from the jth subpopulation, while $\mathbf{p_j} = (p_{j1}, p_{j2}, ..., p_{jr})$ and $\mathbf{y_j} = (y_{j1}, y_{j2}, ..., y_{jr})$ represent the dispersion and location parameters, respectively, of the jth subpopulation.

The R package `disclapmix` (Andersen, 2013) allows one to estimate all the parameters of the model using the EM algorithm (Dempster et al., 1977), with initial subpopulation centres chosen via PAM algorithm (Kaufman and Rousseeuw, 2009), while their number is chosen by the BIC criteria (Schwarz, 1978).

1.3 Graphical models

The well-known *probabilistic graphical models* are graph-based representations that consist of a qualitative part, where features from graph theory are used, and a quantitative part that specifies the probability distributions over the nodes of the graph. They are defined as a "marriage between probability theory and graph theory" by Jordan (1998, 2004). Such network structures are formulated in a graphical communication language and can be seen as a compact representation of (conditional) dependencies and independencies between variables.

The most famous type of graphical models are the Bayesian networks, described in details in Section 1.3.1. The main goal of such models is to use conditional independences to find factorizations of the distribution over the graph structure, with the ultimate aim of computing efficiently the probabilities of interest.

Bayesian networks were first introduced by Pearl (1982), but other detailed accounts can be found in specialized literature (Pearl, 1988; Neapolitan, 1990; Jordan, 1998; Jensen and Nielsen, 2007; Cowell et al., 2007a). They are used in many fields where reasoning under uncertainty plays a central role (Pearl, 1988; Neapolitan, 1990; Gómez, 2004; Pourret et al., 2008) and they are thus becoming a more and more regularly used approach for analyzing problems in forensic science, where they are now part of well established literature (Aitken and Gammerman, 1989; Dawid and Evett, 1997; Dawid et al., 2002; Garbolino and Taroni, 2002; Taroni et al., 2004, 2014; Cowell et al., 2006b; Biedermann, 2007; Fenton and Neil, 2012). For all these reasons, Bayesian networks (and their object-oriented extension described in Section 1.3.2) are retained as the general modeling framework in this research.

The evaluation of DNA results via the likelihood ratio (described in Section 1.2), requires the calculation of the likelihood for data given the alternative hypothesis, which may be challenging. When adopted for kinship analyses, for example, likelihood ratio formulae become considerably complex, depending on parameters such as the supposed degree of relatedness and the number of individuals one needs to account for. Moreover, formulae may vary according to the genotypic configurations of the target individuals and the chosen genotyping technique. This computational burden can − as shown by foundational works of Dawid et al. (2002) − be approached and safely handled through Bayesian networks to obtain the same results as those obtained by Essen-Möller's formulaic approach (Essen-Möller, 1938). In fact, Bayesian networks allow one to obtain numerator and denominator of the likelihood ratio in few simple steps. Moreover, they prove to be a highly versatile framework that can accommodate analysts and reasoners with differing inferential interests (Taroni et al., 2014; Kjærulff and Madsen, 2008).

1.3.1 Bayesian networks: formal definition

A directed acyclic graph (DAG) is a graph formed by a collection of vertices, and by directed edges connecting one vertex to another, in a way that makes it impossible to start at some vertex v and follow a sequence of edges that eventually loops back to v again. Under its formal definition (e.g. Jensen and Nielsen, 2007), a Bayesian network is a DAG, whose vertices (also called nodes) form a finite set V and correspond to random variables $\mathcal{X} = \{X_v,\ v \in V\}$, while directed edges represent dependencies between variables. A node v is called a *parent* of a node w if there is a directed edge from v to w in the graph. Each node has a finite set of states and is equipped with a conditional distribution (given the parent nodes). For any ordering $X_1, X_2, ..., X_n$ of the random variables in $\mathcal{X}$, the edge set specifies the following factorization of the joint density p:

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} \Pr(x_i | \mathrm{Pa}(x_i)), \tag{1.4}$$

where $\mathrm{Pa}(x_i)$ represents the set of the parent nodes of the node x_i. In case a node has no parent nodes, by $p(x_i|\emptyset)$ we mean $p(x_i)$. Equation (1.4) is called a *recursive factorization of p according to the DAG*, and formally defines Bayesian networks.

Bayesian networks are typically used for probabilistic inference. Each node represents a proposition or an assertion, such as those that an individual forms during the reasoning task. The probabilistic inference amounts to update the degree of belief on the truth of a particular proposition in the light of new information. Practically, the network is used to update knowledge about the state of a subset of variables when other variables are observed. The use of a factorization as that defined by Equation (1.4) reduces the computational effort of the procedure. In forensic applications, Bayesian networks are used to update the probabilities of the hypotheses of the prosecution (h_p) and of the defence (h_d) when relevant data for the case is obtained, but also the other way around: to update the probabilities of observing the data, under the two hypotheses h_p and h_d, (i.e., to calculate numerator and denominator of the

likelihood ratio). For this research we used the software Hugin,[4] specifically designed to work with Bayesian networks, along with the R package `RHugin`, which allows one to integrate the two platforms. A simple example of Bayesian network is the one for reasoning about diseases and symptoms, represented in Figure 1.3. Also known in specialized literature as the 'Classical diagnostic problem', it refers to a situation in which particular symptoms can arise as a consequence of different causes (or, diseases). This is a hypothetical, but widely applicable, scheme of representation, which is retained here for the sole purpose of remaining on a general level of discussion.

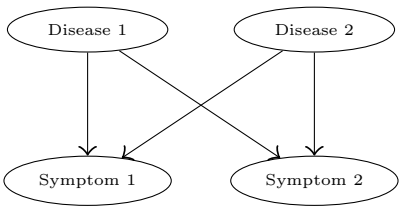

Figure 1.3: Bayesian network for generic reasoning about diseases and symptoms. Each node is Boolean with the state *True* representing the presence of the described disease or symptom, and the state *False* representing its absence.

The network models a situation in which there are two possible diseases and two symptoms. Both the diseases and the symptoms are not mutually exclusive. Each node is Boolean and the associated CPTs are shown in Tables 1.3 and 1.4.

Disease 1 = *False*	0.9
Disease 1 = *True*	0.1
Disease 2 = *False*	0.3
Disease 2 = *True*	0.7

Table 1.3: CPTs of the disease nodes. These two tables express the view that, initially, there is a probability of 0.1 for an individual to have one disease, and a probability of 0.7 to have the other.

Disease 1 =	*False*		*True*	
Disease 2 =	*False*	*True*	*False*	*True*
Symptom 1 = *False*	1	0.7	0.4	0.1
Symptom 1 = *True*	0	0.3	0.6	0.9
Disease 1 =	*False*		*True*	
Disease 2 =	*False*	*True*	*False*	*True*
Symptom 2 = *False*	1	0.9	0.4	0.05
Symptom 2 = *True*	0	0.1	0.6	0.95

Table 1.4: CPTs of the nodes representing the two symptoms. These tables contain the probability that the analyst assigns to the presence of the symptoms, given each possible configuration of the two disease nodes.

This network allows an analyst to revise his beliefs about the different diseases given the presence (or, absence) of one or both symptoms. For example, the analyst may ask questions

[4] http://www.hugin.com.

of the following kind: "If symptom 2 (but not symptom 1) is observed, what should be the probability that disease 1 affects the patient X?", "Does symptom 2 allow me to discriminate between the two 'causes', disease 1 and disease 2?", or "What if symptom 1 is absent?".

1.3.2 Object-orientation

The Bayesian network formalism offers interesting modeling capacity, but it is not always efficient or straightforward in its process of manual construction. For example, in case the model is composed of the repetitive use of some submodels, the 'copy and paste' system may be demanding, because all the submodels have to be updated to integrate new information (or evidence). In order to deal with this problem, object-oriented Bayesian networks (OOBN) can be employed (Koller and Pfeffer, 1997; Laskey and Mahoney, 1997; Bangsø and Wuillemin, 2000; Kjærulff and Madsen, 2008; Korb and Nicholson, 2011).

An object-oriented Bayesian network (also called a *class*) is a network that, in addition to regular nodes, contains *objects*. These are subnets that encapsulate themselves multiple objects (or subnetworks), giving rise to a composite hierarchical structure. Objects become part of a class under the form of so-called *instance nodes*.

One of the main advantages of using object-oriented Bayesian networks is that they are well suited for problem domains containing repetitive patterns. This happens typically when dealing with DNA evidence, in particular with DNA mixtures: in this case several contributors are represented in the network, each with a similar network substructure: with classical Bayesian networks, one is forced to use repetitive network fragments. With OOBNs, the use of instance nodes simplifies the problem, since the subnet is built only once, and then integrated in the class of interest through instances.

Each instance node is connected to other nodes of the 'external' class through the so-called *interface nodes*, divided into *input* and *output nodes*. As the name suggests, input and output nodes are the only nodes which directly connect any instance of the class to the external network: to the connected node(s), they hold the role of children or parent, respectively. Input nodes are actually placeholders for their parent nodes in the external network: they have the same states and the arrow only serves the technical purpose of establishing a logical equivalence. Throughout this text, instances are drawn as rounded rectangles, while interface nodes are grey: input nodes have a dashed outer border whereas output nodes have a solid line.

To show the utility of OOBNs one can consider again the generic problem of reasoning about diseases and symptoms, introduced earlier in Section 1.3.1. Imagine that the analyst wants to describe the progression in time of this setting. To do this with a Bayesian network, one possibility is to use a network structure as shown in Figure 1.4, where the same substructure is repeated twice. If several such periods need to be represented, this way of building the network can be time-demanding, especially if one needs to modify the CPT in the nodes of each repeated structure. With object-oriented Bayesian networks, one can create a class to model a single period, and use it in the form of instance nodes, where input nodes **Previous Disease 1** and **Previous Disease 2** are designed to be bound to nodes **Disease 1** and

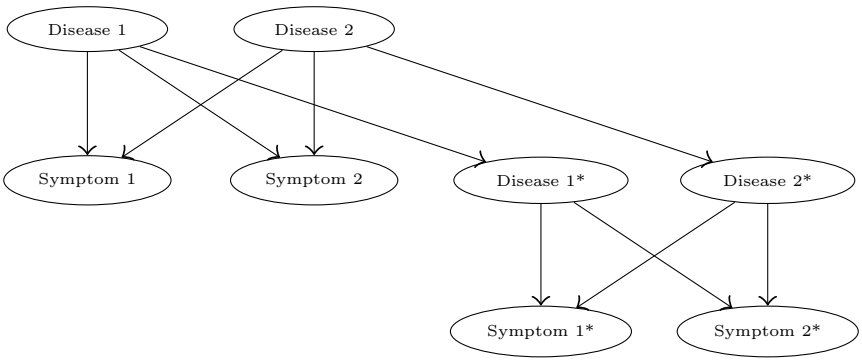

Figure 1.4: Bayesian network for reasoning about diseases and symptoms over time. Nodes **Disease 1** and **Disease 2** refer to the first period, while nodes **Disease 1*** and **Disease 2*** refer to the second period.

Disease 2 in the instance representing the previous period of time. This class is named `Period`, and is shown in Figure 1.5.

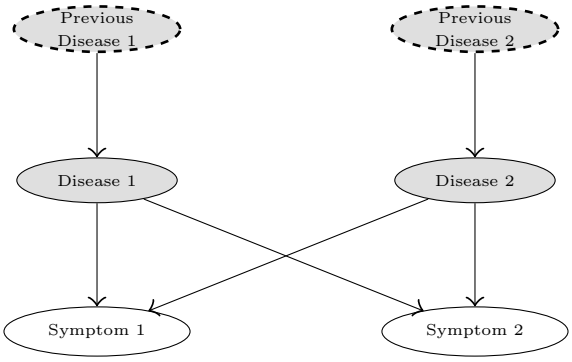

Figure 1.5: Representation of the class `Period`.

A model covering three (or more) periods can now be built simply by creating three instances of the class `Period`, as shown in Figure 1.6. As can be seen, nodes **Symptom 1** and **Symptom 2** do not appear in the external network, since they are not interface nodes.

Generally, an object-oriented Bayesian network can be used in the same way as a Bayesian network, instantiating some nodes to obtain the conditional probabilities of other nodes of interest.

Existing forensic literature considers the object-oriented Bayesian networks a particularly useful approach for addressing many evaluative aspects that are associated with evaluation of complex patterns of evidence (Dawid et al., 2007; Hepler et al., 2007; Taroni et al., 2014), or of results of DNA profiling analyses including mixtures, mutations, inference of source or kinship analyses. This is the reason why we chose OOBNs as the relevant methodological choice for the problem of interpreting DIP-STR results, which will be studied in this research.

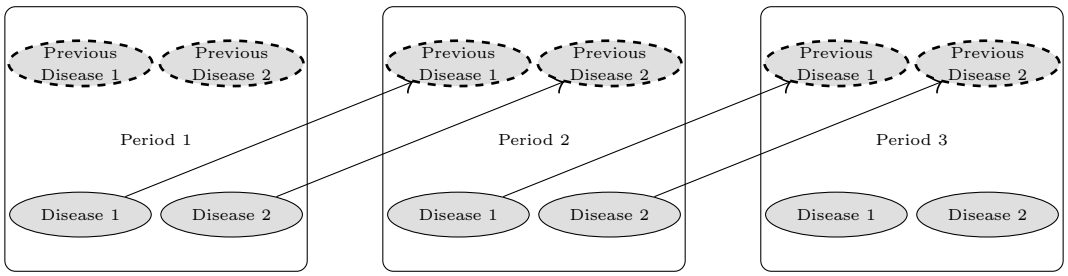

Figure 1.6: Expanded representation of an OOBN with three periods for reasoning about diseases and symptoms.

1.3.3 Bayesian networks in forensic DNA literature

The paper of Dawid and Evett (1997) is the first that introduces the use of Bayesian networks for DNA evidence evaluation, through an example involving blood stains.

Basic structures to deal with some aspects of DNA analyses have been proposed in the literature of the last decade.

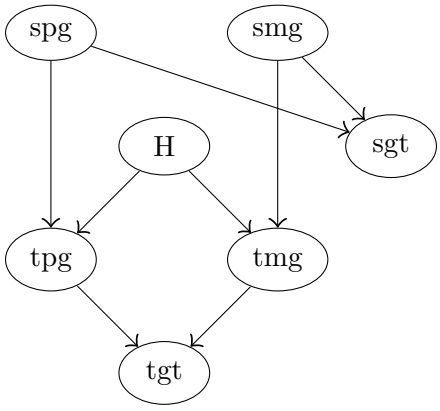

Figure 1.7: Bayesian network used for the evaluation of DNA results obtained from a crime stain (not mixed) and a suspect's stain.

Figure 1.7 shows a simple example of a Bayesian network to evaluate the correspondence between the DNA profile of a suspect and of a crime stain, focusing on individual genes and genotypes, presented in Dawid et al. (2002), and further discussed in Taroni et al. (2006). **H** is the hypothesis node, regarding whether or not the suspect is the source of the stain; **sgt** and **tgt** represent the genotype of the suspect and of the crime stain, respectively. Similarly, **spg** and **smg** represent paternal and maternal gene of the suspect, while **tpg** and **tmg** represent the paternal and maternal gene of the donor of the trace.

Another structure, which deals with the possibility of finding small quantities of DNA is proposed in Evett et al. (2002). A further important contribution in this context has been provided by Dawid et al. (2002), where the authors show the passage from initial pedigree representation of forensic identification problems to appropriate Bayesian networks representation.

A classical structure for the evaluation of DNA mixtures is shown in Figure 1.8 (Mortera

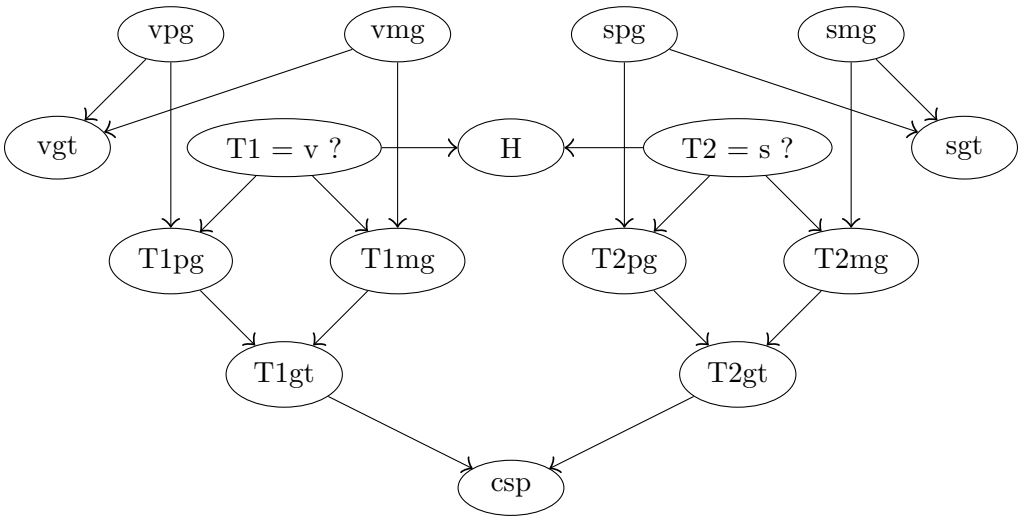

Figure 1.8: Bayesian network used for the evaluation of DNA results from a mixed DNA stain, recovered on a crime scene (Mortera et al., 2003).

et al., 2003). The network can be used to evaluate results from a crime stain (represented by **csp**), which contain two or more alleles per marker. **H** represents the hypotheses regarding the contributor of the stain, which may be (i) the suspect and the victim, (ii) the victim and an unknown individual, (iii) the suspect and an unknown individual, and (iv) two unknown individuals. The labels 'T1' and 'T2' denote, respectively, the first and the second contributors to the mixture. Nodes **vpg**, **vmg**, **spg** and **smg** represent the paternal and maternal genes of the victim and of the suspect, respectively, while **vgt** and **sgt** represent the victim's and the suspect's genotypes. Boolean nodes **T1=v?** and **T2=s?** model whether the victim has contributed to the crime stain or not, and whether the suspect has contributed to the crime stain or not, respectively. Nodes **T1pg**, **T2pg**, **T1mg** and **T2mg** represent their paternal and maternal genes, while **T1gt** and **T2gt** their genotypes.

Other more detailed structures are proposed in Mortera et al. (2003) and Biedermann et al. (2011b), the latter handling situations for which the number of contributors is not available. In Cowell et al. (2011), a network which models results from a quantitative approach and takes into account allelic drop-outs, stutters and silent alleles is presented. A specific overview of scientific literature about the use of Bayesian networks in forensic DNA applications can be found in Biedermann and Taroni (2012).

Bayesian networks have also been used for kinship analyses (Dawid et al., 1999, 2002), but more has been developed using object-oriented Bayesian networks (e.g. Hepler and Weir, 2008).

1.3.4 Object-oriented Bayesian networks for DNA evidence

It is in recent years that considerable research has been devoted to the application of object-oriented Bayesian networks to the evaluation of DNA evidence. As already mentioned, they are a very valuable tool to be used in this field, with different main perspectives, due to their versatility and modularity. Consider a model that describes the genotype, at a certain

marker, of two different persons: a suspect and an unknown person. The mechanism with which each of their alleles is inherited from their parents is the same for both individuals. Hence, in a Bayesian network it would give rise to four similar substructures (one for each allele). Through the use of an OOBN, it is possible to achieve this by using only one class, invoked in the network via four instances.

Another useful feature of OOBNs is their flexibility. In particular, a given network structure can readily be adapted to model other markers, or other individuals (such as the sibling of a missing suspect). As examples, an OOBN proposed in Green and Mortera (2009) is explained in details here (see Figure 1.9), to be used to evaluate DNA mixture evidence.

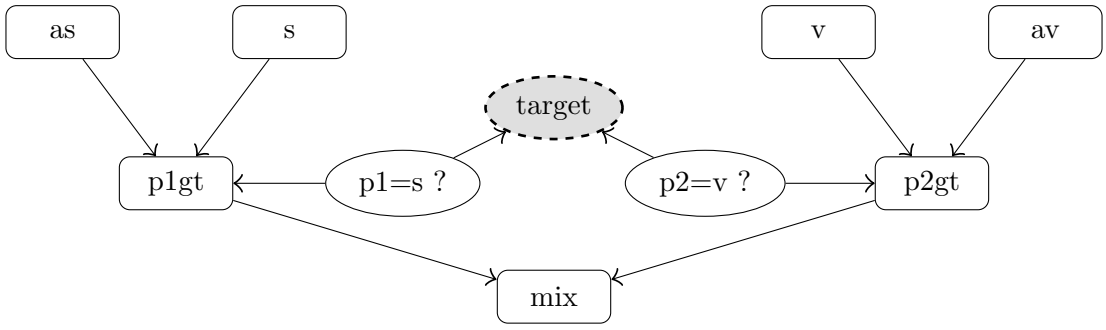

Figure 1.9: The network `Mixture` to be used for the evaluation of results obtained from a DNA mixture of two contributors.

The network `Mixture` of Figure 1.9 is used to calculate the strength of the evidence regarding a mixture of two contributors, the first of which can be either the suspect or an unknown person, while the second can be either the victim or an unknown person.

This network is made of instances of the classes `Genotype` and `Trace`, shown in Figure 1.10 and 1.11. The class `Genotype`, represented in the class `Mixture` by the instances **s**, **as**, **v**, and **av** is designed to represent the genotype, at a specific marker, of each individual.

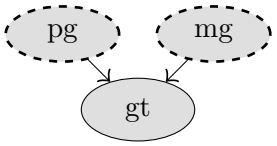

Figure 1.10: The class `Genotype`. The nodes **pg**, **mg** and **gt**, represent the paternal allele, the maternal allele, and the genotype itself, which is a logical combination of the parents' alleles.

The class `Trace` is represented in the class `Mixture` by the instances **p1gt** and **p2gt**. Ordinary nodes **p1=s?** and **p2=v?** of the class `Mixture` are boolean: they are *True* if, respectively, the suspect and the victim are contributors to the mixture. In the interaction with instance nodes **p1gt** and **p2gt**, respectively, they have the same function that node **p in mixture?** has, in the interaction with node **trace** of the class `Trace`.

The node **target** is the logical combination of the nodes **p1=s?** and **p2=v?** into four hypotheses:
h_1: Victim and suspect contributed to the mixture,

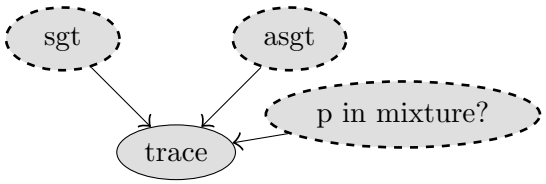

Figure 1.11: The class `Trace`. The output node **trace** represents the crime scene trace, and is identical to the nodes **sgt** (or **asgt**), depending on the state of the input node **p in mixture?**. For example, $\Pr(\textbf{trace} = \{a,b\} |\ \textbf{sgt} = \{a,b\}, \textbf{asgt} = \{c,d\}, \textbf{p in mixture?} = \textit{True}) = 1$, for any a, b, c, d. Nodes **sgt** and **asgt** are input nodes, bounded to nodes **gt** of the instances **s** and **as** of the class `Genotype`. Information about the genotype of the suspect is entered by fixing values of the node **gt** of the instance **s**.

h_2: Victim and an unknown person contributed to the mixture,
h_3: An unknown person and the suspect contributed to the mixture,
h_4: Two unknown persons contributed to the mixture.

However, by fixing the node **p2=v?** to its state *True*, the network can be used to evaluate the strength of the evidence with respect to the hypotheses h_1 and h_2, corresponding to h_p and h_d, respectively.

Node **mix** represents the alleles present at the specific marker of the mixed trace. It is the logical combination of the alleles from the two contributors. Information about the genotype of the suspect and the victim is entered by fixing the values for node **gt** in the instances **s** and **v** of the class `Genotype`. The likelihood ratio is then obtained by consecutively selecting the two hypotheses of interest (h_1 and h_2) in node **target** and reading the posterior probability corresponding to the observed state of node **mix**.

Again, this network models a single marker, but an OOBN to take into account results from all markers can be constructed, as shown in Figure 1.12. All markers (here only six for the sake of example) are modelled through instances **Marker i** ($i = 1, ..., 6$) of the network `Single Trace`. The input node **target** of each instances is bounded to the node **target** of the external class, as shown in Figure 1.12.

In order to support scientists in the evaluation of DNA mixture evidence from a quantitative point of view, object-oriented Bayesian network have been proposed in Cowell et al. (2007b, 2008), further developed in Cowell (2009), to take into account additional aspects, such as allelic drop-outs, stutter bands and silent alleles. Further uses of object-oriented Bayesian networks can be found in Cavallini and Corradi (2006), in the context of database searching problems (Gittelson et al., 2012) and in Green and Mortera (2009), where they are used to explore what happens when standard assumptions about the founder genes − such as the knowledge of the allele population proportion or the Hardy-Weinberg equilibrium − are violated.

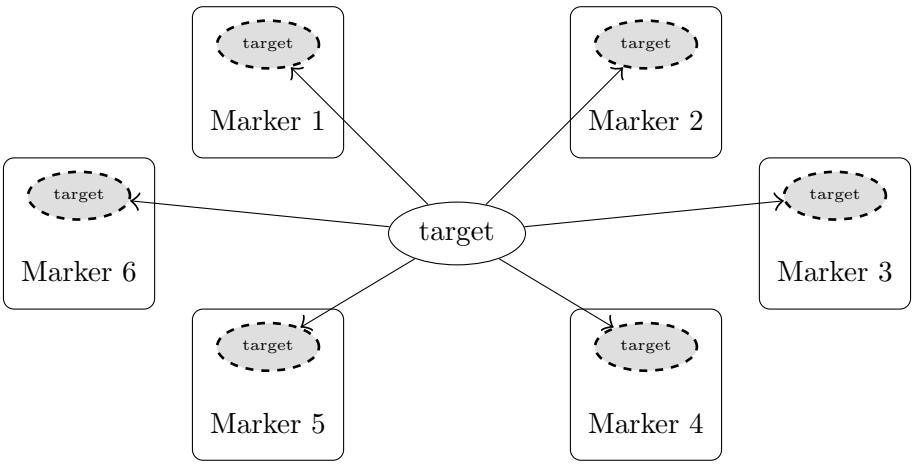

Figure 1.12: Expanded representation of an OOBN combining several markers.

1.4 Nonparametric Bayesian priors

A Bayesian nonparametric model is a Bayesian model on an infinite-dimensional parameter space. For our applications, the unknown parameter is typically represented by the vector containing the frequencies of some genetic characteristics (single STR alleles at one locus, or entire Y-STR haplotypes) in the population of possible perpetrators. The parameter space can be thought of as infinite dimensional, assuming that any STR number can potentially be observed. Even though this is not realistic (there may be biological reasons that bound the STR range) an infinite dimension can be a good approximation of a large dimension. This allows us to use nonparametric priors, in particular the two-parameter Poisson Dirichlet prior, which shows many interesting properties.

1.4.1 The two-parameter Poisson Dirichlet distribution

The two-parameter Poisson Dirichlet distribution, first introduced in Pitman (1992), is a distribution over ∇_∞, the infinite simplex of the form $\nabla_\infty = \{(p_1, p_2, ...) \mid p_1 \geq p_2 \geq ... > 0, \sum p_i = 1\}$. It is, in practice, a distribution on distributions. It can be constructed in two steps: first by simulating another distribution and then by sorting the results.

1. Given $0 \leq \alpha < 1$, and $\theta > -\alpha$, the vector $\mathbf{W} = (W_1, W_2, ...)$ is said to be distributed according to the GEM(α, θ), if $\forall i, W_i = V_i \prod_{j=1}^{i-1}(1-V_j)$, where the V_i are independently distributed as Beta$(1-\alpha, \theta + i\alpha)$. The GEM distribution (short for 'Griffin-Engen-McCloskey distribution') is well-known in literature as the "stick breaking prior", since it measures the random sizes in which a stick is broken iteratively.

2. The random vector $\mathbf{P} = (P_1, P_2, ...)$ obtained sorting the elements in $\mathbf{W}$ in decreasing order, has the two-parameter Poisson Dirichlet distribution PD(α, θ). Parameter α is called *discount parameter*, while θ is the *concentration parameter.*

For $\alpha = 0$, we obtain the well-known Poisson Dirichlet distribution, first introduced in Kingman (1975) as the infinite dimensional generalization of the classical Dirichlet distribution.

Notice that also $\alpha < 0$ and $\theta = -m\alpha$ for some $m \in \mathbb{N}$ is allowed: this is used for a model with only finitely (m) many DNA types, where a Dirichlet prior is put over the probability vector $\mathbf{P}$, with all hyperparameters equal to $-\alpha$. This case is not of interest for our research.

1.4.2 Pitman sampling formula

Partitions of the set $[n] = \{1, ..., n\}$ will be denoted with $\pi_{[n]}$, random partitions of $[n]$ will be denoted as $\Pi_{[n]}$. The different subsets forming a partitions are called 'blocks' of the partition.

Given a sequence of integer-valued random variables $X_1, ..., X_n$, consider the equivalence relation $i \sim j$ if and only if $X_i = X_j$. The equivalence classes of this relation form a random partition of $[n]$, which will be denoted as $\Pi_{[n]}(X_1, X_2, ..., X_n)$.

It holds that, if

$$\mathbf{P} \mid \alpha, \theta \sim PD(\alpha, \theta), \quad \text{and} \quad X_1, X_2, ... \mid \mathbf{P} = \mathbf{p} \sim_{\text{i.i.d}} \mathbf{p}, \tag{1.5}$$

then, for all $n \in \mathbb{N}$, the random partition $\Pi_{[n]} = \Pi_{[n]}(X_1, ..., X_n)$ has the following distribution:

$$\Pr(\Pi_{[n]} = \pi_{[n]} | \alpha, \theta) = \frac{[\theta + \alpha]_{k-1;\alpha}}{[\theta + 1]_{n-1;1}} \prod_{i=1}^{k} [1 - \alpha]_{n_i - 1;1}, \tag{1.6}$$

where n_i is the size of the ith block of $\pi_{[n]}$ (the blocks are here ordered according to the least element), and $\forall x, b \in \mathbb{R}, a \in \mathbb{N}$,

$$[x]_{a,b} := \begin{cases} \prod_{i=1}^{a-1} (x + ib) & \text{if } a \in \mathbb{N} \backslash \{0\} \\ 0 & \text{if } a = 0 \end{cases}.$$

This model can be used for partitions formed by a sample of individuals from a population with an infinite number of species, where the distribution of the vector containing the ordered population frequencies is assumed as $\text{PD}(\alpha, \theta)$ distributed. This is also known as the *Pitman sampling formula*, further studied in Pitman (1995). Notice that for $\alpha = 0$ we obtain the famous Ewens's sampling formula (Ewens, 1972).

1.4.3 The two-parameter Chinese restaurant process

Consider a restaurant with infinitely many tables, each one infinitely large. Let $Y_1, Y_2, ...$ be integer valued random variables that represent the seating plan: tables are ranked in order of occupancy, and $Y_i = j$ means that the ith customer seats at the jth table to be created. The two-parameter Chinese restaurant process is described by the following transition matrix:

$$Y_1 = 1,$$

$$\Pr(Y_{n+1} = i | Y_1, ..., Y_n) = \begin{cases} \dfrac{\theta + k\alpha}{n + \theta} & \text{if } i = k + 1 \\ \\ \dfrac{n_i - \alpha}{n + \theta} & \text{if } 1 \leq i \leq k \end{cases} \tag{1.7}$$

where k is the number of tables occupied by the first n customers, and n_i is the number of customers that occupy table i. The process depends on two parameters α and θ with the same conditions $0 \leq \alpha < 1$, $\theta > 0$ valid for the two-parameter Poisson Dirichlet distribution of Section 1.4.1. First described in case $\alpha = 0$ by Aldous (1985), this process is studied in details in Pitman and Picard (2006).

This process is deeply related to the two-parameter Poisson Dirichlet distribution thanks to the following results:

- if (1.5) and (1.7) hold, then for all $n \in \mathbb{N}$ the random partition $\Pi_{[n]} = \Pi_{[n]}(X_1, ..., X_n)$ has the same distribution as $\Pi_{[n]}(Y_1, ..., Y_n)$. They are both distributed according to the Pitman sampling formula (1.6). Stated otherwise, we can use the seating plan of n customers to obtain the same partition $\pi_{[n]}$ obtained by a sample $X_1, ..., X_n$ from a population whose probabilities are distributed according to a two-parameter Poisson Dirichlet distribution.
- the distribution of the infinite vector containing the relative sizes of the tables when infinite many customers have taken seats is $\text{PD}(\alpha, \theta)$.

This result is the key to use the two-parameter Poisson Dirichlet distribution as a Bayesian nonparametric prior for the rare type match case. Indeed, despite the rather complex definition presented in Section 1.4.1, it allows simplifying the definition of the only two probabilities of interest: given a database of size n, the probability (of interest for the prosecution) of observing a not yet observed Y-STR haplotype, and the probability (of interest for the defence) of observing the same not yet observed Y-STR haplotype twice. This will be explained in Section 2.7, and in details in Chapter 7, but the reader may already foresee that, by discarding enough information, the whole story can be described in terms of a customer entering into a restaurant with n customers already seated, and choosing an unoccupied table, or in terms of two customers entering the restaurant and choosing the same unoccupied table. These two probabilities can be easily obtained with formula (1.7).

1.4.4 The hyperparameters

To use the two-parameter Poisson Dirichlet prior, one has to deal with the presence of the two hyperparameters α and θ. This can be done either by choosing an hyperprior for them, or doing an assignment based on data (hence choosing an empirical Bayesian approach).

Regarding the possibility to infer their real values by analzying enough data, it is important to know that for a fixed $\alpha \in (0, 1)$, the $\text{PD}(\alpha, \theta)$ (for different θ) are all mutually absolutely continuous. Thus, θ cannot be consistently estimated. This is not much of a problem, since when n increases, the parameter θ becomes less and less important. It describes how much "social" are the customers: the smaller θ the more the customers tend to seat to already

occupied tables. It defines the size of the tables created first. On the other hand, α can be consistently estimated, as shown by the power law behavior, and there exists at least one consistent estimator for α (Carlton, 1999), namely:

$$\hat{\alpha} = \frac{\log K_n}{\log n}.$$

1.4.5 Power law behavior

Sampling from a two-parameter Poisson Dirichlet prior, using the marginalization described in (1.7), shows the rich-get richer clustering property (Teh et al., 2006): the more customers have been assigned to a table the more likely subsequent customers will be assigned to that table. Moreover, the more customers are assigned to unoccupied tables, the more the next customer will be assigned to an unoccupied table. The consequence of these two effects is that few tables will be occupied with many customers, and many tables will be occupied by very few customers. More precisely, the behaviour of the p_i follows the so-called *power law behaviour* (Newman, 2005), very common in many natural phenomenon, such as Y-STR data. This is one of the main reasons why we decided to use this prior for Y-STR data.

More precisely:

- Let K_n denote the random number of blocks of a partition $\Pi_{[n]}$ distributed according to the Pitman sampling formula with parameters α and θ. There exists a positive random variable S_α such that
$$\lim_{n\to+\infty} \frac{K_n}{n^\alpha} = S_\alpha, \quad \text{a.s.} \tag{1.8}$$
The distribution of S_α is a generalization of the Mittag Leffler distribution (Gorenflo et al., 2014).

- If $\mathbf{P} \sim \mathrm{PD}(\alpha, \theta)$, then
$$\frac{P_i}{Z i^{-1/\alpha}} \to 1, \quad \text{a.s., when } i \to +\infty \tag{1.9}$$
for a random variable Z such that $Z^{-\alpha} = \Gamma(1-\alpha)/S_\alpha$.