Object-oriented Bayesian networks for evaluating DIP-STR profiling results from unbalanced DNA mixtures

G. Cereda[a,*], A. Biedermann[a], D. Hall[b], F. Taroni[a]

[a] *University of Lausanne, School of Criminal Justice, Institute of Forensic Science, le batochime, 1015 Lausanne-Dorigny, Switzerland*
[b] *University of Lausanne, University Center of Legal Medicine Lausanne and Geneva, Rue du Bugnon 21, 1011 Lausanne, Switzerland*

Abstract

The genetic characterization of unbalanced mixed stains remains an important area where improvement is imperative. In fact, with current methods for DNA analysis (Polymerase Chain Reaction with the SGM Plus™ multiplex kit), it is generally not possible to obtain a conventional autosomal DNA profile of the minor contributor if the ratio between the two contributors in a mixture is smaller than 1:10. This is a consequence of the fact that the major contributor's profile 'masks' that of the minor contributor. Besides known remedies to this problem, such as Y-STR analysis, a new compound genetic marker that consists of a Deletion/Insertion Polymorphism (DIP), linked to a Short Tandem Repeat (STR) polymorphism, has recently been developed and proposed elsewhere in literature [1]. The present paper reports on the derivation of an approach for the probabilistic evaluation of DIP-STR profiling results obtained from unbalanced DNA mixtures. The procedure is based on object-oriented Bayesian networks (OOBNs) and uses the likelihood ratio as an expression of the probative value. OOBNs are retained in this paper because they allow one to provide a clear description of the genotypic configuration observed for the mixed stain as well as for the various potential contributors (e.g., victim and suspect). These models also allow one to depict the assumed relevance relationships and perform the necessary probabilistic computations.

Keywords: Unbalanced DNA mixture, Bayesian networks, Object-orientation, Deletion/Insertion Polymorphism, Likelihood ratio.

1. Introduction

The common way to analyze DNA mixtures for forensic purposes is to use the Polymerase Chain Reaction (PCR) and STR markers [2]. This has proven to be a very successful technique, both for its speed and its high discriminating power. But besides its many advantages, this technique has also some drawbacks. When dealing with mixtures of two contributors, for example, the method will generally not work successfully if the proportion between the DNA of the two contributors is more extreme than 1:10 [3].[1] These situations are quite common, such as in cases of sexual assaults

*Corresponding author
Email address: `giulia.cereda@unil.ch` (G. Cereda)

[1] Here, the threshold of 10% is retained as the limit of detection of the minor DNA for blood:blood mixtures. This value varies depending on the types of biological fluids which constitute the mixture and the specific combination of genotypes present in the mixture (as reported in [4]) and should be assessed in the validation procedure [2].

when the victim's DNA is largely predominant, or in case of microchimerism during pregnancy or following organ transplant. To address this constraint, an alternative analytical method has recently been developed and proposed [1]. This method is based on the use of a new compound marker, formed by an STR marker coupled to a marker in which a Deletion/Insertion Polymorphism (DIP) is known to be present.

DIPs as such have previously been discussed in biological and biostatistical literature (e.g., identification and characterization of di-allelic polymorphisms and allelic frequencies in particular ethnies and in natural population [e.g., 5, 6, 7]), genetics (e.g., identification of DIPs as causes of genetic diseases [e.g., 8]), and forensic science (e.g., use of DIPs for analysing highly degraded DNA [e.g., 9]). The novelty of the paper here is to present an interpretative model that represents an essential element for rendering the potential of a new compound marker formed by a DIP marker coupled to an STR marker operationally useful for practitioners. The discussion will mainly concentrate on the coherent combination of the advantages of the two kinds of polymorphism, and on how this may be formally achieved through an interpretative model. In particular, this paper aims to develop and describe a probabilistic framework for the assessment of profiling results obtained with this novel typing technique, applied in the particular context of unbalanced DNA mixtures of two contributors. The approach relies on probabilistic graphical models, in particular object-oriented Bayesian networks (OOBNs). The paper also includes a discussion of this framework for two casework examples.

Section 2 provides a short description of the DIP-STR method from a biological point of view, while Section 3 describes the generic structure of the probabilistic model (i.e., OOBN) that has been built to evaluate DIP-STR profiling results. More detailed descriptions of the different structures composing the proposed OOBN are confined to Appendix A. Section 4 presents two casework examples to illustrate the kind of calculations that can be performed with the proposed graphical network environment (i.e., to obtain likelihood ratios for particular DIP-STR profiling results).They also exemplifiy the flexibility of graphical models, which are readily adapted to different scenarios. The last section presents a discussion and conclusions.

2. Genetic background

The standard method for the analysis of DNA mixtures relies on STR primers as part of a procedure that seeks to amplify only selected portions of DNA, that is regions where particular STR markers are located. STR primers are only locus-specific, not allele-specific. This means that, as the DNA of both contributors have the same loci, these primers should, in theory, anneal to both the markers of the major and to those of the minor contributor. This is, in fact, what happens whenever the minor contributor's DNA represents more than (about) the 10% of the major contributor's DNA. But, below this threshold, the minor contributor's DNA is generally not detected, as it is "masked" by the DNA of the major contributor. The difficulties, in this case, include the detection threshold of most capillary electrophoresis equipments, possible amplification biases and low template amplification conditions for the minor contributor's DNA. As a result, its signal is lost under major alleles, stutters and background noise with the consequent failure in retrieving important information.

This problem can be addressed with the use of primers that are allele-specific, to assure that – each time the two contributors have different genotypes in some marker – the primers will anneal to different alleles. This thus would avoid situations that involve competition. Based on these considerations, DIP-STRs were recently proposed as novel type of genetic marker [1]. The novelty consists on pairing a Deletion/Insertion Polymorphism (DIP) [e.g., 5] with a standard STR, to form a superlocus where the two component loci are not independent (less than 500bp apart).[2] In this way, it is possible to design two alternative allele-specific primers overlapping the DIP locus, denoted L-DIP primer and S-DIP primer. Each of these is to be used together with a primer downstream the STR region.

Hence, DIP-STR genotyping allows the selected amplification of the minor contributor's genotype (DIP-STR genotypes of minor contributors were successfully typed at ratios as low as 1:1000), as long as it has alleles that are absent in the major contributor's genotype. The best scenario is when the DNA of, respectively, the major and minor contributor 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 different minor DNA haplotypes or one, depending on the STR-homozygosis or heterozygosis of

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

DIP genotype of major/minor contributor	Number of haplotypes retrieved from minor contributor's DNA	Informativeness of genotypic configuration
Hom/Hom (different allele)	2 (if STR het)	Yes completely
	1 (if STR hom)	Yes
Hom/Het	1 (regardless STR)	Yes
Hom/Hom (same allele)	0 (regardless STR)	No
Het/Hom	0 (regardless STR)	No
Het/Het	0 (regardless STR)	No

Table 1: Informativeness of genotypic configurations. 'Hom' denotes homozygous and 'Het' heterozygous.

the minor contributor. On the other hand, when the major contributor's DNA is DIP-homozygous and the minor contributor's DNA is DIP-heterozygous, only one haplotype of the minor DNA can be retrieved (i.e., the one concerning the DIP allele opposite to the DIP allele of the major contributor's DNA).

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 have any information about the minor contributor's genotype, because both DIP primers (S and L), if used, will anneal to the major contributor's DNA. For such cases, the term *uninformative genotype* is used here. Table 1 summarizes the possible outcomes.

As a side note, it is worth mentioning that there is a traditional way to overcome the problem of strongly unbalanced mixtures in some cases. The use of Y-STR markers, for example, is of great help for cases where a male component is detected in DNA mixtures with a high female background [10]. However, Y haplotypes can be quite common in a population [11] and, if no mutations occur, patrilineal relatives of a suspect cannot be excluded as being the contributors of the stain. Recently, a panel of 13 rapidly mutating (RM) Y-STR markers has been identified [12], which successfully differentiates between closely and distantly related males. However, both the classical and the RM Y-STR techniques are useful only for a specific sex mismatch, that is if the major contributor is a women and the minor contributor is a man.

One of the advantages of the DIP-STR method over the classic STR method is that, whenever it is feasible, it detects alleles that can directly be related to the second contributor. Conversely, with the classical STR method used for mixtures of two contributors, if in some locus less than four different alleles are present, it is impossible (unless the height of the peak is taken into account and this information is reliable) to discern the alleles that belong to the second contributor, despite knowing the genotype of the first. It can only be assessed that, if the second contributor is heterozygous, he shares one allele with the first contributor, but not to decide which one. With this new method, in a case of completely informative genotypic configurations (see first row of Table 1), the complete genotype of the minor contributor can be obtained, even if this individual shares some STR alleles with the main contributor. In addition, even in case of only partially informative configurations – in which only one allele is observed (see the second and third row of Table 1) – it is certain that the detected allele belongs to the second contributor.

Finally it is important to note that, in order to carry on a DIP-STR analysis on the mixture, the only information needed from the main contributor's DNA is its DIP-heterozygosis or homozygosis.

3. An object-oriented Bayesian network (OOBN) for results of DIP-STR analyses

3.1. Evaluation of DNA profiling results using graphical models

Given a mixed DNA stain from two contributors, of which only one can be taken as known (say, the victim), and a suspect who shares alleles with the stain profile in some appropriate way, two main hypotheses may generally be of interest if the evaluation is addressed at source level [13]. One, usually that referred to as the prosecution hypothesis (H_p), asserts that the mixture originates from the victim and the suspect (if the case is such that a suspect is available for comparative examinations). A second proposition, typically put forward by the defense (H_d), states that the mixed stain comes from the victim and an unknown person.[3] In order to assess the degree to which the profiling results

[3]Here, that unknown person will be considered as unrelated to the victim.

allow one to discriminate between the latter two propositions, scientists need to focus on the likelihood ratio, defined as follows:

$$LR = \frac{P(E \mid H_p, I)}{P(E \mid H_d, I)}, \tag{1}$$

where E represents the profiling results (e.g., the genotypes of the stain, of the victim and of the suspect) and I represents the background information (i.e., the circumstances of the case). Two casework examples are proposed later in the paper, involving two different sets or profiling results. The first case covers the genotypes of a stain, a victim, and a suspect. The second case involves the genotypes of a stain and of a victim, while the suspect is supposed to be unavailable. Only his brother is available for profiling analyses.

A common interpretation of the likelihood ratio is to say that a value greater than 1 supports the prosecution hypothesis H_p, and that a value lower than 1 is in favour of the alternative hypothesis H_d. A value of 1 does not allow one to discriminate between the competing propositions of interest. As part of Bayes' theorem, the likelihood ratio connects prior odds to posterior odds in the following way:

$$\underbrace{\frac{P(H_p|E,I)}{P(H_d|E,I)}}_{\textit{Posterior odds}} = \underbrace{\frac{P(H_p \mid I)}{P(H_d \mid I)}}_{\textit{Prior odds}} \underbrace{\frac{P(E|H_p,I)}{P(E|H_d,I)}}_{\textit{Likelihood ratio}}. \tag{2}$$

This formula clarifies that the likelihood ratio, which is a measure of the probative value of the findings E with respect to two alternative hypotheses, is to be distinguished from the conditional degree of belief on the same hypotheses (represented by posterior odds). Notice that Equation (2) is a general formulation of Bayes' theorem.

When E refers to results of DNA profiling analyses, the calculation of the components of the likelihood ratio ($P(E|H_p, I)$ and $P(E|H_d, I)$) can be challenging. In relatedness testing cases, for example, the complexity of likelihood ratio formulae may be considerable, depending on parameters such as the supposed degree(s) of relatedness and the number of individuals that need to be accounted for. Moreover, formulae may vary according to genotypic configurations of the target individuals. However, this computational burden can − as shown by the foundational works by Dawid et al. [14] − be approached and safely handled through Bayesian networks to obtain the same results as those obtained by Essen-Möller's formulaic approach (focusing on posterior probabilities). In fact, Bayesian networks allow one to obtain any component defined by Equation (2). Thus, they prove to be a highly versatile framework that can accommodate analysts and reasoners with differing inferential interests [e.g., 15, 16]. Detailed accounts on Bayesian networks can readily be found in specialized literature [e.g., 17, 18, 19]. In forensic science, they are now part of well established literature as illustrated by several reports on their application for evaluating results of forensic DNA profiling analyses [e.g., 20, 14, 21, 22].

For these reasons, Bayesian networks and their object-oriented extension (i.e., OOBNs) are retained as the general modeling framework in this paper. On a practical account, the models described throughout this study have been constructed with Hugin 7.4[4] (i.e., for building OOBNs and performing calculations). The forthcoming parts of this section describe the definition of two OOBNs (i.e., the main classes) to be used to approach the probabilistic evaluation of results of the particular kind of DNA profiling analyses presented earlier in Section 2 (i.e., the findings for the two casework examples). These two OOBNs, called here `Marker` and `Marker for brother`, will focus on, respectively, two-person mixtures with a suspect being available in the first case, and the suspect being missing in the second case. Elements of the general logic underlying the structure of the proposed OOBNs are inspired by [23], but with some definitional differences to reflect the particular mechanism of functioning of the DIP-STR typing technique.

3.2. The main class `Marker`

The class `Marker` (see Figure 1) represents the main class of the OOBN proposed to model a situation in which the evidence is given by the genotypes of the stain, of the victim and of the suspect. Its main purpose is to model observations on a DNA mixture from two contributors when one of them, typically the victim, contributed more than 90% to the mixture. The remaining part is due to either the suspect or an unknown person. A collapsed version of this OOBN is given later, in Figure 3.

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

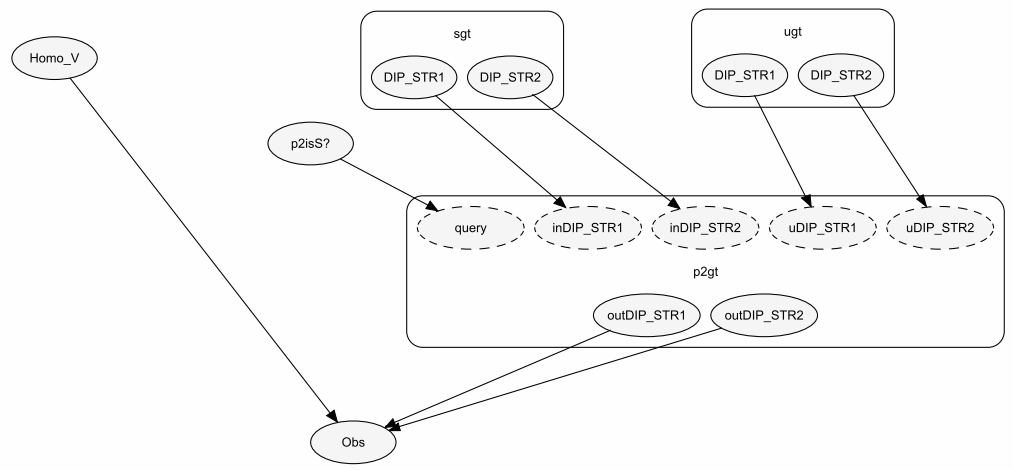


Figure 1: Expanded representation of the class `Marker`.

The left part of the network refers to the victim, represented by the single node **Homo_V**. This node has three states, *HomL*, *HomS* and *Hetero*, depending on whether the main contributor is homozygous or heterozygous for the DIP allele. As noted earlier in Section 2, this is actually the only information needed about the main contributor's genetic constitution. For purely technical reasons, the CPT of this node is completed with equal probabilities.[5]

The right part of the network (i.e., all components other than **Homo_V** and **Obs**) models the minor contributor, that could be either the suspect or an unknown person. In particular, nodes **sgt** and **ugt** are instances of the class `Genotype` (see Appendix A.2) and represent the genotype of, respectively, the suspect and an unknown person.

The Boolean node **p2isS?** addresses the question of whether the second contributor is the suspect or an unknown person. Again, for technical reasons and invoking the same arguments as in footnote 5, equal probabilities are assigned to the CPT of this node. Node **p2gt** is an instance of the class `Pgt` (see Appendix A.4) and represents the genotype of the actual second contributor to the mixture.

Node **Obs**, with states *La, Lb, Lab, Sa, Sb, Sab, X, nr*,[6] represents the observed (minor contributor's) DIP-STR allele(s) in the trace. These states, except *nr*, represent the results obtained when analysing the trace using the DIP primer opposite to the DIP allele of the major contributor's genotype and when one of the situations described in the first three rows of Table 1 holds. In turn, the state *nr* (short for 'not revealed') represents a not observed genotype, that is a result obtained whenever the major contributor is DIP-heterozygous or both contributors are DIP-homozygous of the same type.[7]

The state of the node **Obs** depends on the state of the parent node **Homo_V** that, combined with the state of the parent nodes **outDIP_STR1** and **outDIP_STR2**, determines how many STR alleles of the second contributor will appear in the result. If **Homo_V** is in the state *Hetero*, no DIP-STR profiling results can be obtained for the mixed stain. The same is the case if the first and the second contributor are DIP-homozygous of the same type. In these cases, the node **Obs** will assume the state *nr*. If the node **Homo_V** is in a state other than *Hetero*, and the second contributor is DIP-heterozygous, then only the DIP-STR allele with the DIP allele opposite to that of the first contributor is revealed. The last case is the one in which both contributors are DIP-homozygous of different type: in this case the observed minor contributor's genotype is composed by a couple of different DIP-STR alleles if it is STR

[5] As will be shown in Section 4, this node will be instantiated when evaluating components of the likelihood ratio so that the initial values in its CPT are no longer of importance. The initial probabilities in the CPT are only needed from a definitional point of view, in order to build a complete model.

[6] The meaning of the letters *a*, *b* and *x* will be explained in further detail in Appendix A.1.

[7] The term *nr* represents a situation in which no alleles of the minor contributor are revealed, due to a particular combination of the genotypes of the two contributors (see rows 3, 4 and 5 of Table 1). Situations in which no alleles are revealed, due to low-template traces or other problems with the PCR process, are not taken into account in the paper.

heterozygous, otherwise is composed by a single DIP-STR allele as in the previous case. Table 2 shows part of the CPT for the node **Obs**.[8]

3.3. The main class `Marker for brother`

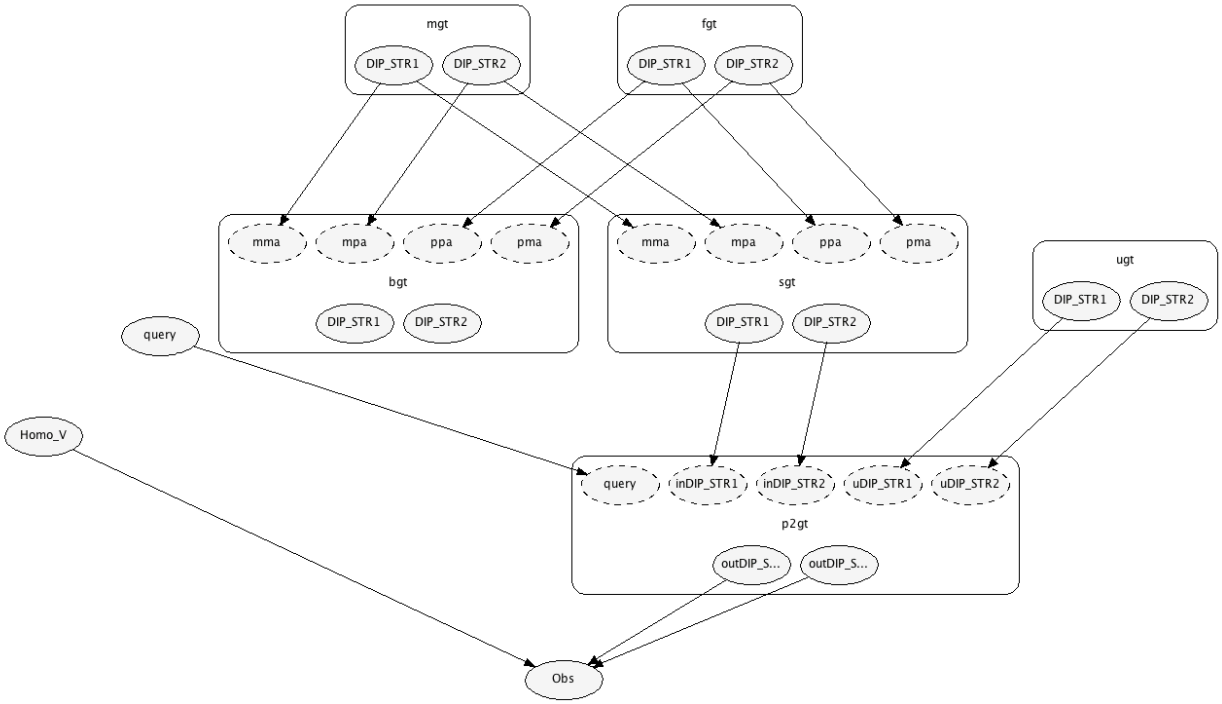


Figure 2: Expanded representation of the class `Marker for brother`.

The class `Marker for brother` (see Figure 2) represents the overall structure of the OOBN, proposed to model a situation in which profiling results consist of the genotypes of the stain, the victim, and the brother of the suspect. Its main purpose is to model profiling results for a DNA mixture from two contributors when one of them, typically the victim, contributed more than 90% to the mixture, and the suspect's DNA is not available for comparison. To deal with these missing data, nodes for the suspect's (full) brother genotype, supposed to be known, have been added. These additional nodes have been logically combined with the network through nodes representing the genotype of the brother's parents (which are also the parents of the suspect). A collapsed version of this OOBN is given later in Figure 5. Nodes **mgt**, **fgt** and **ugt** are instances of the class `Genotype` (see Appendix A.2) and represent the genotype of, respectively, the suspect's mother, the suspect's father and an unknown person. Most of the nodes which are in common with the class `Marker` have the same definition, except for the node **sgt**, which, together with node **bgt**, is an instance of the class `Child` (see Appendix A.3).

4. Casework examples

4.1. General case description and DIP-STR analyses

Suppose a case in which the body of a dead women is found [1]. Circumstantial evidence leads to three suspects: a man and his two sons. Other information supports the possibility of a single perpetrator, and this information is used

[8]The state *X* of the node **Obs** summarizes all cases in which the results of the analysis of the mixed stain show an allele with letter *x*: *Lx, Lax, Lbx, Sx, Sax, Sbx*, as explained in further detail in Appendix A.1.

Homo_V	*HomL*																								
outDIP_STR1	*La*						*Lb*						*Lx*						*Sa*						...
outDIP_STR2	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	...
La	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
Lab	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
Lb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
Sa	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1	1	1	0	0	...
Sab	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	...
Sb	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	...
X	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	...
nr	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	...

Homo_V	*HomS*																								
outDIP_STR1	*La*						*Lb*						*Lx*						*Sa*						...
outDIP_STR2	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	...
La	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	...
Lab	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
Lb	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	...
Sa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
Sab	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
Sb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
X	0	0	1	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1	0	0	1	0	0	0	...
nr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	...

Homo_V	*Hetero*																								
outDIP_STR1	*La*						*Lb*						*Lx*						*Sa*						...
outDIP_STR2	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	...
La	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Lab	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Lb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Sa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Sab	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Sb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
nr	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	

Table 2: Partial representation of the way in which the CPT of node **Obs** is completed, depending on the possible state configurations of the parental nodes **Homo_V**, **outDIP_STR1** and **outDIP_STR2**.

as an assumption in the subsequent evaluation of analytical results. A relevant blood stain − denoted A here − was collected on the victim's body. Blood of the victim and of the three suspects was also available for analysis. Using the standard protocols (autosomal STR multiplex and Y-STR), the analyses led (i) to a complete autosomal STR profile matching the victim's DNA profile (without any indication of a mixed profile), and (ii) to a complete Y-STR profile, matching all the three suspects. None of the three suspects can thus be excluded as a potential contributor to the detected DNA stain. Further, it is assumed here that there are only two contributors to the DNA trace.

With the aim of discriminating between the three male suspects, it has been decided to analyze three DIP-STR loci: MID1950-D20S473, MID1107-D5S1980, MID1013-D5S490, called here Marker 1, Marker 2 and Marker 3, respectively. Table 3 summarizes the DIP-STR profiles of the victim and of the three suspects.

	Marker 1	Marker 2	Marker 3
Victim	S12-S13	L13-L14	S14-S14
Father	S11-S11	L13-L13	S14-S14
Son 1	S11-L12	L13-L13	S14-S14
Son 2	S11-S12	S19-L13	S14-S14

Table 3: DIP-STR genotypes of the victim and the three suspects.

Since the victim is DIP-homozygous (S-S, L-L and S-S) in the three selected loci, it is possible to genotype the mixture with the opposite DIP-alleles: L for Marker 1, S for Marker 2 and L for Marker 3. The results are as follows:

$$\text{Stain A:} \quad \text{Marker 1}=\{\text{L12}\}, \quad \text{Marker 2}=\{nr\}, \quad \text{Marker 3}=\{nr\}.$$

Comparing these results for DIP-STR markers to the DIP-STR genotypes of the three suspects, it can be seen that, at Marker 3, the DIP-genotypes of all suspects are compatible with the result *nr* for the bloodstain. At Marker 2, this happens only for the DIP-alleles of Father and Son 1. Indeed, if Son 2 contributed to the mixture, then S19 would appear in the results. Using a similar argument for Marker 1, only the DIP-genotype of Son 1 is compatible with the result for the blood stain.

Based on these DIP-STR results, Son 2 and Father can thus be excluded as contributors to the mixed stain recovered on the victim's body. This leaves only Son 1 as a potential contributor (among the individuals for which analyses have been performed), and this leads to questions of the following kind: What is the meaning of such a non-exclusion? What is the degree of support for the proposition according to which Son 1 contributed to the crime stain? The forthcoming sections will approach such questions through the use of OOBNs and two distinct case settings. Propositions of interest can now be expressed as 'the mixed stain is made up of the DNA of the victim and Son 1' (H_p) and 'the mixed stain is made up of the DNA of the victim and an unknown person, unrelated to Son 1' (H_d).

4.2. Case 1: suspect available

4.2.1. Preliminaries

In order to build a network for illustrating the application of the OOBN modeling procedure to Case 1, instances of the class networks defined in Appendix A.1, Appendix A.2 and Appendix A.4 have been combined to form the overall network called `Marker` (as introduced earlier in Section 3.2). Figure 3 shows this network in a collapsed version.

In this first example, the genotype of the suspect is considered as known. The available items of evidence for which the probative value is to be calculated, are shown in Table 4.

	Marker 1	Marker 2	Marker 3
Victim	S12-S13	L13-L14	S14-S14
Suspect (Son 1)	S11-L12	L13-L13	S14-S14
Stain	L12	nr	nr

Table 4: Profiling results for Case 1.

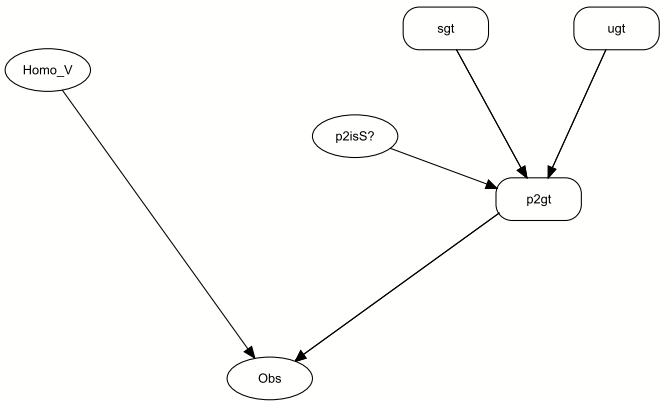


Figure 3: Collapsed representation of the proposed OOBN for the evaluation of DIP-STR profiling results.

Before using this OOBN for inference, one needs to decide about the definition of the states of the node **DIP_STR**, depending on the genotype of the suspect and on the results obtained after the analysis of the trace. To do this, one should take into account both the alleles observed in the suspect's profile and those in the profile of the mixture. There are 4 possible situations, summarized in Table 5:

- If only one allele is observed (say L12 for Marker 1) in both analyses (i.e. for the stain and the suspect), then one can refer to this with the state *La* and put probability 0 to the states relative to the alleles *Sa Sb* and *Lb* in the node **DIP_STR** of the class `DIP_STR`. This is shown in the first row of Table 5.
- If two alleles are observed, such as L12 L11, L12 S12 or L12 S11, the states *La Lb*, *La Sa* or *Sa Sb* can be used to represent them. In such a case, a probability equal to zero would be set for the remaining *a* and *b* alleles.

The assignment of probabilities to states *Sx* and *Lx* follows the explanations given in Section Appendix A.2. Reminding how DIP-STR analyses work, if three (or more) different alleles are observed following the analysis of the suspect and the mixture, then the suspect can be excluded from being a contributor (under the assumption of a two person mixture). In the case of Marker 1, for example, one can use the probability of the allele L12 for the state *La* of **DIP_STR**, and the probability of the allele S11 for the state *Sb*. The probability for states *Lb* and *Sa* will be set to 0, while the probability for the states *Lx* and *Sx* will be calculated by adding the probabilities of all the other L- (and S-) alleles. A summary of this is given in row three of Table 5.

The described procedure of assigning letters *a*, *b* and *x* and the correct probabilities to the states of node **DIP_STR** may be viewed as time-demanding or prone to errors. It is for this reason that an R function, written with the package RHugin [24] is available for the interested readers (requests by e-mail to the Corresponding author). The function is called `Dipstr_LR` and is of the following form:

```
Dipstr_LR<-function(Marker_name, vgt1, vgt2, sgt1, sgt2, Obs1, Obs2)
```

where one has only to specify the marker to be considered, the genotype of the victim, of the suspect and the two alleles observed in the trace.

4.2.2. Likelihood ratios for DIP-STR profiling results: instantiating the OOBN

To obtain a likelihood ratio, scientists need to evaluate two conditional probabilities: $P(E|H_p, I)$ and $P(E|H_d, I)$. Here, the variable E refers to the results for the trace (E_{obs}) and the genotype of the victim and the suspect. The latter two genotypes will be denoted E_g for short. The likelihood ratio (LR) can thus be written as follows:

$$LR = \frac{P(E \mid H_p, I)}{P(E \mid H_d, I)} = \frac{P(E_{obs}, E_g \mid H_p, I)}{P(E_{obs}, E_g \mid H_d, I)} = \frac{P(E_{obs} \mid H_p, E_g, I)P(E_g|H_p, I)}{P(E_{obs} \mid H_d, E_g, I)P(E_g|H_d, I)} = \frac{P(E_{obs} \mid H_p, E_g, I)}{P(E_{obs} \mid H_d, E_g, I)}.$$

Observation for the stain	Suspect genotype	La	Lb	Lx	Sa	Sb	Sx
L12	L12 L12	γ_{L12}	0	$\sum_{j\in J\setminus\{12\}}\gamma_{Lj}$	0	0	$\sum_{k\in K}\gamma_{Sk}$
L12	L12 S12	γ_{L12}	0	$\sum_{j\in J\setminus\{12\}}\gamma_{Lj}$	γ_{S12}	0	$\sum_{k\in K\setminus\{12\}}\gamma_{Sk}$
L12	L12 S11	γ_{L12}	0	$\sum_{j\in J\setminus\{12\}}\gamma_{Lj}$	0	γ_{S11}	$\sum_{k\in K\setminus\{11\}}\gamma_{Sk}$
L12 L11	L12 L11	γ_{L12}	γ_{L11}	$\sum_{j\in J\setminus\{11,12\}}\gamma_{Lj}$	0	0	$\sum_{k\in K}\gamma_{Sk}$

Table 5: Possible probability assignments for the states La, Lb, Lx, Sa, Sb et Sx of the node **DIP_STR**. J is the set of all the possible STR alleles linked to the DIP allele L, that can be present in the marker of interest and K is the set of all the possible STR alleles linked to the DIP allele S. The choice of the numbers 11 and 12 serves the sole purpose of illustration. An analogous table can be built to model situations in which DIP alleles of the type S are observed in the stain.

The last equality is obtained by invoking the assumption that the victim's and the suspect's genotype (represented by E_g) do not depend on whether the suspect is or is not a contributor to the mixed stain, given the background information I.

To obtain a value for the numerator of the likelihood ratio, instantiations should be made in the nodes **Homo_V?**, **DIP_STR1** and **DIP_STR2** of the class `sgt`, and **p2isS?**. In particular, the latter node needs to be set to the state *True*, because under H_p it is the suspect who is assumed to be the second contributor. The probability of observing a given DIP-STR configuration for the mixed stain − under this conditioning − is then read from the node **Obs**. In turn, a value for the denominator is obtained by instantiating the node **p2isS?** to the state *False* and then, again, reading the required conditional probability in the node **Obs**.[9] The ratio between the two numbers thus found gives the likelihood ratio.

Figure 4 illustrates these instantiations and propagations in terms of the OOBN `Marker`, used to represent the results obtained in the currently discussed casework example for Marker 1. Figure 4(a) illustrates how to obtain the numerator, whereas Figure 4(b) illustrates the evaluation of the denominator. Notice that the instantiations made in the nodes **DIP_STR1** and **DIP_STR2** of the class `sgt` are not visually displayed because they are operated inside instance nodes. Prior to using the OOBN for these propagations, one needs to tailor the probability table of the node **DIP_STR**. In view of the observation that, for Marker 1, the mixed stain shows L12 and the suspect has the genotype S11-L12, probabilities as shown in row three of Table 5 need to be specified. On the basis of internal data collected at the authors' institution (on 100 individuals), the following vector of probabilities was assigned:

$$P(\{La, Lb, Lx, Sa, Sb, Sx\}) = \{0.181, 0, 0.208, 0, 0.259, 0.352\}.$$

The probabilities for the state La of the node **Obs** found in the described way lead to the following likelihood ratio:

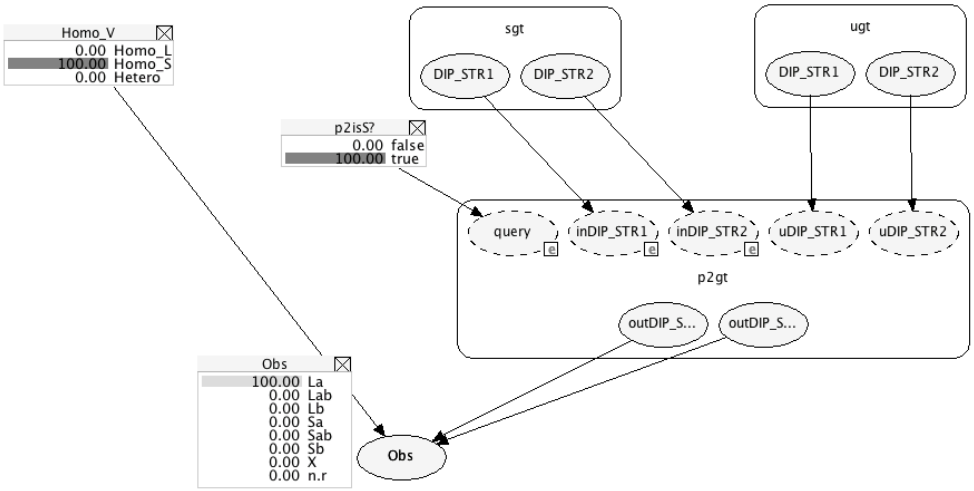
$$LR = \frac{P(E_{obs} \mid H_p, E_g, I)}{P(E_{obs} \mid H_d, E_g, I)} = \frac{P(\textbf{Obs} = La \mid \textbf{P2isV?} = True, \textbf{Homo\_V} = HomS, \texttt{sgt} = \{La, Sb\}, I)}{P(\textbf{Obs} = La \mid \textbf{P2isV?} = False, \textbf{Homo\_V} = HomS, I)} \sim \frac{1}{0.2539} \sim 3.95. \tag{3}$$

The numerator of this result can readily be understood. If the suspect, whose genotype is L12-S11, is truly the second contributor, and the analyst analyses the trace using primers for the L-DIP, then it can reasonably be expected that L12 will be detected in the stain.[10] Assuming no disturbing or otherwise complicating factors during the analyses, a numerator of 1 can thus be found. For the denominator, further considerations are required. Under the assumption of a contributor other than the suspect, one needs to consider several possible genotypes. In fact, the second contributor could have any of the following genotypes: $La - La, La - Sb, La - Sx$, for $a = 12$, $b = 11$ and $x \neq 11, 12$. The value of the denominator is thus given by the sum of the probabilities of these genotypes. This leads to the following expression of the likelihood ratio:

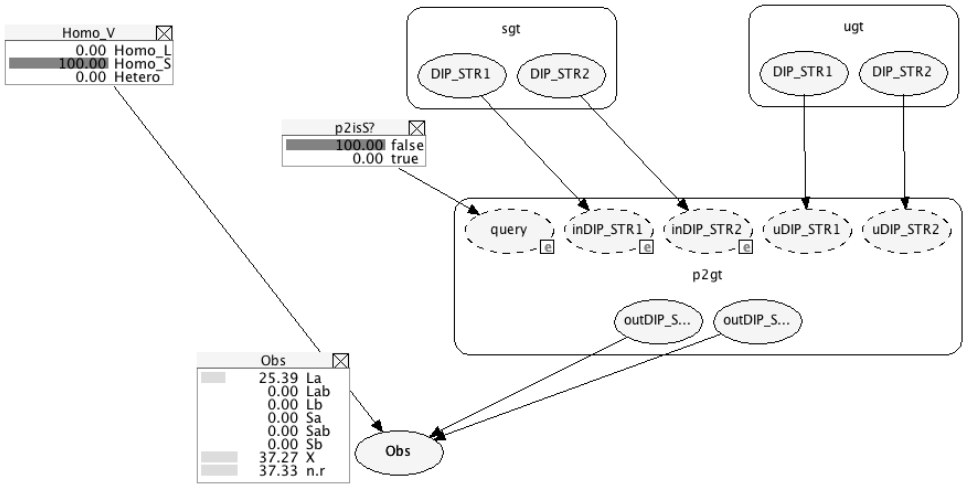
$$LR = \frac{P(E_{obs} \mid H_p, E_g, I)}{P(E_{obs} \mid H_d, E_g, I)} = \frac{1}{\gamma_{La}^2 + 2\gamma_{La}\gamma_{Sb} + 2\gamma_{La}\gamma_{Sx}} \sim \frac{1}{0.2539} \sim 3.95. \tag{4}$$

[9] Notice that given **p2isS?**=*False*, information about the genotype of the suspect becomes irrelevant. That is, any instantiation made in the nodes modeling the suspect's genotype will not affect the probabilities obtained at the node **Obs**.

[10] Recall that no alleles from the victim will be detected because he has only S-DIP alleles.



(a) Under the proposition H_p, the node **p2isS?** is set to *True*.



(b) Under the proposition H_d, the node **p2isS?** is set to *False*.

Figure 4: OOBN `Marker` used to evaluate profiling results obtained for Marker 1 where (a) represents the view under the first proposition (i.e., H_p) and (b) represents the the view under the alternative proposition (i.e., H_d).

Marker:	Marker 1	Marker 2	Marker 3	Combined
Likelihood ratio:	3.9	2.2	1.8	~ 15

Table 6: Summary of the likelihood ratios obtained for profiling results on three DIP-STR markers for Case 1, rounded to one decimal.

This result demonstrates that the OOBN-output is not arbitrary, but can be reproduced as a result of logical considerations. It is also worth mentioning that the result can also be related to existing literature on qualitative mixture assessment as described by Weir et al. [25] (based on Evett [26]). Although the formulae derived in these references are intended to evaluate traditional STR profiling results, their underlying logic also applies to DIP-STR results: the aim is to find the probability that a given number of persons possess – in combination – particular alleles (here: DIP-STR alleles).

In the same way as outlined here above, one can also find likelihood ratios for DIP-STR profiling results on Marker 2 and Marker 3. Table 6 summarizes these results, as well as the overall likelihood ratio. The latter value is obtained by multiplication because the DIP-STR markers are assumed to be independent, in the same way as traditional STR markers. For the time being, the results summarized in Table 6 should be taken as provisional because the collection of relevant data, in a more extensive form, is still underway.

The probabilistic analyses conducted in this section, such as Equation 4, may appear elementary. However, they may become tedious if they need to be done manually. The reason for this is that, for each marker, distinct allelic configurations may be observed so that the formulaic development may take different forms. The advantage of using an OOBN thus becomes immediately clear. Except for the input values (i.e., the initial numerical specification), the model structure remains constant. Moreover, the analyst can confine computations entirely to the model. Thus, the use of an OOBN could also help to make evaluative procedures less prone to possible errors. This is further clarified in the next casework example (Section 4.3), where the suspect's genotype is supposed to be unavailable. Generally, formulae readily become more complicated in such settings, depending on the degree of relatedness between the suspect and the typed individuals.

4.3. Case 2: missing suspect

This example uses some of the data of Case 1. As a main difference, it is supposed that the genotype of the suspect, as well as that of his father, are not available. Table 7 provides a summary of the available profiling results.

	Marker 1	Marker 2	Marker 3
Victim	S12-S13	L13-L14	S14-S14
Brother (Son 2)	S11-S12	S19-L13	S14-S14
Stain	L12	nr	nr

Table 7: Profiling results for Case 2.

The network specifications and the definition of the states of the different nodes of the class `Marker for brother` are as described in Section 3.3, and displayed in collapsed form in Figure 5. Table 8 summarizes the likelihood ratios obtained for Case 2.

4.4. A note on the likelihood ratio results

In Case 1, an overall likelihood ratio of about 15 was obtained. This means that the findings (i.e., results for three markers) support the proposition according to which the victim and the suspect Son 1 are the two contributors to

Marker:	Marker 1	Marker 2	Marker 3	Combined
Likelihood ratio:	0.6	0.6	1.4	~ 0.5

Table 8: Summary of the likelihood ratios obtained for profiling results on three DIP-STR markers for Case 2, rounded to one decimal.

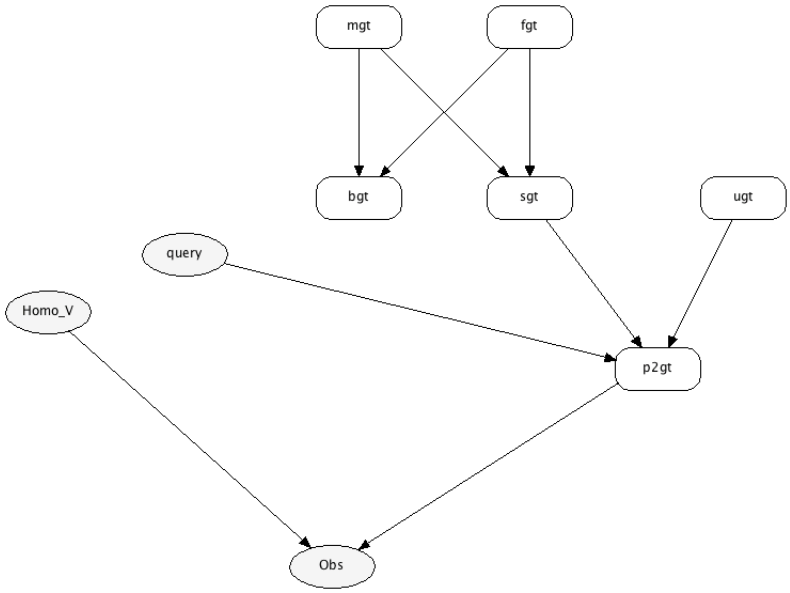


Figure 5: Collapsed representation of the class `Marker for brother`.

the mixture. Although this likelihood ratio is rather moderate, it is useful to note some further aspects of this result. First, when the case was actually examined, only three DIP-STR markers had been used. At the moment, a panel of 9 markers is available which could allow one to obtain higher likelihood ratios. Moreover, the likelihood ratio result is higher than the likelihood ratio of one that would be obtained in situations where the classical STR method does not yield any profiling output, which typically occurs with strongly imbalanced mixtures. Finally, one should also consider that even in presence of a *nr* result for the stain, when the victim is DIP homozygous a likelihood ratio higher than one is obtained. This is so because the result indicates that the victim and the second contributor are DIP homozygous of the same kind (both S-S or L-L).

Often, there are unfortunate expectations in the field that forensic DNA must necessarily be (highly) probative, but this should not distract us from devoting attention to alternative profiling techniques that usefully complement the broad range of approaches available to the forensic practitioner. Moreover, the resulting likelihood ratio can be aggregated with the result obtained from Y-STR profiling analyses, using the same couple of target propositions.

In Case 2, an overall likelihood ratio of about 0.5 was obtained. This means that the findings (i.e., results for three markers) slightly support the proposition according to which Son 1 is not a contributor to the mixture. This result is not unreasonable since markers in which the brother's genotype is incompatible with the stain results tend to lead to a LR< 1: since the suspect is genetically close to his brother, this observation will also hold for the suspect.

5. Discussion and conclusions

Unbalanced DNA mixtures are problematic for traditional STR profiling analyses, in particular when the proportion between the DNA of the two contributors is more extreme than 1:10 [3]. Cases of sexual assaults (where the victim's DNA is predominant and that of the aggressor is present only as a minor quantity) or cases of micro chimerism during pregnancy (where minute quantities of fetal DNA are present in maternal blood) are typical examples for situations in which stains of this type may be found. To cope with this constraint, recent developments focused on alternative analytical methods using a new compound marker, formed by a STR marker coupled to a DIP [1]. A particular feature of DIP-STR markers is that, whenever they can be analyzed, they can detect alleles directly related to the second contributor. However, the successful detection of DIP-STR alleles in an unbalanced mixed stain depends on how the DIP-STR genotypes of the stain contributors compare to each other, as there may be situations in which

none or only part of the target DNA of the individual of interest (i.e., different from the assumed known contributor, such as a victim) can be detected.

In order to make this novel DNA profiling technique applicable in forensic contexts, one needs to be able to assess the meaning of particular profiling results with respect to selected competing propositions. Examples include 'the victim and the suspect contributed to this DNA mixture' versus 'the victim and an unknown person contributed to this DNA mixture'. This paper has investigated the use of graphical probability models (i.e., Bayesian networks), in particular OOBNs, to address questions of this kind. OOBNs have been chosen because they allow one to derive a concise representation of the genotypic configuration of the various (assumed) contributors as well as the mixed stain. Most importantly, such graphical networks allow one to depict the way in which the assumed contributors' genotypes relate to that of the crime stain. On a computational account, such models also allow their user – and this is one of the main features of OOBNs – to find the components of likelihood ratios that express the probative value for particular findings (i.e., DIP-STR profiling results). OOBNs can thus help scientists to deal with the (often complex) calculations that are encountered with DNA mixtures. For example, an OOBN will require only minimal initial specifications in order to approach the typing results for a given marker. Typically, these initial specifications will relate to the probabilities assigned for the various alleles, but with regards to the qualitative graph structure, the model should not require any changes. This is different for purely formulaic approaches to evaluation because these may take various different forms, depending on the particular profiling results (for both the potential contributors as well as the mixed crime stain), and may thus be less practical in their application – eventually also more prone to error (if they need to be done manually). Moreover, as pointed out in [14], the advantage of using a graphical probabilistic approach becomes evident in cases where genetic information of further individuals (other than the suspect) need to be considered (typically when the suspect is missing, and information on his genotype is not available). A purely arithmetic solution to such problems may become increasingly challenging. Such a situation is encountered in Case2. The corresponding OOBN shows how the brother's genotype can be considered through a very straightforward modification of the OOBN's structure.

The rather moderately sized likelihood ratio values obtained for the reported casework examples should not be interpreted as a limiting factor in principle. Indeed, it is worthwhile to emphasize that (i) the described profiling technique (DIP-STR) works with particularly high reliability under special circumstances implied by unbalanced mixtures (at least in the case in which the two contributors are of the same gender, or the minor contributor is a female), (ii) potential stain contributors could be excluded, and (iii) the probative value for non-excluded individuals can be characterized probabilistically. Future research in the authors' institution will focus on investigating further markers of this kind, as well as the generation of relevant population data to improve the numerical specification of the proposed OOBN-approach.

6. Acknowledgements

This research was supported by the Swiss National Science Foundation, through grant no. 105311- 1445570.

7. References

References

[1] V. Castella, J. Gervaix, D. Hall, DIP-STR: Highly sensitive markers for the analysis of unbalanced genomic mixtures, Human Mutation 34 (2013) 644–654.
[2] J. Butler, Advanced Topics in Forensic DNA Typing: Methodology, Elsevier Academic Press, New York, 2011.
[3] T. Clayton, J. Buckleton, Mixtures, in: J. Buckleton, C. Triggs, S. J. Walsh (Eds.), Forensic DNA evidence interpretation, CRC Press, Boca Raton, 2005, Chapter 7, pp. 217–274.
[4] Applied Biosystems, AmpFlSTR Profiler Plus PCR Amplification Kit User's Manual, Foster City, California (2012).
[5] J. L. Weber, D. David, J. Heil, Y. Fan, C. Zhao, G. Marth, Human diallelic insertion/deletion polymorphisms, The American Journal of Human Genetics 71 (2002) 854–862.
[6] U. Vali, M. Brandstrom, M. Johansson, H. Ellegren, Insertion-deletion polymorphisms (indels) as genetic markers in natural populations, BMC Genetics 9 (2008) 8.
[7] A. M. Neuvonen, J. U. Palo, M. Hedman, A. Sajantila, Discrimination power of investigator dipplex loci in Finnish and Somali populations, Forensic Science International: Genetics 6 (2012) e99 – e102.
[8] D. N. Cooper, M. Krawczak, Mechanisms of insertional mutagenesis in human genes causing genetic disease, Human Genetics 87 (1991) 409–415,

[9] P. A. da Costa Francez, E. M. Ribeiro Rodrigues, A. M. de Velasco, S. E. Batista dos Santos, Insertion-deletion polymorphisms-utilization on forensic analysis, International Journal of Legal Medicine 126 (2012) 491–496.
[10] L. Roewer, Y chromosome STR typing in crime casework, Forensic Science, Medicine, and Pathology 5 (2009) 77–84.
[11] M. Vermeulen, A. Wollstein, K. van der Gaag, O. Lao, Y. Xue, Q. Wang, L. Roewer, H. Knoblauch, C. Tyler-Smith, P. de Knijff, M. Kayser, Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms, Forensic Science International: Genetics 3 (2009) 205–213.
[12] K. N. Ballantyne, V. Keerl, A. Wollstein, Y. Choi, S. B. Zuniga, A. Ralf, M. Vermeulen, P. de Knijff, M. Kayser, A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages, Forensic Science International: Genetics 6 (2012) 208–218.
[13] R. Cook, I. Evett, G. Jackson, P. Jones, J. Lambert, A hierarchy of propositions: deciding which level to address in casework, Science & Justice 38 (1998) 231–239.
[14] A. P. Dawid, J. Mortera, V. L. Pascali, D. van Boxel, Probabilistic expert systems for forensic inference from genetic markers, Scandinavian Journal of Statistics 29 (2002) 577–595.
[15] U. B. Kjærulff, A. L. Madsen, Bayesian networks and Influence Diagrams, A Guide to Construction and Analysis, Springer, New York, 2008.
[16] F. Taroni, C. Aitken, P. Garbolino, A. Biedermann, Bayesian networks and probablistic inference in forensic science, John Wiley & Sons, Chichester, 2006.
[17] R. E. Neapolitan, Probabilistic Reasoning in Expert Systems, John Wiley & Sons, Inc., New York, 1990.
[18] M. Jordan, Learning in graphical models, MIT Press, Cambridge, 1998.
[19] J. Pearl, Reverend Bayes on inference engines: A distributed hierarchical approach, in: Proceedings of the National Conference on Artificial Intelligence (AAAI'82), Morgan Kaufmann, Pittsburgh, 1982, pp. 133–136.
[20] A. Biedermann, F. Taroni, Bayesian networks for evaluating forensic DNA profiling evidence: A review and guide to literature, Forensic Science International: Genetics 6 (2012) 147–157.
[21] R. Cowel, S. Lauritzen, J. Mortera, Object-oriented Bayesian networks for DNA mixture analyses (2006).
[22] R. Cowell, S. Lauritzen, J. Mortera, Probabilistic expert systems for handling artifacts in complex DNA mixtures, Forensic Science International: Genetics 5 (2011) 202 – 209.
[23] A. P. Dawid, J. Mortera, P.Vicard, Object-oriented Bayesian networks for complex forensic DNA profiling problems, Forensic Science International, 169 (2007) 195–205.
[24] K. Konis, RHugin: RHugin, R package version 7.4/r247 (2010).
URL `http://rhugin.r-forge.r-project.org/`
[25] B. Weir, C. Triggs, J. Buckleton, K. Walsh, L. Stowell, L. Starling, Interpreting DNA mixtures, Journal of Forensic Sciences 42 (1997) 213–222.
[26] I. W. Evett, C. Buffery, G. Willott, D. Stoney, A guide to interpreting single locus profiles of DNA mixtures in forensic cases, Journal of the Forensic Science Society 31 (1991) 41–47.
[27] I. Evett, B. Weir, Interpreting DNA evidence: Statistical Genetics for Forensic Scientists, Sinauer Associates, Sunderland, 1998.
[28] F. Taroni, S. Bozza, A. Biedermann, P. Garbolino, C. Aitken, Data Analysis in Forensic Science: A Bayesian Decision Perspective, Statistics in Practice, Wiley, 2010.
[29] A. C. Brandwein, W. Strawderman, Bayesian estimation of multivariate location parameters, in: Handbook of Statistics, Vol. 25, Elsevier, 2005, pp. 221–244.

Appendix A. Details on the OOBN to model DIP STR results

This appendix describes in details the classes which are contained in the main class `Marker` (see Section 3.2)

Appendix A.1. The class `DIP_STR`

Figure A.6: Representation of the class `DIP_STR`.

The class `DIP_STR` (see Figure A.6) consists of a single output node **DIP_STR** whose states represent the different DIP-STR alleles, denoted here *La, Lb, Lx, Sa, Sb* and *Sx*. The CPT contains the probability of occurrence of these alleles in the relevant population. In order to be coherent with a Bayesian approach, a Bayesian estimation is used here for the allele proportions, based on a prior Dirichlet distribution for those proportions [e.g., 27, 28, 29].[11] At this point, the assumption of non-independence between DIP and STR markers, made earlier in Section 2, becomes explicit.

[11] In what follows, the Bayesian estimate for the *Li* allele proportion is referred to as γ_{Li}.

This understanding is conveyed by using only a single node **DIP_STR**, rather than two separate nodes. Letters *a*, *b* and *x* are used instead of the list of the actual STR allele numbers in order to use the model for different markers, and to facilitate computational tasks.

Note that, according to currently available data, there are loci with more than six different alleles. To handle this, states *La*, *Lb*, *Sa* and *Sb* are used to represent the two alleles that, at most, could be observed with the DIP-STR method,[12] while states *Lx* and *Sx* represent all the alleles that, in principle, may appear in that locus but are not actually observed. This means that, before deciding which alleles to associate with the states of the node **DIP_STR**, one will consider the results of the analysis performed on the DNA mixture (see Section 4.2). The probability of states *Lx* (and *Sx*) is set as the sum of all the probabilities of the unobserved L-STR alleles (S-STR alleles). As already explained, the state *X* of the node **Obs** is reserved for results that show an allele corresponding to letter *x*. Practically, this will not be the case, since *x* represents the not observed STR alleles, so that the state *X* may be said to have a dummy function. It is needed only for structural reasons, in order to complete the network from a definitional point of view.

Appendix A.2. The class `Genotype`

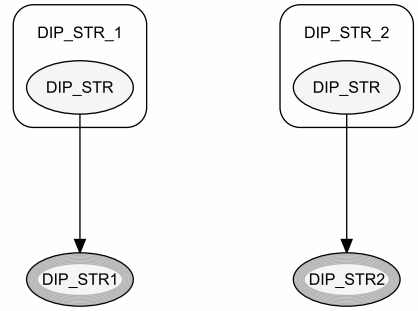


Figure A.7: Representation of the class `Genotype`.

The class `Genotype` (see Figure A.7) is used in the main class `Marker` to represent genotypes of different individuals involved in the case, when there is no need to include an explicit representations of their parents. This class is itself composed by instance nodes, namely **DIP_STR1** and **DIP_STR2**. These are instances of the class `DIP_STR` and represent the allelic constitution, on the two chromosomes, of the given person. The class `Genotype` also contains output nodes, called **DIP_STR1** and **DIP_STR2**, that are copies of their parent nodes.[13] This class is used in the main class `Marker` through two instances (**sgt** and **ugt**) representing, respectively, the genotype of the suspect and of an unknown person. It is also used in the main class `Marker for brother` through the instances **mgt**, **fgt**, and **ugt**.

Appendix A.3. The class `Child`

The class `Child` (FigureA.8) is used in the main class `Marker for brother` to represent genotypes of individuals for which it is necessary to include their parents explicitly (i.e., the genotype of a person of interest is presented as a child variable depending on the genotypic configuration of the parents). This class contains (i) two output nodes, called **DIP_STR1** and **DIP_STR2**, that represent the allelic constitution (on the two chromosomes) of the given person, (ii) two input nodes **mpa** and **mma** which represent the two alleles possessed by the mother (one of which is inherited by the child), and (iii) two input nodes **fpa** and **fma** which represent the two alleles possessed by the father (one of which is inherited by the child). This class is used in the main class `Marker for brother` through two instances (**sgt** and **bgt**) representing, respectively, the genotype of the suspect and of the brother of the suspect.

Figure A.8: Representation of the class `Child`.

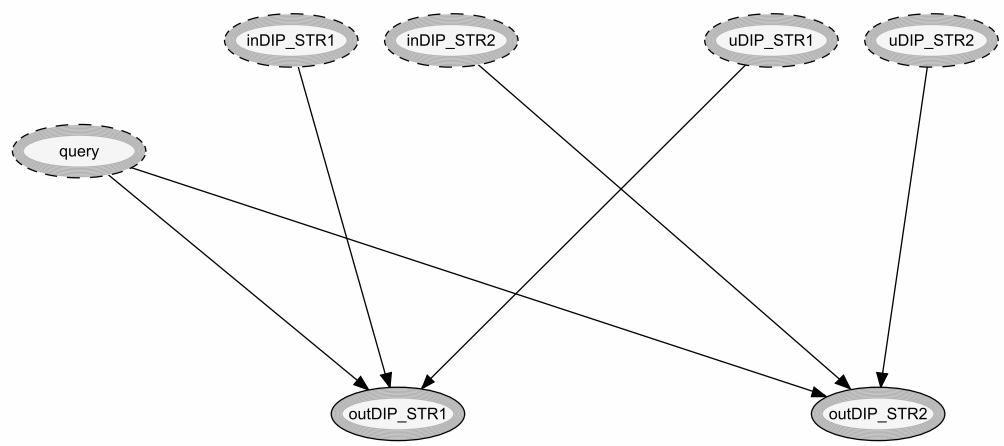


Figure A.9: Representation of the class `Pgt`.

Appendix A.4. The class `Pgt`

The class `Pgt` (see FigureA.9) is used to represent the allelic configuration, on the two chromosomes, of the actual second contributor to the mixture. It is composed by input and output nodes. The input nodes are **query**, that is a Boolean node, bounded to the node **p2isS?** of the external network (e.g., Figure 1). The nodes **inDIP_STR1** and **inDIP_STR2** are bound, respectively, to the nodes **DIP_STR1** and **DIP_STR2** of the instance **sgt** of the class `Genotype`. The nodes **uDIP_STR1** and **uDIP_STR2** are related, respectively, to the nodes **DIP_STR1** and **DIP_STR2** of the instance **ugt** of the class `Genotype`. The output nodes **outDIP_STR1** and **outDIP_STR2** are copies of the nodes **inDIP_STR1** and **inDIP_STR2** if the node **query** is in the state *True*, otherwise they are copies of the nodes **uDIP_STR1** and **uDIP_STR2**. This represents the understanding that, if the second contributor is the suspect (i.e., the node **p2isS?** is in the state *True*), then the genotype of the second contributor (modeled by **outDIP_STR1** and **outDIP_STR2**) should reflect the genotype of the suspect. Otherwise it should be equal to the genotype of an unknown person from the relevant population (represented by the nodes **uDIP_STR1** and **uDIP_STR2**). The CPT of the nodes **outDIP_STR1** and **outDIP_STR2** thus are completed as follows:

$$\begin{cases} P(\textbf{outDIP\_STR}i = j \mid \textbf{uDIP\_STR}i = j, \textbf{query} = False) = 1, \\ P(\textbf{outDIP\_STR}i = j \mid \textbf{inDIP\_STR}i = j, \textbf{query} = True) = 1, \end{cases} \quad \text{for } i = \{1,2\},\ j = \{La, Lb, Lx, Sa, Sb, Sx\}.$$

[12] Four different states are needed even if at most two alleles can be observed at a given marker. This is because – taking into account the two different DIP alleles – four different combinations can appear.

[13] The purpose of this is to have the nodes **DIP_STR1** and **DIP_STR2** as output nodes in the main class `Marker`.