Università degli Studi "Roma Tre"

Scuola dottorale in "Economia e metodi quantitativi"

Tesi di dottorato

OBJECT-ORIENTED BAYESIAN NETWORKS FOR ANALYSIS OF DNA MIXTURES

Candidata:
Sara Merigioli

Relatore:
Prof.ssa Julia Mortera

XX Ciclo

This thesis is dedicated to Pier Francesco Sammartino. Thanks for everything, thanks for having supported me all the way since the beginning. I am greatly in debt to you. THANKS.

Contents

1 Introduction 6

1.1 Objectives and main aims . . . 9

1.2 Layout . . . 12

2 Theoretical aspects on graph theory 14

2.1 Graph notions . . . 14

2.1.1 Basic aspects . . . 14

2.1.2 DAGs, chain graphs and moralization . . . 17

2.1.3 Chordal and decomposable graphs . . . 20

2.1.4 Junction trees . . . 23

2.2 Conditional independence . . . 28

2.3 Markov Properties . . . 31

2.3.1 Markov Properties for undirected graphs . . . 32

2.3.2 Markov Properties for DAGs . . . 37

2.3.3 Markov Properties for chain graphs . . . 44

3 Bayesian Networks for discrete variables 47

3.1 Bayesian Networks . . . 47

3.1.1 Definition of Bayesian Networks . . . 48

3.1.2 Inference in Bayesian networks . . . 49

3.1.3 Propagation algorithm applied to the Asia example . . . 57

4 Genetic Background 63

4.1 DNA Biology 63

4.1.1 Basic DNA principles 63

4.1.2 DNA structure 64

4.1.3 Chromosomes 66

4.1.4 Genes 67

4.1.5 DNA markers nomenclature 68

4.2 DNA amplification and STR markers 69

4.2.1 Polymerase chain reaction (PCR) process 69

4.2.2 Short tandem repeat analysis 70

4.3 Alleles drop-out and stutter 73

5 The experimental design procedure 74

5.1 Data for 2-person mixtures 77

5.2 Data for two traces 2-person mixtures 79

5.3 Data for 3-person mixtures 81

6 Introduction to DNA mixtures 88

6.1 Representation of a DNA mixture 90

6.2 Basic framework 92

6.3 OOBN for analysing DNA mixtures 95

7 Identification and separation for 2-person DNA mixtures 100

7.1 Identifying the genotype each of the possible contributors to the mixture 100

7.2 Analysis of DNA mixtures using peak area information 104

7.2.1 Basic model assumptions 105

7.2.2 Identifying the genotype each of the possible contributors to the mixture 110

7.2.3 Separation of genotypes 114

8 Analysis of two DNA mixed traces 121

8.1 Two mixed traces analysis using allele's repeat number information 122

8.1.1 Identification of the suspects' genotypes 122
8.2 Two mixed traces analysis adding peak ares information . . . 127
8.2.1 Identification of the suspects' genotypes 128
8.2.2 Separation of mixtures 133

9 Identification of DNA mixtures involving more than two contributors 138
9.1 Model assumptions 139
9.2 Advantages and disadvantages of each network and their efficiency 142
9.3 Bounding the total number of contributors 144
9.3.1 Theoretical aspects 144
9.3.2 Bounding the number of contributors for the 3-person mixture analysed in § 9.4 146
9.4 Forensic identification problems 148
9.5 Separation of genotypes 156

10 Conclusions and further investigations 161

Bibliography 165

Appendices 167

A Details of the object-oriented Bayesian network for 2-person mixtures 169
A.1 OOBN for 2-person mixtures using alleles' repeat number information only 169
A.2 OOBN for 2-person mixtures including peak area information 178

B Details of the object-oriented Bayesian network for two mixed traces 188
B.1 OOBN for two DNA mixed traces using alleles' repeat number information only 188
B.2 OOBN for two DNA traces including peak area information . 196

C Details of the object-oriented Bayesian network for 3-person mixtures 201

C.1 OOBN for 3-person DNA mixtures using alleles' repeat number information only 201

C.2 OOBN for 3-PERSON DNA mixtures including peak area information 209

Chapter 1

Introduction

The core of DNA testing is founded on Statistics. Genetic evidence is often used to solve problems concerning criminal cases as well as disputed paternity using statistical tools such as probabilistic expert systems (PESs). In the years, this topic has been being investigated and a number of tools have been developed and updated for both carrying out numerical computations and analysing cases involving traces of DNA. DNA evidence plays a large role in criminal cases as a tool for convicting or discharging. These cases often involve murders, rapes or robberies. DNA is an important investigative tool since, with the exception of homozygote twins (or triplets etc.), two people in the world cannot have the same DNA, so it can confirm or not the guilt of a suspect. A common scenario is that a DNA trace is found on the crime scene, afterwards a suspect is identified and his DNA profile is gathered and matched to the acquired sample in order to investigate the compatibility of the suspect's genotype to the trace and therefore to verify his guilt. Whenever a suspect is not recognized, the genotypes of the contributors can be predicted and, if a database of DNA profiles is available, matched to the profiles of the individuals in the database in order to identify possible suspects. These databases might be extremely wide (3 million profiles are preserved in the UK database from December 2005) and DNA samples from suspects are kept even in the case, after investigations, they are proved not to be guilty (Mortera and Dawid 2007). In more complex cases the trace found in the crime scene is made up of more than two profiles, *i.e.* at least three persons are involved in the crime. This is a typical example when multiple perpetrators took part in a rape or before a rape the victim had a consensual partner. Thus, biological material from the victim, the perpetrator/s and the consensual partner are contained in the trace and have to be processed properly in order to be distinguished.

Any type of organism can be identified through the examination of DNA sequences that are unique for that species. DNA (DeoxyriboNucleic Acid) is a nucleic acid carrying the genetic code which determines individual characteristics of each person. Genetic markers are particular portions of DNA, used to investigate the relationships between individuals. The most common for forensic purposes are the short tandem repeat (STR) markers. Every single DNA profile is composed of several STR markers. For each marker a genotype is organized in an unordered pair of alleles that are represented as positive integers. According to Mendelian segregation, each parent transmits to his or her child just one of the two alleles possessed. Thus, each genotype is made of one allele from the father and one from the mother, but there is an ambiguity who a certain allele comes from, as there is not a specific order in their displacement. When mother and father transmit the same allele value the individual is called homozygous, otherwise heterozygous. Thus, in a homozygous individual just one single allele value is observed. Whereas each single individual can possess at most two distinct alleles on any marker, whenever an observed crime scene trace presents more than two alleles at any marker, the trace must be clearly a mixture of DNA profiles from two or more individuals. Obviously, the greater the number of alleles observed in the mixture (and therefore the number of contributors to the mixture), the greater the complexity of the problem, because a greater number of combinations of the genotypes must be considered.

The results of a DNA analysis can be represented as an *electropherogram* (EPG) which reproduces the alleles in the mixture through peaks having a specific height and area around the allele. The *peak areas* are an extremely important quantities since they are approximately proportional to amount of DNA in the mixture and therefore provide important information on the composition of the mixture.

Complications arise as soon as the possibility of artifacts, such as allelic drop-out or stutter, are considered. Allelic drop-out are due to equipment failure when the low DNA level is insufficiently amplified to give a detectable signal. This is often due to reduced quantities of DNA, so that they are not detectable. In particular, they occur especially in presence of extremely unbalanced contributions to the mixture. For example, suppose to observe a 2-person mixture where the DNA proportions are $10:1$, *i.e.* 10 parts of DNA come from a contributor and 1 part comes from the other. Moreover, suppose to observe in the mixture the alleles (A, B, C) and that profiles of the contributors are (A, B) and (C, D). In this scenario, the allele D, present in the genotype of the second contributor, is not observed in the mixture since it is a drop-out allele. Other frequent artifacts are stutters. These are due to a slippage of the DNA during the replication process. They are spurious products with extremely small peaks and they contain one repeat unit less than the corresponding main allele peak.

The statistical tools that are used in this thesis are the object-oriented Bayesian networks (OOBNs), developed using the software package Hugin[1]. They have been developed and introduced by Koller and Pfeffer (1997); Laskey and Mahoney (1997). The first time Bayesian networks have been introduced to analyse DNA evidence was by Dawid *et al.* (2002).

OOBNs are a recent extension of the Bayesian networks (BNs); they are blocks of BNs combined in a hierarchical form, where Bayesian networks are Direct Acyclic Graphs (DAGs) used to build Bayesian models including a high number of variables. Each variable is described with a node. Nodes are connected by directed links that express probabilistic casual relationships between variables.

OOBNs implement numerical computations in order to evaluate the likelihood ratio in favour of guilt. After propagating the evidence, the posterior probabilities of the hypotheses on the individuals involved in the mixture

[1] See www.hugin.com

are computed in a target node. When prior probabilities are uniform, as we assume in this thesis, no prior information is added to the data, so the ratio of the posterior probabilities can be interpreted as the likelihood ratio in favour of guilt. In effect, forensic experts are often induced to formulate the reasonable assumption that the prior probabilities for each hypothesis H are equal leaving to adjudicators, judges or juries to formulate the prior assessments. When the likelihood ratio has a high enough value we can conclude there is enough evidence for the null hypothesis , *i.e.* “the suspect is guilty”. An example will clarify the method to build these likelihood ratios. Suppose to observe a 2-person DNA mixture having alleles' repeat number $\{A, B, C\}$ for a specific marker and suppose to observe the following profiles from a suspect and a victim: $s = \{A, B\}$, $v = \{B, C\}$. We are interested in testing the hypotheses $H_0 : v\&s$ versus $H_1 : v\&u$. Whereas for the hypothesis H_1 the profile of u can be either $\{A, A\}$, $\{A, B\}$, $\{B, A\}$, $\{A, C\}$, $\{C, A\}$, the likelihood ratio is expressed as:

$$\frac{\mathrm{pr}(H_0|\mathcal{E})}{\mathrm{pr}(H_1|\mathcal{E})} = \frac{1}{p_A^2 + p_A p_B + p_A p_C},$$

where p_i, for $i = A, B, C$, is the frequency of allele i in the population.

1.1 Objectives and main aims

We report a methodology, based on Probabilistic Expert Systems (PESs) for analysing and solving complex problems involving DNA mixtures using both allele repeat number and peak area information. A PES using information about the alleles present in the mixture was introduced by Mortera *et al.* (2003). Cowell *et al.* (2007b) showed how object-oriented Bayesian networks (OOBNs) can be used to analyse peak area information in 2-person mixtures.

Here the main aim is to extend their statistical model in order to analyse two traces (`T1` and `T2`) simultaneously. Both identification and separation of the DNA mixtures are analysed on different laboratory prepared mixtures on two independent traces. In particular, for the identification problem we discriminate between two different situations: when allele repeat number only is available, and when peak areas are also observed. Furthermore, we show that sometimes an investigation based on allele repeat number only can lead to erroneous inference, whereas the inclusion of the peak area information in the analysis gives the correct result.

In particular, we consider a robbery case where we suppose that some tools for breaking into an apartment have been handled by more than one individual. Thus, mixed traces of DNA samples are left at the scene of crime and two suspects are identified. We suppose to be interested in two particular traces. In identification problems, the main aim is to investigate whether the DNA genotypes of the suspects match those who contributed to the mixtures. Thus, the evidence could consist of the two mixed traces and DNA profiles extracted from two suspects, `s1` and `s2`. In this scenario, for each trace inference is made on the total number of contributors to the mixture. As a consequence, the posterior probabilities are evaluated for each hypothesis concerning the total number of contributors. Thus, if we suppose to obtain high posterior probability associated to the hypothesis that the total number of contributors is two in each trace, then we compare, for each trace, the hypotheses in Table 1.1.

However, in a courtroom two hypotheses are considered and the likelihood ratio is evaluated in favour of the hypothesis that both the suspects contributed to the mixture: H_0 :$s1\&s2$, versus the hypothesis that two unknown individuals `u` contributed to the mixture: H_1 :$2u$. Additionally, in a courtroom we could be interested in investigating whether each suspect contributed to both traces or one of them. Thus, we compute the posterior probabilities for the hypotheses in Table 1.2.

Hypotheses under test	
s1&s2	both suspects contributed to the trace1/trace2
s1&u	suspect1 and an unknown individual contributed to the trace1/trace2
s2&u	suspect2 and an unknown individual contributed to the trace1/trace2
2u	two unknown individuals contributed to the trace1/trace2

Table 1.1: Hypotheses under test.

The analysis is developed using the alleles' repeat number only and then

Hypotheses under test	
s1 in T1 or T2	first suspect in at least one trace
s1 in T1&T2	first suspect in both traces
s2 in T1 or T2	second suspect in at least one trace
s2 in T1&T2	second suspect in both traces

Table 1.2: Hypotheses under test.

adding peak area information, where peak area delivers additional information, since it is approximately proportional to the amount of DNA in the mixture. The main aim is to show that peak area information should be taken into account, since it allows to obtain more accurate probabilities.

Peak area information also allows to solve problems concerning separation of the mixture. In this scenario, the evidence could be given by the two traces only, whereas we assumed that profiles from any suspects are not available. Thus, the main aim is to predict the DNA profiles of the unknown contributors for each trace by separating the mixtures into its individual components. This allows to compare the single components with those available, for instance, in a database. Peak areas are modeled with a conditional-Gamma model, as in Cowell *et al.* (2006), but we show that also a conditional-Gaussian model is a good approximation.

Moreover, the statistical model of Mortera *et al.* (2003); Cowell *et al.* (2007b) is extended in order to analyse mixtures involving three contributors. In particular, we consider a rape case where we suppose that a sample contains biological material from a victim and two perpetrators. For this statistical model we solve only identification problems and the mixture is analysed using as evidence, before allele repeat number only, and then adding also peak areas information. Here, also the conditional-Gamma model of Cowell *et al.* (2006) is extended for a case involving three contributors to the mixture and is applied for the rape case. Although the network used for 3-person mixtures answers much more complex problems, it is computationally more elaborated than the network for 2-person mixed traces. Unfortunately, this complexity represents a strong limitation of computer for the analysis of 3-person mixtures and this is the reason why we cannot consider a high number of markers in the identification analysis and we cannot predict the genotypes of the unknown contributors to the mixture.

Concluding, the main aim of this thesis is to show the efficiency of both extended statistical models and to prove that peak weights need to be taken into account since they improve the performance by increasing the likelihoods and the posterior probabilities.

1.2 Layout

This thesis is organized into eleven main sections. Chapters 2 and 3 provide some theoretical aspects on the graph theory and Bayesian networks, whilst a genetic background is given in chapter 4. Data used for the analyses in the rest of the chapters are shown in chapter 5. Here we give details on the laboratory prepared mixtures with the corresponding DNA proportions. Moreover for each marker we further provide the allele repeat numbers, peak areas, relative peak weights, genotypes of suspects and victims and the population gene frequencies. In chapter 6 we introduce and explain DNA mixtures, together with the issues under investigation in this thesis and the tools used to overcome them, then in chapter 7 a murder case with a 2-person DNA sample is examined. In particular, in this chapter we discuss in detail the statistical model and we solve identification and separation problems using as evidence allele repeat numbers and peak area information as evidence. Details on the network used for this chapter are given in Appendix A. In chapter 8 the statistical model of Mortera *et al.* (2003); Cowell *et al.* (2007b) applied in the previous chapter is extended in a way that allows to analyse two mixed traces simultaneously. We assume a robbery case where some tools used to beak into an apartment are left on the crime scene and two independent DNA samples are analysed. Details on the network are given in Appendix B. On the contrary, in chapter 9 the statistical model of Mortera *et al.* (2003); Cowell *et al.* (2007b) is extended in order to solve identification problems for 3-person mixtures. Here, we consider a rape case where a sample containing biological material from the victim and two perpetrators is examined and compared to the genotypes of two suspects. Details on the network are finally provided in Appendix C.

Chapter 2

Theoretical aspects on graph theory

2.1 Graph notions

2.1.1 Basic aspects

In this section we introduce the basic theory about graphs. Graph theory is an abstract mathematical subject and is an extremely useful tool when applied to probabilistic expert systems for its ability to present a representation of expert knowledge about the subject.

A graph $\mathcal{G}$ is constituted by the pair (V, E), where V is a set of vertices, called *nodes*, and E is a subset $V \times V$ of ordered pairs of vertices called *edges* or *links*. Nodes are represented by circles, directed edges by arrows, and undirected edges by lines. Figure 2.1 shows an example of a graph having four vertices, with two directed edges from node A to B and from node A to C, and two undirected edges between vertices (B,D) and (C,D).

A graph is called *directed* if all its edges are represented by arrows, whilst it is termed *undirected* if all its edges are undirected. The undirected version of a graph $\mathcal{G}^{\sim}$ has all the arrows replaced by undirected edges and the undirected version $\mathcal{G}^{\sim}$ is an undirected graph.

In a graph, if both $(a,b) \in E$ and $(b,a) \in E$, the edges between the two vertices a and b are undirected, and a and b are *joined*. In this case a and b are said to be *neighbours*, therefore a is neighbour of b and b is neighbour of

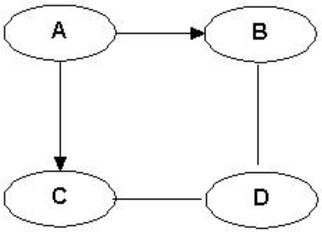


Figure 2.1: Example of graph

a. Two joined nodes are denoted by $a \sim b$ and the set of neighbours of a is denoted by ne(a). For example, in Figure 2.1 the nodes B and D are joined. Conversely, if both (a,b) $\notin E$ and (b,a) $\notin E$, then a and b are not joined, this is denoted by $a \nsim b$. In this case there is neither a line nor an arrow between a and b and they are said to be *non-neighbours*. Similarly, if (a,b) $\in E$ but (b,a) $\notin E$, then it can be written $a \rightarrow b$, and if (a,b) $\notin E$, then $a \nrightarrow b$.

The relations in a directed graph are denoted using the terms commonly referred to family relations. Nodes, with arrows starting from them, are called *parents*, whilst nodes, with arrows pointing into them, are called *children*. For example, in Figure 2.1, A is a parent of B and B is a child of A. In addition, we refer to (i) the set of vertices parents of b as pa(b), (ii) the collection of children of a node a as ch(a), and (iii) the family of b as the collection of b and its parents as fa(b) = $b \cup$ pa(b). In a directed graph, nodes that have no parents are called *founder nodes* whilst those that have no children are called *terminal nodes*.

Consider a subset W of V, $W \subseteq V$, we have:

$$\begin{aligned} \mathrm{pa}(W) &= \textstyle\bigcup_{w\in W} \mathrm{pa}(w)\backslash W \\ \mathrm{ne}(W) &= \textstyle\bigcup_{w\in W} \mathrm{ne}(w)\backslash W \\ \mathrm{ch}(W) &= \textstyle\bigcup_{w\in W} \mathrm{ch}(w)\backslash W. \end{aligned}$$

Thus, pa(W), ne(W) and ch(W) indicate, respectively, the collection of parents, neighbours and children of W excluding any vertex in W. For example, in Figure 2.1, the set of parents of (B,C) is represented by the node A, *i.e.* pa({B,C}) = {A}.

The collection of parents and neighbours of a node a is called *boundary* bd(a), whilst the boundary of a subset $W \subset V$ is the set of parents or neighbours of the elements in W excluding any element in W, *i.e.* bd(W) = pa(W) ∪ ne(W). For example, in Figure 2.1 bd(B)=$\{A, D\}$. The *closure* of W, cl(W), is defined as the set formed by W and its boundary, *i.e.* cl(W) = W∪bd(W). In Figure 2.1, cl(B)={A,B,D}.

A *path* of length n from a_1 to a_n, $a_1 \mapsto a_n$, is a sequence of distinct vertices a_1, a_2,... a_n belonging to E such that the direction of arrows is always followed and the path never crosses itself. In this case it is said that a_1 *leads to* a_n. When, in the path, two or more consecutive vertices are connected by an arrow, the path is directed, *i.e.* there exists at least one $i \in$ {1,2,...,n} such that $a_i \rightarrow a_{i+1}$. If there is a path in both directions from a to b and from b to a, *i.e.* $a \mapsto b$ and $b \mapsto a$, a and b are *connected* and this is denoted by $a \rightleftharpoons b$. Connectivity forms equivalence classes $[a]$, called *strong components* of $\mathcal{G}$, such that $b \in [a] \Leftrightarrow a \rightleftharpoons b$. In Figure 2.1 the nodes (B,D) and (C,D) are strong components. Considering a graph $\mathcal{G}$ and its undirected version $\mathcal{G}^{\sim}$, if there is a path between every pair of vertices in $\mathcal{G}^{\sim}$, then $\mathcal{G}$ is connected. The strong components of $\mathcal{G}^{\sim}$ are *connected components.*

A *trail* of length n from a_1 to a_n is a sequence of distinct vertices a_1, a_2,... a_n belonging to E such that, for all i=1,2,...,n, $a_i \rightarrow a_{i+1}$, or $a_{i+1} \rightarrow a_i$, or $a_i \sim a_{i+1}$. In contrast to a path, a trail can pass against the direction of the arrows.

A *subgraph* of $\mathcal{G}$ is the graph $\mathcal{G}_{\mathcal{W}}$ = (W,E_W), where $W \subseteq V$ and $E_W \subseteq E \cap (W \times W)$. $\mathcal{G}_{\mathcal{W}}$ is a subset of vertices of $\mathcal{G}$ that may contain the same vertices in $\mathcal{G}$ but fewer edges. If E_W =$E \cap (W \times W)$, $\mathcal{G}_{\mathcal{W}}$ is called subgraph *induced* by W. Examples are shown in Figures 2.2 (b) and (c) that are subgraphs of (a).

A graph is said to be *complete* if all vertices are joined by an arrow or a line. A complete subgraph which is maximal with respect to $\subseteq$ is called a *clique*. Figure 2.3 (a) shows an example of complete graph, whilst Figure 2.3 (b) shows an undirected graph with two cliques represented by the nodes

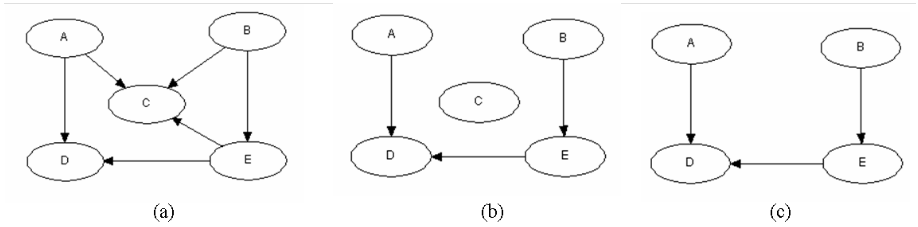


Figure 2.2: (a) a graph; (b) subgraph of (a); (c) induced subgraph of (a).

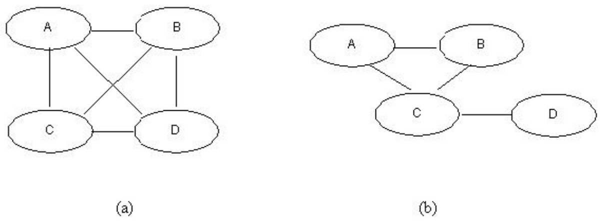


Figure 2.3: (a) a complete graph; (b) an undirected graph formed by the cliques (A,B,C) and (C,D).

(A,B,C) and (C,D).

2.1.2 DAGs, chain graphs and moralization

A particular kind of directed graph is a DAG (Directed Acyclic Graph). An important requirement for a DAG is that E has to comprise distinct vertices so that loops, or cycles, are not allowed, *i.e.* a directed path that starts and ends at the same vertex is not permitted. A cycle is such that following the direction of the arrows it is possible to return to the node of departure (see Figure 2.4)

A DAG can always be *well-ordered* providing a linear ordering or numbering such that, if two nodes are connected, it is possible to pass from a node with lower number or order (a node where the edge starts from) to a node with higher number or order (a node where the arrow points to). For example, Figure 2.5 shows a DAG with a unique well-ordering given by the sequence of nodes (A,B,C,D). The well-ordering may not be unique. In a well-ordered DAG the *predecessors* of a, $\mathrm{pr}(a)$, are the vertices with lower

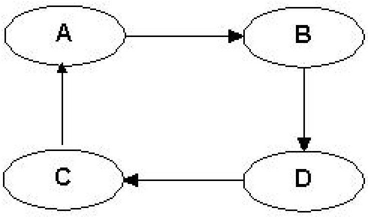


Figure 2.4: Example of cyclic graph.

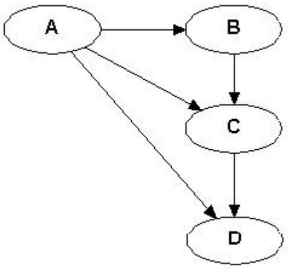


Figure 2.5: Example of graph with order A,B,C,D.

order number than a.

The concept of *chain graph* $\mathcal{K}$ is now introduced. This is a graph where the set of vertices V can be partitioned into numbered subsets $W(t) \subseteq V$ forming a *dependence chain* $V = W(1) \cup \cdots \cup W(T)$ such that the vertices in the same subset are joined by undirected edges whilst different subsets are connected by arrows. Chain graphs have no directed cycles and its connected components are termed *chain components*. The chain components can be easily found removing all the arrows in the chain graph. For example, both undirected graphs and DAGs are special cases of chain graphs and in the directed acyclic graph the chain components are given by single vertices. For example, the graph in Figure 2.6 is a chain graph having chain components $\{X_1, X_2, X_3\}$, $\{X_4, X_5, X_6, X_7\}$, $\{X_8\}$.

In a chain graph, the set of vertices a_1 such that $a_1 \mapsto a_n$ but not viceversa $a_n \not\mapsto a_1$, are termed *ancestors* of a_n, $\mathrm{an}(a_n)$, whilst, the set of vertices a_n are termed *descendants* of a_1, $\mathrm{de}(a_1)$. The *non-descendants*,

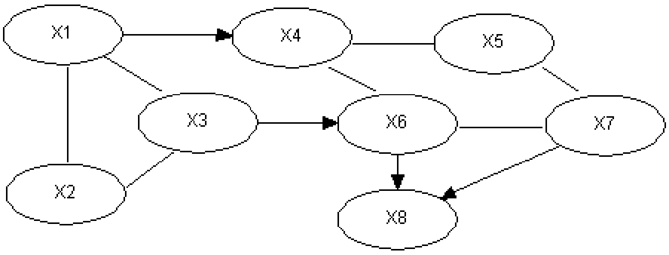


Figure 2.6: A chain graph having chain components $\{X_1, X_2, X_3\}$, $\{X_4, X_5, X_6, X_7\}$, $\{X_8\}$.

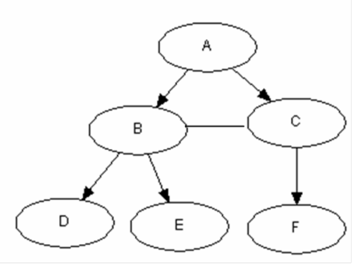


Figure 2.7: A chain graph: an(F)=(A, B, C); de(B)=(D, E); An(C)=(A, B)

nd(a_1), are the set of vertices in V excluding the descendants of a_1 and a_1 itself, *i.e.* nd(a_1)=V \(de(a_1) $\cup$ a_1). Consider $a_1 \in A$, where A is a subset of V, if bd(a_1) $\subseteq A$, then A is an *ancestral set*, and the smallest ancestral set containing A is denoted by An(A). For example, in Figure 2.7 the set of ancestors of F is (A,B,C), thus An(F)={A,B,C}; the set of descendants of B are (D,E), thus de(B)={D,E}; the set of ancestors of C is (A,B), thus An(C)={A,B}.

The *moral graph* of a chain graph $\mathcal{K}$ is now considered. This is defined as the undirected graph $\mathcal{K}^m$ obtained through the following two steps: a) we add undirected edges in $\mathcal{K}$ between nodes that have a common child and that are not already joined (this is called "marrying" two nodes); b) dropping all directions of arrows and obtaining the undirected version of the resulting graph. If $\mathcal{K}$ is a DAG the process is the same and all the pairs of parents are married and the arrows are replaced by undirected edges. An example of *moralization* process is reported in Figure 2.8 where a DAG is displayed with its moral graph.

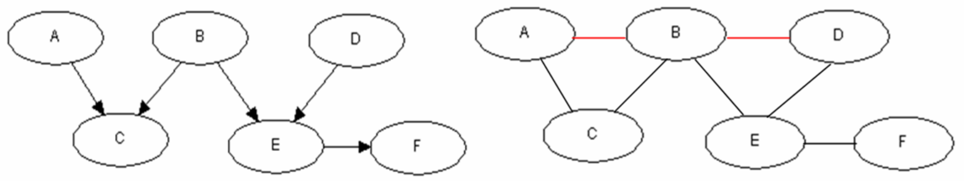


Figure 2.8: Moral graph of a DAG obtained by marrying the nodes (A,B) and (B,D) and replacing the arrows with undirected edges.

2.1.3 Chordal and decomposable graphs

Let $\mathcal{G}$ be an undirected graph, it is called *chordal* or *triangulated* if every one of its cycles of length $\geq$4 contains a chord. A chord of an n-cycle in $\mathcal{G}$ is an arc between two non-consecutive vertices in that cycle. An example is shown in Figure 2.9 where the edge $B \sim C$ is a chord. If $\mathcal{G}$ is chordal and $A \subset V$, then $\mathcal{G}_{\mathcal{A}}$ is also chordal. A graph $\mathcal{G} = (V, E)$ can be always made chordal by adding extra edges F to form $\mathcal{G}' = (V, E')$, where $E' = E \cup F$. The edges in F are referred to as *fill-in* edges. If $\mathcal{G}'$ is chordal, then it is called a *triangulation* of $\mathcal{G}$. An important type of graph is a *decomposable* graph.

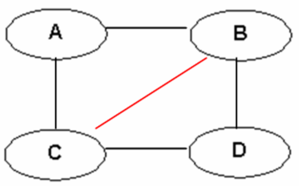


Figure 2.9: A chordal graph. The edge $B \sim C$ is a chord.

In order to define a decomposable graph we need to introduce the concept of a *separator*: let S be a subset of V, $S \subseteq V$, S is an $(a{,}b)$-*separator* if all trails from a to b intersect S. If S is an $(a{,}b)$-separator for every $a \in A$ and $b \in B$, then S *separates* A from B, where A, B and S are disjoint subsets of V. An $(a{,}b)$-separator S is *minimal* if there are no subsets of S that are $(a{,}b)$-separators. For example, in Figure 2.10 the set of nodes $(C{,}D)$ is $(B{,}E)$-separator and it is also minimal.

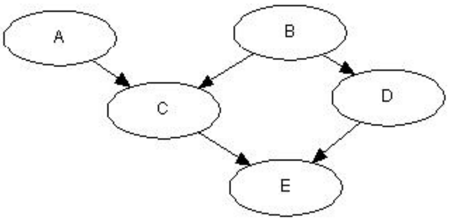


Figure 2.10: Graph with the set of nodes (C,D) as minimal (B,E)-separators.

Let $\mathcal{G}$ be an undirected graph, a triplet (A,B,S) of disjoint subsets of V is a *decomposition* of $\mathcal{G}$, if V=$A \cup B \cup C$ and the following two conditions hold: (i) S separates A from B; (ii) S is a complete subset of V. An undirected graph is *decomposable* if either (i) it is complete or (ii) it contains a proper decomposition (A,B,S) that defines the decomposable subgraphs $\mathcal{G}_{\mathcal{A}\cup\mathcal{S}}$ and $\mathcal{G}_{\mathcal{B}\cup\mathcal{S}}$. Any graph can be decomposed into its connected components. The smallest non decomposable graph is a 4-cycle. A connection between decomposability and chordality is shown through the following theorem (Lauritzen 1996). The proof is taken from Cowell *et al.* (1999).

Theorem 2.1 Let $\mathcal{G}$ be an undirected graph, it holds equivalently that:
1. $\mathcal{G}$ is decomposable;
2. $\mathcal{G}$ is chordal;
3. every minimal (a,b)-separator is complete.

Proof. The three conditions are proved by induction on the number of vertices $|V|$ of $\mathcal{G}$. For a graph with no more than three vertices they hold automatically. Thus, assuming these results for all graphs with $|V| \leq n$, it has been proved that they hold also for all graphs $\mathcal{G}$ with $|V| = n + 1$.
First we show that $1 \Rightarrow 2$. Let $\mathcal{G}$ be a decomposable graph. For definition of decomposable graph, $\mathcal{G}$ is either complete, and thus it is obviously chordal, or has a proper decomposition (A, B, S) such that both subgraphs $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ are decomposable with fewer vertices. These subgraphs are chordal by inductive hypothesis. In only one case we have a chordless cycle: when the cycle intersects both A and B. But, if the cycle intersects both A and B,

then it intersects S at least twice because S separates A from B. Then, the cycle is chordal since S is complete.

We prove now that 2 $\Rightarrow$ 3. Let $\mathcal{G}$ be a chordal graph and S be a minimal (a,b)-separator in $\mathcal{G}$. If S contains only one node, it is complete. If S has at least two vertices, for example g_1 and g_2, since it is a minimal separator, there will be paths from a to b via g_1 and back via g_2, (a,..,g_1,...,b,...,g_2,...a). These paths produce a cycle which can have repeated points. The cycle can be shorten through the repeated points or adding a chord (other than the one connecting the vertices g_1 and g_2) and leaving at least one vertex in the connected components $[a]_{V\setminus S}$ and $[b]_{V\setminus S}$ of $\mathcal{G}_{V\setminus S}$ containing a and b respectively. Therefore, cycles of length at least 4 are produced and these must have a chord obtaining $g_1 \sim g_2$. Repeating the process for every pair of vertices in S we obtain that S is complete.

Finally we prove that 3 $\Rightarrow$ 1. Assume the third condition, *i.e.* every minimal (a, b)-separator is complete. If $\mathcal{G}$ is complete then is automatically decomposable, otherwise it has at least two non-adjacent vertices not joined (a and b say). Assume that the result holds for every proper subgraph of $\mathcal{G}$. Let S be a minimal separator of a and b, and partition the vertex set into $[a]_{V\setminus S}$, $[b]_{V\setminus S}$, S and C, where C includes all the remaining vertices. Now, let $A = [a]_{V\setminus S} \cup C$ and $B = [b]_{V\setminus S}$, then the triplet (A, B, S) forms a decomposition of $\mathcal{G}$, since S is complete. Actually, in order to prove the theorem, both the subgraphs $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ must be decomposable. Thus, if $\tilde{S}$ is a minimal $(\tilde{a}, \tilde{b})$-separator in $\mathcal{G}_{A\cup S}$, then it is also a minimal separator in $\mathcal{G}$, and therefore complete by assumption. As a consequence, $\mathcal{G}_{A\cup S}$ is decomposable by inductive hypothesis. Similarly it has been proved that $\mathcal{G}_{B\cup S}$ is decomposable. Now, since $\mathcal{G}$ has been decomposed into decomposable subgraphs, we can conclude that $\mathcal{G}$ is decomposable.

□

A DAG is defined *perfect* if its parent nodes form a complete set. For an undirected graph a numbering of its vertices, (v_1, v_2, ..., v_n), is said perfect if the neighbours having lower numbers, *i.e.* $\mathrm{ne}(v_i)\cap\{v_1, v_2, ..., v_{i-1}\}$, induce a complete subgraph.

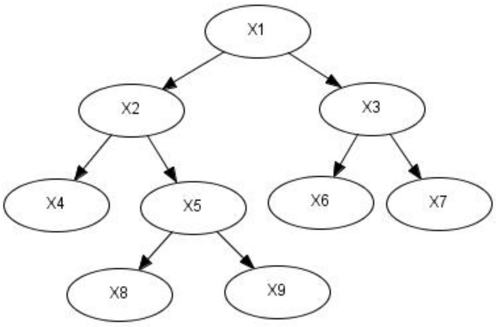


Figure 2.11: Example of a tree where the node X1 is a root node, whilst the nodes X4, X6, X7, X8 and X9 are leaf nodes

The undirected version $\mathcal{G}^{\sim}$ of a well-ordered perfect directed graph $\mathcal{G}$ is a chordal graph where the ordering (v_1, v_2, ..., v_n) forms a perfect numbering. This can be proved by induction since for all i the triplet (W_i, $V_i{-}1$, S_i) forms a decomposition $\mathcal{G}_{\tilde{V_i}}$ (where $V_i = (v_1, v_2, ..., v_i)$, $W_i = \text{cl}^{\sim}(v_i) \cap V_i$, $S_i = W_i \cap V_{i-1}$, and cl$^{\sim}$ indicates closure relative to the undirected graph $\mathcal{G}^{\sim}$). A perfect directed graph can be construct by directing the edges from lower to higher numbered vertices of an undirected graph $\mathcal{G}$ having a perfect numbering of its vertices. Additionally, Lauritzen (1996) proved that an undirected graph is chordal if and only if it admits a perfect numbering.

2.1.4 Junction trees

Another type of graph is a *tree* $\mathcal{T}$. A tree is a connected graph $\mathcal{G}$ (*i.e.* a graph where there is a path between every pair of vertices) and its undirected version $\mathcal{G}^{\sim}$ has no cycles. Thus, in a tree any two vertices are connected by exactly one trail. An example of tree is the graph in Figure 2.11. In a tree a *root* node is a node at the top level, and it has no parents, whilst, a *leaf* node is a node at the bottom level, and it has no children. A tree has a *diameter* represented by the length of longest trail between two leaf nodes.

Definition 2.2 Let $\mathcal{T}$ be a tree formed by a collection of cliques $\mathcal{C}$ as its node set, $\mathcal{T}$ is a *junction tree* (or *join tree*) if, for any pair $C_v = (X_1, X_2, ..., X_j)$ and $C_w = (X_{j-i}, X_{j-i+1}, ..., X_n)$ in $\mathcal{C}$ for all $i < j < n$, the intersection $C_v \cap C_w = (X_{j-i}, X_{j-i+1}, ..., X_j)$ is contained in every node on the unique path in $\mathcal{T}$ between C_v and C_w. This intersection $C_v \cap C_w$ corresponds to the set of nodes that separates C_v from C_w.

□

Similarly, for any vertex v in $\mathcal{G}$, the set of subsets in $\mathcal{C}$ containing v induces a connected *subtree* $\mathcal{T}'$ of $\mathcal{T}$. Let $\mathcal{G}$ be an undirected graph having $\mathcal{C}$ as the family of its cliques. Then, $\mathcal{T}$ is a junction tree for $\mathcal{G}$, if $\mathcal{T}$ is a junction tree containing $\mathcal{C}$ as its node set.

Theorem 2.3 A junction tree $\mathcal{T}$ of cliques for a graph $\mathcal{G}$ exists if and only if $\mathcal{G}$ is decomposable.

Proof. If $\mathcal{G}$ contains at most two cliques the result clearly holds. Thus, the theorem is proved proceeding by induction on the number k of cliques.

Let $\mathcal{T}$ be a junction tree of cliques for $\mathcal{G}$ having $k+1$ cliques and let C_1 and C_2 be two adjacent cliques in $\mathcal{T}$. If the link $C_1 \sim C_2$ is cut, then $\mathcal{T}$ is separated into two subtrees $\mathcal{T}_1$ and $\mathcal{T}_2$. Now, the union of the nodes in $\mathcal{T}_i$ is denoted by V_i, for $i = 1, 2$, and let $\mathcal{G}_i = \mathcal{G}_{V_i}$. The nodes in $\mathcal{T}_i$ are the cliques of $\mathcal{G}_i$, and $\mathcal{T}_i$ is a junction tree for $\mathcal{G}_i$. Both the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are decomposable by the inductive hypothesis. Now, proving that $S := V_1 \cap V_2$ is complete and separates V_1 from V_2, then the theorem holds. If we take $v \in V_1 \cap V_2$, then there exists in $\mathcal{G}_i$ a clique C'_i for $i = 1, 2$ containing v, *i.e.* $v \in C'_i$. Clearly the path in $\mathcal{T}$ joining C'_1 and C'_2 passes through both C_1 and C_2. As a consequence, $v \in C_1 \cap C_2$ and so we must have $V_1 \cap V_2 \subseteq C_1 \cap C_2$. Whereas $C_1 \cap C_2 \subseteq V_1 \cap V_2$, then $S = C_1 \cap C_2$ and is complete.

Consider now $u \in V_1 \backslash S$ and $v \in V_2 \backslash S$. Furthermore, suppose that there exists a path $u, w_1, w_2, ..., w_k, v$ where $w_i \notin S$. Then, a clique C including the complete set $\{u, w1\}$ also exists. It is clear that $C \subseteq V_1$, so $w_1 \in V_1$, whence $w_1 \in V_1 \backslash S$. Repeating the argument also the other elements in the path, $w_2 \in V_1 \backslash S, ..., v \in V_1 \backslash S$, can be deduced. Since this is a contradiction, it is concluded that S separates V_1 from V_2 and that (V_1, V_2, S) is a decomposition of $\mathcal{G}$. Thus, $\mathcal{G}$ has been decomposed into a number of subgraphs containing

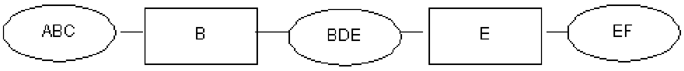


Figure 2.12: Junction tree of the graph in Figure 2.8.

junction trees and thus are decomposable by the inductive hypothesis.

On the contrary, suppose that $\mathcal{G}$ is decomposable and let (W_1, W_2, S) be its decomposition into proper decomposable subgraphs $\mathcal{G}_{V_1}$, $\mathcal{G}_{V_2}$, for each $V_i = W_i \cup S$. Then, either V_1, or V_2, or both has the form $\bigcup_{C \in \mathcal{C}_1} C$ with $\mathcal{C}_1 \subset \mathcal{C}$. If we suppose $V_1 = \bigcup_{C \in \mathcal{C}_1} C$, then V_2 is redefined as $\bigcup_{C \in \mathcal{C}_2} C$ with $\mathcal{C}_2 = \mathcal{C} \backslash \mathcal{C}_1$ and there is still a decomposition. Now, let $C_i \in \mathcal{C}_i$ with $S \subseteq C_i$. Then, there exists a junction tree $\mathcal{T}_i$ for $\mathcal{G}_i$ by hypothesis, (where, as said previously, $\mathcal{G}_i = \mathcal{G}_{V_i}$) and we form $\mathcal{T}$ by linking C_1 in $\mathcal{T}_1$ to C_2 in $\mathcal{T}_2$.

It is considered now $v \in V$, if $v \notin V_1$, then v is in the cliques contained in $\mathcal{C}_2$. Such cliques are also connected in $\mathcal{T}_2$, hence in $\mathcal{T}$. It holds similarly if $v \notin V_2$. Otherwise $v \in S$. Thus, in general, the cliques in C_i containing v are connected in $\mathcal{T}_i$, and include C_i. Now, the theorem is proved whereas C_1 and C_2 are connected in $\mathcal{T}$.

□

By this theorem it follows that the intersection $S = C_1 \cap C_2$ between two neighbouring nodes C_1 and C_2 in a junction tree of cliques $\mathcal{C}$ is a minimal separator which separates the decomposable graph $\mathcal{G}$. Additionally, S is said *separator* associated with the edge between C_1 and C_2 of the junction tree and this term separator is used even if the nodes of the junction tree are not all cliques. Sometimes distinct edges may have identical separators and the set of all separators is denoted by $\mathcal{S}$. It can be shown that, if $\mathcal{G}$ admits more than one junction tree of cliques, then $\mathcal{S}$ is the same for all of them.

As shown in Figure 2.12, in a junction tree, separators are drawn as rectangles, whilst nodes formed by cliques are displayed as ovals.

A clique $C^* \in \mathcal{C}$ is called *extremal* if the triplet $(C^* \backslash V_2,\ V_2 \backslash C^*,\ C^* \cap V_2)$ is a decomposition of $\mathcal{G}$, where $V_2 = \bigcup_{C \in \mathcal{C} \backslash \{C^*\}} C$.

Corollary 2.4 If a chordal graph $\mathcal{G}$ has at least two cliques, then it has at least two extremal cliques.

Proof. The proof is due directly to the fact that any junction tree of $\mathcal{G}$ has at least two leaves.

□

A property characteristics of junction trees is the *running intersection property.* The running intersection property is such that, let $(C_1, C_2, ..., C_k)$ be a sequence of cliques of a junction tree if, for all $1 < j \leq k$, there exists an index i < j such that $C_j \cap (C_1 \cup ... \cup C_{j-1}) \subseteq C_i$, then such a sequence $(C_1, C_2, ..., C_k)$ satisfies the running intersection property. In other words, the intersection between the nodes of a clique and the nodes of all the previous cliques are contained in one of the previous cliques and this intersection is represented by the separating nodes.

Let $\mathcal{T}$ be a junction tree for a decomposable graph $\mathcal{G}$, by well-ordering the junction tree, also the cliques of the decomposable graph can be ordered to have the running intersection property.

Algorithm 2.5 - **Junction tree construction**. Let $\mathcal{G}$ be a chordal graph, and let $(C_1, ..., C_p)$ be a sequence of cliques of $\mathcal{G}$ ordered to satisfy the running intersection property. Then

(i) each clique C_i is associated to a node of the tree;

(ii) for $i = 2, ..., p$, an edge between C_i and C_j is added, where j is any one value in $\{1, ..., i-1\}$ such that $C_i \cap (C_1 \cup ... \cup C_{i-1}) \subseteq C_j$.

□

A chain graph $\mathcal{K}$ is now considered. If $\mathcal{K}$ is a probabilistic network, we shall see that, in order to make inference, the first stage is to form the moral graph $\mathcal{K}^m$. The moral graph is an undirected graph but may not be chordal. However, we can make it so and this process allows finding all of the cliques in $\mathcal{G}$. In general, given any ordering of the nodes of an undirected graph $\mathcal{G}$, for example $(v_1, ..., v_k)$, all the cliques C_i in $\mathcal{G}$ are identified by a successive vertices elimination process. Each node is examined in turn in reverse order, *i.e.* beginning by the last v_k. The node v_k is eliminated if all its neighbours are already joined. Otherwise, an extra edge is *filled-in* joining those pairs of neighbours that appear earlier in the ordering and are not already joined. Then, the eliminated vertex v_k and its neighbours form a clique. This process is repeated for all vertices. When all the vertices are eliminated, all the cliques are identified. The resulting graph, which is the undirected graph $\mathcal{G}$ including the extra edges F, is a triangulated graph $\mathcal{G}' = (V, E')$, where $E' = E \cup F$. The given ordering $(v_1, ..., v_k)$ is a perfect numbering for the triangulation $\mathcal{G}'$ of $\mathcal{G}$.

Example 2.6 Consider the undirected graph in Figure 2.13 (a) with ordering (A, B, C, D, E). We examine each node in turn in reverse order. Thus, we start from the last node E. It can be directly eliminated since its neighbours, A and D, are already joined and no fill-in edges are therefore required. Then, E and its neighbours form the clique (A, D, E). Consider now the remaining graph given by the nodes (A, B, C, D). The next node to examine is the node D. Since its neighbours A and C are not already joined we need to add an extra edge between them. Thus, also the vertex D is eliminated and the cliques (A, C, D) and (A, B, C) are identified. The resulting graph, given by the original undirected graph including the extra edges, is shown in Figure 2.13 (b) and is a triangulated graph. The three cliques and its separators will be associated with the nodes of the junction tree as shown in Figure 2.13 (c).

□

It is worth noting that, in the example 2.6, if a different order were given to the graph, for example (D, C, B, A, E), after eliminating the node E, the algorithm would examine the node A rather than D and add an extra edge between its neighbours B and D. In this alternative case, the resulting junction tree would be different. Thus, for any undirected graph there is a number of possible junction trees that can be obtained according to the

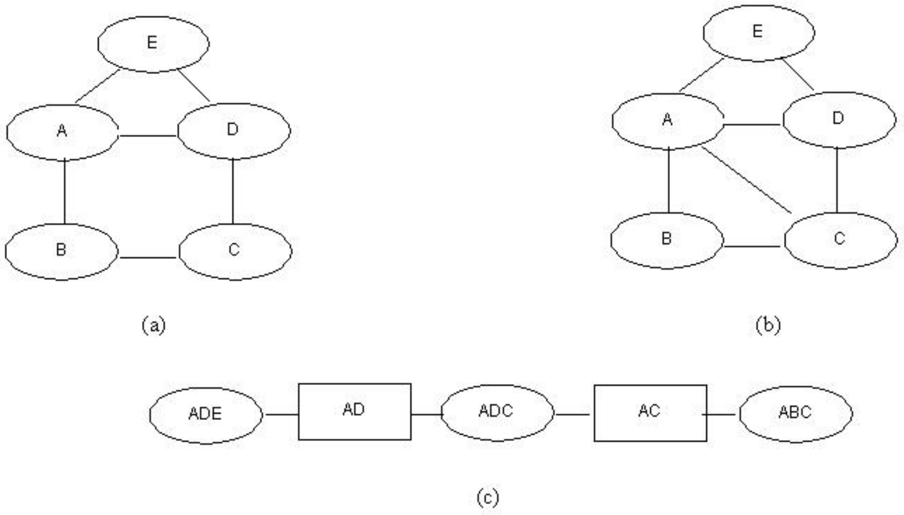


Figure 2.13: Example 2.6.

starting elimination order.

Tarjan and Yannakakis (1984) developed an algorithm to test the triangulatedness of an undirected graph. This algorithm is called *maximum cardinality search* and runs in $O(n + e)$ time, where n is the number of nodes, whilst e is the number of edges. The algorithm works as follows: (i) number 1 is given to an arbitrary node; (ii) the next node to number is the one consecutive. If there are more than one consecutive node, we choose the one with maximum number of previously numbered neighbours. If the ordering so obtained is perfect, the graph is triangulated. Even though the maximum cardinality search algorithm demonstrated efficiency in testing the chordality of a graph, it requires more fill-in edges than are necessary, producing a number of cliques higher than the minimum. This reduces the efficiency of the algorithm for probabilistic computations.

2.2 Conditional independence

In this section the notion of *conditional independence* of random variables is introduced allowing to justify local computations developed in inference processes with Bayesian Networks treated in this chapter.

Definition 2.7 Let X,Y and Z random variables with joint distribution P, then X is *conditionally independent* of Y *given* Z, denoted by $X \perp\!\!\!\perp Y \mid Z$, if, for any set A of possible values for X, the conditional distribution $P(X{\in}A \mid Y, Z)$ does not depend on Y.

□

Let X, Y and Z to be discrete random variables, if $X \perp\!\!\!\perp Y \mid Z$ we can write

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z), \quad (2.1)$$

for all z such that $P(Z = z) > 0$. On the contrary, if X, Y and Z are continuous random variable with joint density f the independence condition implies

$$f_{XY|Z}(x, y \mid z) = f_{X|Z}(x \mid z)f_{Y|Z}(y \mid z), \quad (2.2)$$

for all z such that $f_z(z) > 0$. In particular, if $X \perp\!\!\!\perp Y$, then we can write $P(X|Y = y) = P(X = x)$, *i.e.* the conditional distribution of X given $Y = y$ is equal the marginal distribution of X and this expression holds for any value y of Y. As a consequence, X and Y are said to be (marginally) independent.

Let $t(X)$ denote a generic function defined on X, the relation of conditional independence, $X \perp\!\!\!\perp Y \mid Z$, respects the following four properties:

(C1) if $X \perp\!\!\!\perp Y \mid Z$, then $Y \perp\!\!\!\perp X \mid Z$;

(C2) if $X \perp\!\!\!\perp Y \mid Z$ and $U = t(X)$, then $U \perp\!\!\!\perp Y \mid Z$;

(C3) if $X \perp\!\!\!\perp Y \mid Z$ and $U = t(X)$, then $X \perp\!\!\!\perp Y \mid (Z, U)$;

(C4) $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (W, Y) \mid Z$.

For the sake of simplicity, suppose that the three variables are discrete with density p respect to a product measure, so that $p(x, y|z)$ indicates $P(X = x,$ $Y = y|Z = z)$. Then the ternary relation $X \perp\!\!\!\perp Y \mid Z$ holds if and only if also the following below statements are true:

(S1) $p(x \mid y, z) \equiv p(x \mid z)$, if $p(y, z) > 0$;

(S2) $p(x \mid y, z)$ can be written as $l(x, z)$, if $p(y, z) > 0$;

(S3) $p(x, y \mid z) \equiv p(x \mid z)p(y \mid z)$, if $p(z) > 0$;

(S4) $p(x, y \mid z)$ can be written as $l(x, z)k(y, z)$, if $p(z) > 0$;

(S5) $p(x, y, z) \equiv p(x \mid z)p(y \mid z)p(z)$;

(S6) $p(x, y, z) \equiv p(x, z)p(y, z) \diagup p(z)$, if $p(z) > 0$;

(S7) $p(x, y, z)$ can be written as $l(x, z)k(y, z)$.

In these statements l and k are two generic functions respectively of (x, z) and (y, z). Another property of the conditional independence relation that holds only under additional conditions is:

(C5) if $X \perp\!\!\!\perp Y \mid (Z, W)$ and $X \perp\!\!\!\perp Z \mid (Y, W)$, then $X \perp\!\!\!\perp (Y, Z) \mid W$.

The condition (C5) does not hold universally but it is required a non-strict logical relationships between Y and Z.

Proposition 2.8 The condition (C5) holds if the joint density p of all variables is strictly positive.

Proof. Suppose $p(x, y, z, w) > 0$, $X \perp\!\!\!\perp Y \mid (Z, W)$ as well as $X \perp\!\!\!\perp Z \mid (Y, W)$, then the equivalent statement (S7) can be applied and

$$p(x, y, z, w) = a(x, y, w)b(y, z, w) = h(x, z, w)k(y, z, w) \tag{2.3}$$

for suitable strictly positive functions a, b, h, k. Whereas a continuous density p has been supposed, for all z it holds

$$a(x, y, w) = \frac{h(x, z, w)k(y, z, w)}{b(y, z, w)}.$$

Therefore, fixing $z = z_0$, we can write

$$a(x, y, w) = \tau(x, w)\phi(y, w)$$

with $\tau(x, w) = h(x, z_0, w)$ and $\phi(y, w) = k(y, z_0, w)/b(y, z_0, w)$. Thus the equation (2.3) becomes

$$p(x, y, z, w) = \tau(x, w)\phi(y, w)b(x, z, w),$$

and hence $X \perp\!\!\!\perp (Y, Z) \,|W$.

□

Lauritzen (1996) provided a clarification of the conditions (C1)-(C5) thinking them as formal expressions with a meaning not strictly related to probability. It is supposed that the three random variables represent the events: knowledge of a subject and reading a book. Thus, the expression $X \perp\!\!\!\perp Y \,|Z$ can be translated as: "known Z, reading the book Y is irrelevant for reading the book X". Similarly the four conditions (C1)-(C4) become:

(i) if, knowing Z, reading Y is irrelevant for reading X, then reading X is irrelevant for reading Y;

(ii) if, knowing Z, reading Y is irrelevant for reading X, then reading Y is irrelevant for reading any chapter U of the book X;

(iii) if, knowing Z, reading Y is irrelevant for reading X, then reading Y is still irrelevant for reading X even if any chapter U of X has been read;

(iv) if, knowing Z, reading Y is irrelevant for reading X, and knowing Y besides Z, reading W is irrelevant for reading X, then knowing Z reading both W and Y is irrelevant for reading X.

The condition (C5) is not treated in this sense because slightly more subtle.

2.3 Markov Properties

In this section it is considered Markov properties relative to graphs, widely discussed by Cowell *et al.* (1999). Henceforth, it is taken into account the conditional independence applied to a collection of random variables X_v,

$v{\in}V$ that take values in probability spaces $\mathcal{X}_v$. Additionally, it is defined A a subset of V, and let $\mathcal{X}_A = \mathrm{x}_{v\in A}\mathcal{X}_v$, and $\mathcal{X}{=}\mathcal{X}_V$. Elements of $\mathcal{X}_A$ are denoted $x_A = (x_v)_{v\in A}$.

2.3.1 Markov Properties for undirected graphs

Let $\mathcal{G}$ be an undirected graph representing the collection of random variables X_v, for $v \in V$. Let $\mathcal{B}$ be a collection of subsets of V, and finally let $\psi_B(x)$, for $B \in \mathcal{B}$, be a non-negative function of x such that $x_{\mathcal{B}} = (x_v)_{v\in B}$.

Definition 2.9 It is defined a $\mathcal{B}$-hierarchical distribution, the joint distribution P for X such that its probability density p can be factorized in the following way

$$p(x) = \prod_{B\in\mathcal{B}} \psi_B(x). \tag{2.4}$$

□

Consider a graph $\mathcal{G}$ represented by the set of nodes $V = (A, B, C)$ and with hierarchical distribution $\mathcal{B}{=}\{(A, B), (B, C)\}$, then the joint density function can be factorized as $p(x_A, x_B, x_C) = \pi(x_A, x_B)\tau(x_B, x_C)$. Thus, (S7) gives $X_A \perp\!\!\!\perp X_C \mid X_B$. Similarly, suppose that $V = (A, B, C)$ and $\mathcal{B} = \{(A, B), (B, C), (A, C)\}$, then the joint density function can be factorized as $p(x_A, x_B, x_C){=}\pi(x_A, x_B)\ \tau(x_B, x_C)\ \psi(x_A, x_C)$. In this situation, looking at the equivalent undirected graph in Figure 2.14 (b), X_A and X_C cannot be said to be independent given X_C, thus not all factorizations produce conditional independence.

Although any subset in $\mathcal{B}$ is obviously a complete subset of $\mathcal{G}$, a graph $\mathcal{G}$ can contain other complete sets not belonging to $\mathcal{B}$. An example is represented by the clique $\{A, B, C\}$ in Figure 2.14 (b). Now, let $\mathcal{C}$ denote the collection of cliques of $\mathcal{G}$, it can be concluded that every $\mathcal{B}$-hierarchical distribution is also $\mathcal{C}$-hierarchical because any subset in $\mathcal{B}$ is included in some cliques in $\mathcal{C}$. For this reason the cliques are preferred to be considered when it is referred to the conditional independence properties of the hierarchical

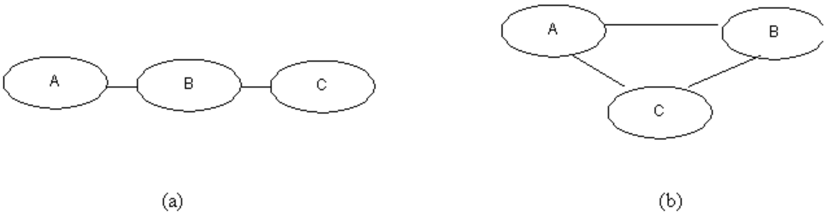


Figure 2.14: Undirected graphs derived by the factorization $p(x_A, x_B, x_C) = \pi(x_A, x_B)\tau(x_B, x_C)$ (a) and $p(x_A, x_B, x_C) = \pi(x_A, x_B)\ \tau(x_B, x_C)\ \psi(x_A, x_C)$ (b).

distributions.

Consider the class $\mathcal{A}$ of complete subsets of $\mathcal{G}$. If a non-negative functions ψ_A exists, that depend on x through x_A, for all $A \in \mathcal{A}$, and there exist a product measure $\mu = \otimes_{v \in V} \mu_v$ on $\mathcal{X}$, such that the probability measure P on $\mathcal{X}$ has density p with respect to μ with the following form

$$p(x) = \prod_{A \in \mathcal{A}} \psi_A(x_A), \tag{2.5}$$

then we can say that P is defined $\mathcal{A}$-hierarchical and factorizes according to $\mathcal{G}$. It is worth noting that μ can be chosen with arbitrariness and there are different ways to multiply groups of functions ψ_A. Thus, the functions ψ_A are not uniquely determined but they are considered as factor potentials of P.

Factorization. Suppose, without loss of generality, that $\mathcal{A}$ is represented only by the set of cliques $\mathcal{C}$ of $\mathcal{G}$. In this situation the factorization becomes

$$p(x) = \prod_{C \in \mathcal{C}} \psi_C(x). \tag{2.6}$$

When the previous equation holds, P is said to be $\mathcal{C}$-hierarchical and it satisfies the factorization property (F).

For example, in the graph in Figure 2.18 four cliques can be recognized, *i.e.* $C_1 = (x_1, x_2, x_3)$, $C_2 = (x_2, x_3, x_4)$, $C_3 = (x_4, x_5)$, $C_4 = (x_4, x_6)$; applying factorization it can be written $f(x_1, x_2, ..., x_6) = \prod_{i=1}^{4} \psi_{C_i}(x)$

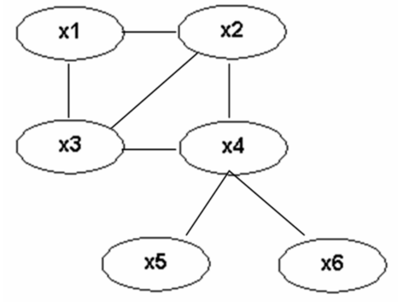


Figure 2.15: Undirected graph that satisfies the Markov properties.

□

There are three Markov properties associated with the undirected graph $\mathcal{G}$. A probability measure P on $\mathcal{X}$ satisfies:

the **pairwise Markov property** (P), relative to $\mathcal{G}$, if given two non-adjacent vertices $(\alpha, \beta) \in V$, it can be written

$$\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\};$$

the **local Markov property** (L), relative to $\mathcal{G}$, if given a vertex $\alpha \in V$, it can be written

$$\alpha \perp\!\!\!\perp V \setminus \mathrm{cl}(\alpha) \mid \mathrm{bd}(\alpha);$$

the **global Markov property** (G), relative to $\mathcal{G}$, if given three disjoint subsets (A,B,S)∈ V, where S is separator of A and B, it can be written

$$A \perp\!\!\!\perp B \mid S.$$

For example, applying the three Markov properties with respect to the graph in Figure 2.18, it can be said that:
$x_2 \perp\!\!\!\perp x_6 \mid \{x_1, x_3, x_4, x_5\}$ - pairwise Markov property;
$x_1 \perp\!\!\!\perp \{x_4, x_5, x_6\} \mid \{x_2, x_3\}$ - local Markov property;
$\{x_1, x_2, x_3\} \perp\!\!\!\perp \{x_5, x_6\} \mid x_4$ - global Markov property.

Proposition 2.10 (Lauritzen 1996). For any undirected graph $\mathcal{G}$ and any probability distribution P on $\mathcal{X}$ it holds that

$$\text{(F)} \Rightarrow \text{(G)} \Rightarrow \text{(L)} \Rightarrow \text{(P)}.$$

Proof. First it is shown that (F) $\Rightarrow$ (G). Consider a triplet (A,B,S) of disjoint subsets of $\mathcal{G}$ such that S is an (A,B)-separator. It is denoted $\tilde{A}$ the connectivity components in $\mathcal{G}_{V\setminus S}$ which contain A and $\tilde{B}$=$V\setminus(\tilde{A} \cup S)$ which contain B. The elements A and B are included in different connectivity components of $\mathcal{G}_{V\setminus S}$ because they are separated by S and, for the same reason, any clique of $\mathcal{G}$ will be either a subset of $\tilde{A} \cup S$ or of $\tilde{B} \cup S$. Thus, from the equation (2.6) it is obtained

$$p(x) = \prod_{C\in\mathcal{C}} \psi_C(x) = \prod_{C\in\mathcal{C}_A} \psi_C(x) \prod_{C\in\mathcal{C}\setminus\mathcal{C}_A} \psi_C(x) = l(x_{\tilde{A}\cup S})k(x_{\tilde{A}\cup S}).$$

Therefore, for the property (S7) of conditional independence it is deduced that $\tilde{A} \perp\!\!\!\perp \tilde{B} \mid S$, whilst applying the (C2) twice it is obtained $A \perp\!\!\!\perp B \mid S$ which is (G).

Now it is shown that (G) $\Rightarrow$ (L). If (G) holds, then (L) holds too because bd(α) separates α from $V\setminus cl(\alpha)$.

Finally, it is shown that (L) $\Rightarrow$ (P). Consider the property (L) and suppose that it holds. Since α and β are non-adjacent vertices, β belongs to $V\setminus cl(\alpha)$ and bd(α)$\cup$($V\setminus$cl(α))$\setminus\{\beta\}$=$V\setminus\{\alpha, \beta\}$. Applying (C3) to property (L) it is obtained $\alpha \perp\!\!\!\perp V\setminus$ cl(α)$|$ $V\setminus\{\alpha, \beta\}$. Now, applying condition (C4), the required result $\alpha \perp\!\!\!\perp \beta \mid V\setminus\{\alpha, \beta\}$ follows.

□

Theorem 2.11 If a probability distribution on $\mathcal{X}$ is such that (C5) holds for disjoint subsets A,B,C,D then

$$\text{(G)} \Leftrightarrow \text{(L)} \Leftrightarrow \text{(P)}.$$

Proof. Assume that (i) both (P) and condition (C5) hold and that (ii) considering three disjoint subsets (A,B,S) of $\mathcal{G}$, S is (A,B)-separator and finally that (iii) A and B are non-empty. It is proved that (P) implies (G) proceeding by reverse induction on the number of vertices n=$|S|$ in S.

If n =$|V| - 2$, then (G) follows directly from (P) because the subsets

A and B has one vertex only. Now, assume that the required conditional independence holds for S with more than n vertices and consider the case $|S| = n < |V| - 2$. Firstly, two different situations have to be considered: $A \cup B \cup C = V$ (so that at least one between A and B has more than one element, for example presume A) and $A \cup B \cup C \subset V$. If $A \cup B \cup C = V$, since S separates A from B, if $\alpha \in A$ then $S \cup \{\alpha\}$ separates B from $A \setminus \{\alpha\}$, and $S \cup A \setminus \{\alpha\}$ separates B from α. As a consequence, it can be written $B \perp\!\!\!\perp A \setminus \{\alpha\} \mid S \cup \{\alpha\}$ and $B \perp\!\!\!\perp \alpha \mid S \cup A \setminus \{\alpha\}$ by the inductive hypothesis. Applying (C5) it follows that $A \perp\!\!\!\perp B \mid S$.

Now, consider the second case, $A \cup B \cup C \subset V$, and select $\alpha \in V \setminus (A \cup B \cup C)$. Then $S \cup \{\alpha\}$ separates A from B and (G) gives $A \perp\!\!\!\perp B \mid S \cup \{\alpha\}$. Further, either $A \cup S$ separates B and $\{\alpha\}$ or $B \cup S$ separates A from $\{\alpha\}$. Assuming the former $\alpha \perp\!\!\!\perp B \mid A \cup S$ is derived, and (C5) gives $B \perp\!\!\!\perp (A \cup \{\alpha\}) \mid S$. Thus, it can be concluded that the required independence, $A \perp\!\!\!\perp B \mid S$, holds. The proof of the latter case is similar.

□

In Hammersley and Clifford (1971) proved that under the assumptions of all discrete state spaces and positive and continuous density of the probability distribution, (P) implies (F) and thus all Markov property are equivalent. It is worth noting that whilst the condition of a continuous density can be considerably relaxed, the positivity is indispensable.

When the triplet (A,B,S) of disjoint subsets of V forms a decomposition of $\mathcal{G}$ also the Markov properties decompose, as seen in the following proposition.

Proposition 2.12 Let $\mathcal{G}$ be a graph decomposed in the triplet (A,B,S), if both $P_{A \cup S}$ and $P_{B \cup S}$ factorizes in (F) with respect to $\mathcal{G}_{A \cup S}$ and $\mathcal{G}_{B \cup S}$, then P factorizes with respect to $\mathcal{G}$ and the density p can be written as

$$p(x) = \frac{p_{A \cup S(x_{A \cup S})} p_{B \cup S}(x_{B \cup S})}{p_S(x_S).} \tag{2.7}$$

Proof. Assume that P factorizes with respect to $\mathcal{G}$ such that

$$p(x) = \prod_{C \in \mathcal{C}} \psi_C(x).$$

Since (A,B,S) is a decomposition of $\mathcal{G}$ by assumption, all cliques are subsets of either $A \cup S$ or of $B \cup S$ and, as a consequence, p is factorized as

$$p(x) = \prod_{C \in \mathcal{A}} \psi_C(x) \prod_{C \in \mathcal{B}} \psi_C(x) = a(x_{\mathrm{A}\cup S})b(x_{B\cup S}).$$

It is defined, by direct integration, the marginal distribution $p(x_{\mathrm{A}\cup S})$ as the product $a(x_{\mathrm{A}\cup S})g(x_S)$, where $g_{(}x_S) = \int g(x_{B\cup S})\gamma_B(dx_B)$. The other marginal distribution $p(x_{B\cup S})$ is defined similarly. Substituting the values of $a(x_{\mathrm{A}\cup S})$ and $b(x_{B\cup S})$ as functions of the marginal densities $p(x_{\mathrm{A}\cup S})$ and $p(x_{B\cup S})$, (2.8) is given.

□

This result holds also for the global Markov property; Lauritzen (1996) gave a proof. In general for a decomposable graph $\mathcal{G}$, if p factorizes as

$$p(x) = \frac{\prod_{C \in \mathcal{C}} p(x_C)}{\prod_{S \in \mathcal{S}} p(x_S)}, \tag{2.8}$$

where $\mathcal{S}$ is the set of separators of the cliques $\mathcal{C}$ of $\mathcal{G}$, then the distribution P is Markov with respect to $\mathcal{G}$.

2.3.2 Markov Properties for DAGs

In this section the Markov properties for directed acyclic graphs $\mathcal{D}$ are considered.

Definition 2.13 A probability distribution P admits a *recursive factorization* according to $\mathcal{D}$ if a (σ-finite) measures μ_v over $\mathcal{X}$ exists and non-negative *kernels* functions $K^v(\cdot,\cdot)$, for all $v \in V$ defined on $\mathcal{X}_v \times \mathcal{X}_{\mathrm{pa}(v)}$ such that

$$\int k^v(y_v, x_{\mathrm{pa}(v)})\mu_v(dy_v) = 1$$

and P has density p with respect to the product measure $\mu = \otimes_{v\in V}\mu_v$ given by

$$p(x) = \prod_{v\in V} k^v(x_v, x_{\mathrm{pa}(v)}).$$

□

Directed factorization (DF). A probability measure P on $\mathcal{X}$ satisfies the property (DF) if it admits a recursive factorization.

□

It is further deduced by induction that if P admits a recursive factorization. Then the conditional distribution of $X_v \mid X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}$ has densities kernels $K^v(\cdot, x_{\mathrm{pa}(v)})$. Hence its density p can be written as

$$p(x) = \prod_{v\in V} p(x_v \mid x_{\mathrm{pa}(v)}). \tag{2.9}$$

Lemma 2.14 If P admits a recursive factorization according to $\mathcal{D}$, then it factorizes according to $\mathcal{D}^m$, where $\mathcal{D}^m$ is the (undirected) moral graph formed from $\mathcal{D}$. Moreover, the probability distribution P satisfies the global Markov property relative to $\mathcal{D}^m$.

Proof. Since the moral graph $\mathcal{D}^m$ is obtained marrying parents (and replacing directed edges by undirected edges) sets $\{v\} \cup \mathrm{pa}(v)$ in $\mathcal{D}^m$ are complete by construction. Thus, it is defined $\psi_{\{v\}\cup\mathrm{pa}(v)} = k^v$. Whereas (F) $\Rightarrow$ (G) in an undirected graph (see Proposition 2.10), also the last statement is proved.

□

Let $\mathrm{bl}(v)$, called *Markov blanket* of v, to denote the set of neighbours of v in $\mathcal{D}^m$ and the set of v's parents, children, and children's parents in $\mathcal{D}$, *i.e.* $\mathrm{bl}(v) = \mathrm{pa}(v) \cup \mathrm{ch}(v) \cup \{w\colon \mathrm{ch}(w) \cap \mathrm{ch}(v) \neq \emptyset\}$. The local Markov property relative to $\mathcal{D}^m$ gives $v \perp\!\!\!\perp V \,|\text{bl}(v)$.

Proposition 2.15 If a probability distribution P admits a recursive factorization according to $\mathcal{D}$, then the marginal distribution P_A on an ancestral set A admits a recursive factorization according to $\mathcal{D}_A$.

Corollary 2.16 - Directed global Markov property (DG) - Let $(X_\alpha$,X_β,$X_S)$ be a triplet of disjoint subsets such that X_S separates X_α from X_β in the moral graph of the smallest ancestral set containing $\{\alpha\} \cup \{\beta\} \cup \{S\}$, *i.e.* $\mathcal{D}^m_{\text{An}(\alpha,\beta,S)}$. If P factorizes according to $\mathcal{D}$, then

$$X_\alpha \perp\!\!\!\perp X_\beta \mid X_S.$$

In this case P is said to be a *directed Markov field* over $\mathcal{D}$.

□

The power of the global Markov property (relative to both undirected and directed graphs) is represented by its ability to provide a general rule to decide whether two groups of variables $X\alpha$ and X_β are conditionally independent given a third group X_S. The global Markov property is further considered the strongest of the Markov properties because the associated list of conditional independence statements strictly includes the statements associated with the other properties.

The concept of conditional independence referred to DAGs was studied by Pearl (1986) who gave an alternative formulation of the global directed Markov property through the concept of directional separation, or *d-separation*. In order to explain the notion of d-separation it is useful to introduce the three kind of connections that can be found in a DAG: a *serial connection*, if a node mediates the communication between other nodes ($\rightarrow X \rightarrow$); *diverging connection*, if a node has two or more children ($\leftarrow X \rightarrow$); *converging connection*, also called V-configuration, if some nodes meet head-to-head at another ($\rightarrow X \leftarrow$).

In order to give a general definition of d-separation let, X_α, X_β and X_S, for all $(\alpha, \beta) \neq S$, be disjoint sets of nodes that belong to V in a directed acyclic graph $\mathcal{D}$, it is said X_S d-separates X_α from X_β if it blocks every trail from X_α to X_β. A trail π between two nodes is blocked by X_S if, either

1. for every node X_γ, with $\gamma \in S$, X_γ has serial or diverging connections,

or

2. X_γ has converging connections, and nor X_γ neither its descendants are in X_S.

Two nodes that are not d-separated are called *d-connected* and it is called *active* a trail that is not blocked by X_S.

Proposition 2.17 Consider a directed acyclic graph $\mathcal{D}$ and its disjoint subsets X_α, X_β and X_S, for all $(\alpha, \beta) \neq S$. X_S d-separates X_α from X_β if and only if X_S separates X_α from X_β in the graph $\mathcal{D}^m_{\mathrm{An}(\alpha,\beta,S)}$, *i.e.* the moral graph of the subgraph induced by $\{\alpha\} \cup \{\beta\} \cup \{S\}$.

Proof. Assume that X_α and X_β are not d-separated by X_S. As a consequence, from X_α to X_β there is a trail not blocked by X_S, thus active. An example is showed in Figure 2.16. Since the trail is active, either there is some vertex X_γ with a converging connection which, either belong to X_S, or has descendants in X_S; otherwise, either of the subpaths away from X_γ either meets another arrow, X_γ has descendants in X_S, or connects all the way to X_α or X_β. Thus, the $\mathrm{An}(X_\alpha \cup X_\beta \cup X_S)$ must contain all the vertices in the trail. The moral graph corresponding to the active chain contains a trail from X_α to X_β in $\mathcal{D}^m_{\mathrm{An}(\alpha,\beta,S)}$ and circumventing S.

On the contrary, assume that X_α and X_β are not separated in $\mathcal{D}^m_{\mathrm{An}(\alpha,\beta,S)}$. Then a trail that circumvents X_S can be found in the graph. This trail contains both edges of the original graph and edges that marry parents. Since marriages derives from converging connection at some node X_γ, if $X_\gamma \in X_S$ or it has descendants in S, the connection does not block the trail. Otherwise, if X_γ is not in X_S or it has not descendants in X_S, a new trail can be drawn with one less head-to-head meeting and using the line

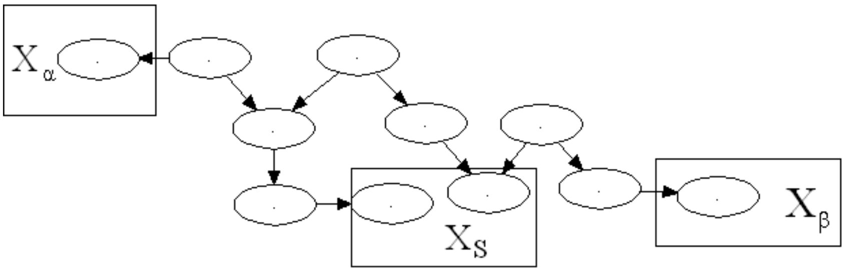


Figure 2.16: An active trail from X_α to X_β.

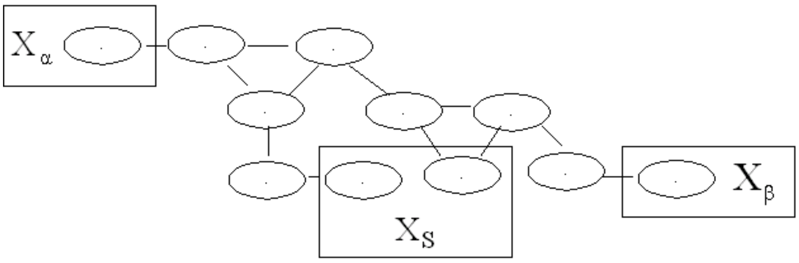


Figure 2.17: Moral graph of Figure 2.15.

of descent. A representation is in Figure 2.17. Repeating the argument an active trail from X_α to X_β can be created in $\mathcal{D}^m_{\mathrm{An}(\alpha,\beta,S)}$.

□

An alternative and more straightforward method to analyze conditional independence in directed acyclic graphs follows by this proposition. In fact if X_S separates X_α from X_β in $\mathcal{D}^m_{An(\alpha,\beta,S)}$, then the global Markov property relative to undirected graphs gives that $X_\alpha \perp\!\!\!\perp X_\beta \mid X_S$. For example, Figure 2.18 (b) shows the moral graph of the smallest ancestral set including all the variables involved. Since S separates X from Y in the moral graph of the subgraph induced then the global Markov property can be applied, and it can be concluded that $X \perp\!\!\!\perp Y \mid S$.

Local directed Markov property (DL). Consider a directed acyclic graph $\mathcal{D}$, if for any vertices $v \in V$

$$v \perp\!\!\!\perp \mathrm{nd}(v) \mid \mathrm{pa}(v), \tag{2.10}$$

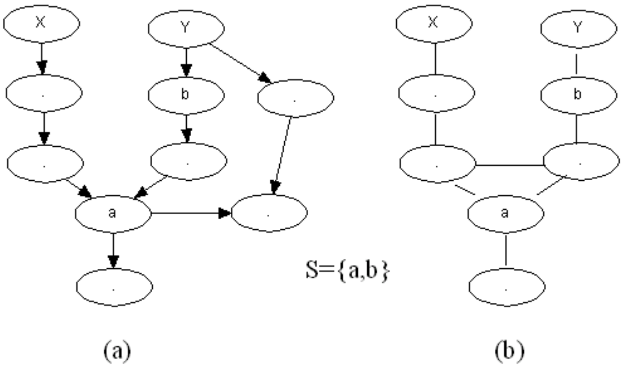


Figure 2.18: Since S separates X from Y in the moral graph of the subgraph induced by $X \cup Y \cup S$ (b), the global Markov property gives $X \perp\!\!\!\perp Y \mid S$.

then P obeys the local directed Markov property. Instead of all non-descendant, consider the predecessors $\mathrm{pr}(v)$ of v in some given well-ordering of the nodes, so that

$$v \perp\!\!\!\perp pr(v) \mid \mathrm{pa}(v). \tag{2.11}$$

Then it is said that P obeys to the *ordered directed Markov property (DO).*

□

Theorem 2.18 Let $\mathcal{D}$ be a directed acyclic graph. For a probability distribution P on $\mathcal{X}$ which has density with respect to a product measure μ, the following conditions are equivalent:

(DG) P admits a recursive factorization according to $\mathcal{D}$;

(DG) P obeys to the global directed Markov property, relative to $\mathcal{D}$;

(DL) P obeys to the local directed Markov property, relative to $\mathcal{D}$;

(DO) P obeys to the ordered directed Markov property, relative to $\mathcal{D}$.

Proof. Corollary 2.16 proves that (DF) implies (DG). Considering a vertex {v} and its non-descendants, $v \cup \mathrm{nd}(v)$ is an ancestral set and $\mathrm{pa}(v)$ separates v from $\mathrm{nd}(v)\backslash\mathrm{pa}(v)$ in $\mathcal{D}^m_{v \cup \mathrm{nd}(v)}$. Thus, (DG) implies (DL). Since $\mathrm{pr}(v) \subseteq \mathrm{nd}(v)$, (DL) implies (DO). The final equivalence is proved by induction on the number of vertices $|V|$ of $\mathcal{D}$. Let v_0 be the last vertex in $\mathcal{D}$ and k^{v_0} be the conditional density of $X_{v_0}|X_{V\backslash\{v_0\}}$. Such conditional density by (DO) can be chosen to depend on $x_{\mathrm{pa}(v_0)}$ only. On the contrary, by inductive hypothesis, the marginal distribution $X_{V\setminus\{v_0\}}$ obeys the ordered directed Markov property admitting a factorization. If this factorization with k^{v_0} is combined, then also P admits the factorization, proving the final condition of the theorem.

□

It is denoted $M(\mathcal{D})$ the set of distributions for a directed acyclic graph $\mathcal{D}$ called *directed Markov distributions* and such that any of the four conditions in Theorem 2.18 is satisfied.

It is considered now a perfect directed acyclic graph $\mathcal{D}$ and its undirected version $\mathcal{D}^{\sim}$. The directed Markov property on $\mathcal{D}$ and the factorization property on $\mathcal{D}^{\sim}$ coincide.

Proposition 2.19 Let $\mathcal{D}$ be a perfect directed acyclic graph and $\mathcal{D}^{\sim}$ its undirected version. Then, a probability distribution P on $\mathcal{X}$ obeys the directed Markov property relative to $\mathcal{D}$ if and only if it admits a recursive factorization according to $\mathcal{D}$.

Proof. If a graph is perfect, for all $v \in V$ pa(v) is complete. Hence, $\mathcal{D}^m = \mathcal{D}^{\sim}$. Applying Lemma 2.14 then any $P \in M(\mathcal{D})$ factorized with respect to $\mathcal{D}^{\sim}$.

For the reverse assumption it is proceeded on induction by the number of vertices $|V|$ of $\mathcal{D}$. If $|V| = 1$ the proof is immediate. Now, it is assumed the proposition holds for $|V| = n$ and it is proved it holds also for $|V| = n + 1$. Let $P \in M(\mathcal{D}^{\sim})$ and a terminal vertex $v \in V$ is considered. This vertex has $\mathrm{pa}_{\mathcal{D}}(v)$ $=\mathrm{bd}_{\mathcal{D}^{\sim}}(v)$ and, being $\mathcal{D}$ perfect, $\mathrm{bd}_{\mathcal{D}^{\sim}}(v)$ is a complete set in both graphs. Hence, the triplet $(V\setminus\{v\}, \{v\}$,bd$(v))$ is a decomposition of $\mathcal{D}^{\sim}$ and for Proposition 2.12 the following factorization holds:

$$p(x) = p(x_{V\setminus\{v\}})p(x_{\mathrm{cl}(v)})/p(x_{\mathrm{bd}(v)}) = p(x_{V\setminus\{v\}})k^v(x_v, x_{\mathrm{pa}(v)}), \qquad (2.12)$$

where $\int k^v(y_v, x_{\mathrm{pa}(v)})\mu_v(dy_v) = 1$, and the first factor factorizes according to

$\mathcal{D}^{\sim}_{V\setminus\{v\}}$. Now, the inductive assumption on this factor gives the full recursive factorization of P.

□

2.3.3 Markov Properties for chain graphs

In this section Markov properties on general chain graphs $\mathcal{K} = (V, E)$ are investigated. We further assume positive density for all probability measures, so that the five conditions (C1) - (C5) on conditional independence hold.

The Markov properties relative on $\mathcal{K}$ are the following. A probability P satisfies:

(CP) the **pairwise chain Markov property**, relative to $\mathcal{K}$, if for any pair (α, β) of non-adjacent vertices with $\beta \in \mathrm{nd}(\alpha)$,

$$\alpha \perp\!\!\!\perp \beta | \mathrm{nd}(\alpha)\setminus\{\alpha, \beta\};$$

(CL) the **local chain Markov property**, relative to $\mathcal{K}$, for any vertex $\alpha \in V$,

$$\alpha \perp\!\!\!\perp \mathrm{nd}(\alpha)\setminus\mathrm{bd}(\alpha)|\mathrm{bd}(\alpha);$$

(CG) the **global chain Markov property**, relative to $\mathcal{K}$, if for any triplet (A, B, S) of disjoint subsets of V

$$A \perp\!\!\!\perp B|S,$$

where S separates A from B in the moral graph of the smallest ancestral set containing $A \cup B \cup S$, $\mathcal{K}^m_{\mathrm{An}(A\cup B\cup S)}$.

As for directed acyclic graphs a definition of d-separation exists for chain graphs. Studený and Bouckaert (1998) introduced a definition of c-separation which is equivalent to the separation property used in the global chain Markov property. These Markov properties have the characteristic to unify the properties relative to undirected graphs with those for directed graphs.

Let $V = V(1)\cup, ..., \cup V(T)$ be a dependence chain that partitions the vertex set. Each set $V(t)$ has lines between vertices only, whilst arrows point from vertices in set with lower number to those with higher number. It is defined $C(t) = V(1) \cup, ..., \cup V(t)$ as the set of *concurrent* variables. It is said that P satisfies the *block-recursive Markov property (CB)* if for any pair (α, β) of non-adjacent vertices

$$\alpha \perp\!\!\!\perp \beta | C(t^*) \backslash \{\alpha, \beta\},$$

where t^* is the smallest t having $\{\alpha, \beta\} \subseteq C(t)$. This property depends on the particular partitioning, but Frydenberg (1990) proved that if P satisfies the condition (C5) for subsets of V, then

$$\text{(CG)} \Leftrightarrow \text{(CL)} \Leftrightarrow \text{(CP)} \Leftrightarrow \text{(CB)}.$$

Now, if $V(1), ..., V(T)$ is a dependence chain of $\mathcal{K}$ or the chain components of $\mathcal{K}$, then any distribution P with density p with respect to a product measure μ factorizes as

$$p(x) = \prod_{t=1}^{T} p(x_{V(t)} | x_{C(t-1)}),$$

where $C(t)$ is defined as previously. This factorization reduces to

$$p(x) = \prod_{t=1}^{T} p(x_{V(t)} | x_{B(t)}) \tag{2.13}$$

if $B(t) = \text{pa}(V(t)) = \text{bd}(V(t))$ and p is Markov relative to $\mathcal{K}$. This factorization, essentially, is the same as that introduced for directed Markov properties even though it does not reveal all conditional independence relationships. This equality is due to the fact that chain graphs form a directed acyclic graph of its chain components. However, it should be intuitable that factorization results are more general for chain graphs than for undirected graphs. On the contrary, chain graphs contain special cases that do not allow this. For example, let $\mathcal{K}^*(t)$ be the undirected graph with vertex set $V(t) \cup B(t)$ and α adjacent to β in $\mathcal{K}^*(t)$ if either $(\alpha, \beta) \in E$ or $(\beta, \alpha) \in E$ or if $\{\alpha, \beta\} \subseteq B(t)$, *i.e.* $B(t)$ is made complete in $\mathcal{K}^*(t)$ by adding all missing edges between these and directions on existing edges are ignored. However, if all variables are discrete we have a result analogous to case of positive density of P and the pairwise Markov property relative to undirected graph that implies factorization.

Theorem 2.20 Let P be a probability distribution on a discrete sample space and with strictly positive density p. This satisfies the pairwise chain graph Markov property with respect to $\mathcal{K}$ if and only if it factorizes as

$$p(x) = \prod_{t=1}^{T} \frac{p(x_{V(t)\cup B(t)})}{p(x_{B(t)})}, \tag{2.14}$$

and each of the numerators factorizes on the graph $\mathcal{K}^*(t)$.

Proof (Lauritzen 1996).

□

Corollary 2.21 If the density p of a probability distribution factorizes as in (2.13), it also factorizes according to the moral graph $\mathcal{K}^m$ and therefore obeys the undirected global Markov property relative to $\mathcal{K}^m$.

Proof See Cowell *et al.* (1999).

□

Chapter 3

Bayesian Networks for discrete variables

3.1 Bayesian Networks

Bayesian networks represent probabilistic models employing graphical structures to describe casual relationships between random variables which can be discrete and/or continuous. Probabilistic networks with discrete random variables only are the simplest form of these systems since they produce an exact analysis. When inference is performed, after observing one or more variables and entering evidence in the domain, the probability of the other variables are updated. This process requires operations, such as marginalization and conditioning as result of compiling the model, and therefore requires also the construction of a junction tree of cliques which are the largest set of variables under investigation. Whereas these cliques are handled simultaneously, they may lead computational problems. For this reason the individual cliques in the triangulated moral graph are required to have a size such to allow the extension of calculations to the complete set of variables.

In this section each fundamental stage is defined to describe in detail the algorithm for propagating information through a junction tree and other operations. Finally, each step is illustrated through an application taken by Lauritzen and Spiegelhalter (1988).

3.1.1 Definition of Bayesian Networks

A *Bayesian network* (BN) can be defined as a pair $(\mathcal{D},\mathcal{P})$ that satisfies the Markov properties for directed graphs. In this notation $\mathcal{D}=(V,E)$ is a DAG and $\mathcal{P}$ is the joint probability distribution of the nodes in the graph. The vertices in V represent random variables and the edges E between the variables indicate conditional probabilistic dependencies.

Random variables in V can be discrete or continuous but for the sake of simplicity in this section it is referred only to the discrete case in which for each variable a finite set of states is defined and a probability measure is associated. Thus, a table specifying the conditional probabilities $p(x_k|x_{\mathrm{pa}(k)})$ is attached to each variable X of $\mathcal{D}$. As previously described in Section 2.3.2, the Markov properties on directed acyclic graphs may lead to factorize the conditional probabilities $p(x_k|x_{\mathrm{pa}(k)})$ in terms of potentials (see S 3.1.3). If $\mathrm{pa}(k) = \emptyset$, the table consists of unconditional probabilities, said prior probabilities.

Let $U \subseteq V$, it is denoted by $\mathcal{U}$ or $\mathcal{X}_U$ the Cartesian product of the state sets of the nodes of U which is the *space* of U. Similarly, the space of V is denoted $\mathcal{V}$ or $\mathcal{X}$, the space of $U \cup W$ by $\mathcal{U} \cup \mathcal{W}$, and that of $W \backslash U$ by $\mathcal{W} \backslash \mathcal{U}$. A *potential* on U is a mapping from $\mathcal{U}$ to the non-negative real numbers $\mathbb{R}_0$. Particularly, the table of conditional probabilities $p(x_v|x_{\mathrm{pa}(v)})$ is a potential on $v \cup \mathrm{pa}(v)$, but with the constrain that, for a fixed parent configuration, the probabilities must be normalized to sum to unity when summed over the states in v. Then, their product gives the joint probability distribution over $\mathcal{X}$.

Definition 3.1 Let $U \subseteq W \subseteq V$, let ϕ be a potential on U and let $x \in \mathcal{W}$. It is defined $\phi(x) = \phi(y)$, where y is the projection of x onto $\mathcal{W}$. Then the potential ϕ is extended to W.

□

Definition 3.2 Let ϕ and ψ be potentials on U and W such that have been extended to $U \cup W$. It is defined their:

(i) *product* $\phi\psi$ on $U \cup W$ by $(\phi\psi)(x)=\phi(x)\psi(x)$;

(ii) *sum* $\phi + \psi$ on $U \cup W$ by $(\phi + \psi)(x) = \phi(x) + \psi(x)$;

(iii) *division* by$(\phi/\psi)(x) = \phi(x)/\psi(x)$ if $\psi(x) \neq 0$, and zero otherwise.

□

Definition 3.3 Let $W \subseteq U \subseteq V$ and let ϕ be a potential on U. It is defined the *margin* $\sum_{U \setminus W} \phi$ of ϕ on W as

$$\left(\sum_{U \setminus W} \phi\right)(x) = \sum_{z \in \mathcal{U} \setminus \mathcal{W}} \phi(z.x)$$

for $x \in \mathcal{W}$ and $z.x \in \mathcal{W}$ with projections x to $\mathcal{U}$ and z to $\mathcal{W} \setminus \mathcal{U}$.

□

3.1.2 Inference in Bayesian networks

The main aim of inference in Bayesian networks is to calculate updated probabilities when a particular information is achieved, *i.e.* evidence is observed. For example, let X be a random variable with n states, it is assumed to get information that X is in the state i. So, all the states of X except i are impossible and probability zero is associated to them. It could be of interest the probability of another node connected with X (*e.g.* its parent) given the new information on X. This probability can be straightforward to calculated applying the Bayes theorem. However, the probabilities of all the nodes in the network can be updated using this method only if the network is small and each node has few states, whilst it becomes difficult to make inference if multiple pieces of evidence are entered. In this case algorithms based on the construction of junction trees can be used. In the following subsections all of the stages are described. Summering, given a Bayesian network, it must be moralized, and then triangulated in order to make it decomposable and to allow that a junction tree exists (see Section 3.1.3). Before the junction tree can be used, it must first be *initialized* to provide a local representation of the overall distribution. Then, after evidence is entered, local computations which yield marginal and conditional distributions are realized.

Initialization

The initial graphical approach to the problem can be examined in terms of conditional independence through the Markov properties relevant to undirected, directed acyclic or chain graphs.

It is denoted $\mathcal{T}$ a junction tree of cliques $\mathcal{C}$ for the triangulated moralized graph $\mathcal{D}^{mt}$ of $\mathcal{D}$. When $\mathcal{T}$ is disconnected it can be easily managed by considering each component, therefore $\mathcal{T}$ is assumed to be connected.

Reminding the equation (2.8) in Section 2.3.1 a factorization over $\mathcal{D}$ of the density $p(\cdot)$ of the probability distribution P on $\mathcal{D}$ is given by

$$\begin{aligned} p(x) = \prod_{v \in V} p(x_v \mid x_{\mathrm{pa}(v)}) = \prod_{v \in V} \psi_{\{v\} \cup \mathrm{pa}(v)}(x) \\ \propto \prod_{v} Z^{-1}(x_{\mathrm{pa}(v)}) \prod_{A \in \mathcal{A}_v} \psi_A(x_A), \quad (3.1) \end{aligned}$$

where $\mathcal{A}_v$ denotes the set of maximal subset of $\{v\} \cup \mathrm{pa}(v)$ that is complete in $\mathcal{D}^m$ of $\mathcal{D}$ and contains at least one child in v.

Now a factorization over $\mathcal{T}$ of the density $p(\cdot)$ of the joint distribution P on $\mathcal{T}$ is considered. It is made as follows:

(i) it is associated a potential ϕ_C to each clique $C \in \mathcal{C}$ and a potential ϕ_S to each separator $S \in \mathcal{S}$ connecting two cliques in $\mathcal{T}$;

(ii) all the potentials, $\{\phi_C,\, C \in \mathcal{C}\}$ and $\{\phi_S,\, S \in \mathcal{S}\}$ are initialized to have value unity;

(iii) for each node v, a clique C of $\mathcal{T}$ such that $\{v\} \cup \mathrm{pa}(v) \subseteq C$ is considered and each factor in (3.1) is multiplied into the potential of any one clique of $\mathcal{T}$. The moralization of the directed acyclic graph ensures that one such clique always exists, and even though there are more than one such cliques it does not make difference which is taken into account;

(iv) the final result is the following factorization of $\mathcal{D}$

$$p(x) = \frac{\prod_{C \in \mathcal{C}} \phi_C(x_C)}{\prod_{S \in \mathcal{S}} \phi_S(x_S)}, \tag{3.2}$$

where $\phi_S \equiv 1$.

Passing flow of information between adjacent cliques and reaching equilibrium

It is called *charge* on $\mathcal{T}$ a set of non-negative potential functions $\Phi = \{\phi_A, A \in \mathcal{C} \cup \mathcal{S}\}$. For any charge, its *contraction* is defined by the right-hand side of (3.2) above. Actually, the initialization phase described previously is an *initial representation*, and its potentials are *initial potentials*. Whilst, Φ are called (*generalized potential*) *representation* of P when the expression (3.2) holds.

The algorithm for propagating information include a sequence of messages, or *flows*, which pass along the edges of $\mathcal{T}$ and involve the potentials on exactly one clique and one separator. It is shown the way in which a flow passes from a clique C_1 in $\mathcal{T}$, called the *source*, to an adjacent clique C_2 in $\mathcal{T}$, called the *sink*, along the edge of the separator S_0 which joins them. It is considered the charge $\Phi = (\{\phi_C, C \in \mathcal{C}\}), \{\phi_S, S \in \mathcal{S}\}$ which, as effect of the flow, is replaced by a new charge $\Phi^* = (\{\phi^*_C, C \in \mathcal{C}\}, \{\phi^*_S, S \in \mathcal{S}\})$. Now, the new potentials on S_0 and C_2 are obtained using Definitions 3.2 (i) and (iii), and Definition 3.3 giving the following expressions

$$\phi^*_{S_0} = \sum_{C_1 \setminus S_0} \phi_{C_1}, \tag{3.3}$$

and

$$\phi^*_{C_2} = \phi_{C_2} \lambda_{S_0}, \tag{3.4}$$

where the *update ratio* λ_{S_0} is given by the ratio

$$\lambda_{S_0} = \phi^*_{S_0} / \phi_{S_0} \tag{3.5}$$

and it derives by passage the flow along S_0 into C_2. A flow is said *consistent* if $\sum_{C \setminus S} \phi_C = \phi_S$ for any $C \in \mathcal{C}$ and neighbouring $S \in \mathcal{S}$. If a flow is consistent, its passage does not affect a charge Φ. Furthermore, passage of flow does not affect the contraction of a charge.

A particular sequence of flows are the *active* flows. The definition of active flows is related to a *schedule*, where a schedule is an ordered list of directed edges of $\mathcal{T}$ which specifies the flows that are to be pass and in what order. Now, a flow is said to be active if, before sending the flow, the source itself has already received active flows from all its neighbours in the tree, with the possible exception of the sink. Thus, an active flow originates by a leaf which is a clique in $\mathcal{T}$ with only one neighbour. A schedule containing only active flows is said *active*. Whilst, if it contains an active flow in all directions it is *full*, and *fully active* if it is both full and active. It has been proved (see Cowell *et al.* (1999)) that there is a fully active flow for any tree.

A subtree is now considered. A subtree $\mathcal{T}'$ of $\mathcal{T}$ is a connected collection of cliques and their edges belonging to $\mathcal{T}$. A clique C is a neighbour of a subtree $\mathcal{T}'$ if C is not a clique of $\mathcal{T}'$ but is a clique of $\mathcal{T}$ and is connected to $\mathcal{T}'$ by an edge in $\mathcal{T}$. Thus, let $\mathcal{T}'$ be a subtree of $\mathcal{T}$ containing vertices $\mathcal{C}' \subseteq \mathcal{C}$ and edges $\mathcal{S}' \subseteq \mathcal{S}$. The set of variables $U' := \bigcup\{C : C \in \mathcal{C}'\}$ associated with $\mathcal{T}'$ is the *base* of $\mathcal{T}'$. If $\Phi = (\{\phi_C, C \in \mathcal{C}\}, \{\phi_S, S \in \mathcal{S}\})$ is a charge of $\mathcal{T}$, then $\Phi' := (\{\phi_C, C \in \mathcal{C}'\}, \{\phi_S, S \in \mathcal{S}'\})$ is a charge of $\mathcal{T}'$. Now, considering a certain schedule of flows, if at a given stage of the schedule the subtree $\mathcal{T}'$ has already received active flows from all its neighbours, then $\mathcal{T}'$ is *live*.

Theorem 3.4 Let $\Phi^0 = (\{\phi^0_C, C \in \mathcal{C}\}, \{\phi^0_S, S \in \mathcal{S}\})$ be a charge for an initial representation and for a function f that factorizes on $\mathcal{T}$. Suppose that a sequence of flows passes according to some schedule, then, whenever $\mathcal{T}'$ is live, the potential on $\mathcal{T}'$ is the sum-margin $f_{U'}$ of f on U'.

Proof. See Cowell *et al.* (1999).

□

Corollary 3.5 Whenever a clique C is live, it has potential f_C.

Corollary 3.6 Whenever active flows have passed in both directions across an edge in $\mathcal{T}$, the potential for the associated separator S is f_S.

Corollary 3.7 Let C_1 and C_2 be two cliques in $\mathcal{T}$ separated by $\mathcal{S}$. Whenever active flows have passed in both directions across the edge between C_1 and C_2, the tree is sum-consistent along S. Thus,

$$\sum_{C_2 \setminus S} \phi_{C_2} = \phi_S = \sum_{C_1 \setminus S} \phi_{C_1}.$$

Corollary 3.8 After a full schedule of flows has passed, the charge changes in the marginal charge Φ_f of f, and the system reaches equilibrium.

This Corollary is our principal results since it shows the ability of the flows propagation process to calculate margins on all cliques and separators.

Corollary 3.9 If f factorizes on $\mathcal{T}$, then Φ_f is a representation for f, and can be expressed as follows

$$f = \frac{\prod_{C \in \mathcal{C}} f_C}{\prod_{S \in \mathcal{S}} f_S}. \tag{3.6}$$

Entering and propagating evidence

The algorithm for propagating information involves two stages of collection to, and distribution from, a root-clique of a flow. Thus, an arbitrary clique $C_0 \in \mathcal{C}$, identified as *root-clique*, is selected. Active flows are initially *collected* toward C_0. Thus, the root-clique C_0 absorbs all information available and its potential becomes f_{C_0} which is a marginal representation of f. Then, all information must be passed to all remaining cliques. Hence, active flows are *distributed* from C_0 back toward the periphery. After the end of the entire process of collection and distribution, each clique has received active flows passed in both directions between every pair of cliques and the resulting charge is a sum-margin of f. This process allows to define a probabilistic network as a dynamic model.

It is formally meant by *evidence* a function $\mathcal{E} : \mathcal{X} \rightarrow \{0,1\}$, where the elements of $\mathcal{X}$ which have assigned value zero are *impossible*. It is called $\mathcal{E}$ a *finding*. The evidence function can be factorized as

$$\mathcal{E}(x) \equiv \prod_{v \in U} l_v(x_v), \tag{3.7}$$

where U is a certain set of nodes. Particularly, if the evidence is given by findings such that X_v has a definite state for each node $v \in U$, then the element $l_v(x_v)$ in (3.7) is

$$l_v(x_v) = \begin{cases} 1 & \text{if } x_v \text{ is the observed state of node } v, \\ 0 & \text{otherwise.} \end{cases}$$

For example, let X be a node with n states $(x_1, x_2, ..., x_n)$. Suppose to get the information that X can be only in state i. The elements of X are zero in all impossible states, except i where have unity value.

Initially, before the evidence is observed, the junction tree is the expression of the overall (prior) distribution of all the variables, *i.e.* it contains a their representation. When evidence is incorporated in the network, it is applied involving the potentials which are modified. These modifications are then propagated through the tree yielding the posterior probabilities. The posterior joint probability function for $\mathcal{E}$ is given by the following product

$$p(x|\mathcal{E}) \equiv kp(x)\mathcal{E}(x), \tag{3.8}$$

where $p(x)$ is the prior probability function for the network and k is a normalizing constant given by the reciprocal of the prior probability of $\mathcal{E}$. The evidence enters into the junction tree multiplying the potential ϕ_C by l_v for some arbitrary clique C containing v and for each $v \in U$. The potential $\phi_C(x_C)$ assumes value 0 when it has observed the state x_v of node v and x_C is a state other than x_v. The modified potentials now constitute a representation of $p^{\mathcal{E}}(x) \equiv p(x\&\mathcal{E}) \equiv p(x)\mathcal{E}(x) \propto p(x|\mathcal{E})$, *i.e.* the contraction of the final charge is equal the joint probability of x and the evidence. Now the passage of a full schedule of flows leads the junction tree to equilibrium and the final charge will be $(\{p^{\mathcal{E}}_C, C \in \mathcal{C}\}, \{p^{\mathcal{E}}_S, S \in \mathcal{S}\})$. Then, the posterior probabilities are obtained normalizing the potentials to sum to unity. The expressions for the joint posterior probabilities is shown below:

$$p(x\&\mathcal{E}) = \frac{\prod_{C\in\mathcal{C}} p(x_C\&\mathcal{E})}{\prod_{S\in\mathcal{S}} p(x_S\&\mathcal{E})}, \tag{3.9}$$

and

$$p(x|\mathcal{E}) = \frac{\prod_{C\in\mathcal{C}} p(x_C|\mathcal{E})}{\prod_{S\in\mathcal{S}} p(x_S|\mathcal{E})}. \tag{3.10}$$

A local application

Here it is shown the general principles of the local computation for a brief illustration applied to a junction of only two cliques. Obviously, the basic idea can be extended to sizer junction trees. Let US and SZ be two cliques separated by S in a junction tree $\mathcal{T}$, where U, S and W are discrete random variables with strictly positive joint density functions which factorize as

$$p(u,s,w) = f(u,s)\frac{1}{h(s)}k(s,w). \tag{3.11}$$

For the condition (S6) in Section 2.2, this factorization holds if and only if $U \perp\!\!\!\perp W|S$.

the marginal density $p(u,s)$ summing over w are now calculated:

$$\begin{aligned} p(u,s) = \sum_w p(u,s,w) = \\ = \sum_w f(u,s)\frac{1}{h(s)}k(s,w) = f(u,s)\frac{1}{h(s)}\sum_w k(s,w). \end{aligned} \tag{3.12}$$

If it is defined

$$h^*(s) = \sum_w k(s,w) \tag{3.13}$$

and

$$f^*(u,s) = f(u,s)\frac{h^*(s)}{h(s)}, \tag{3.14}$$

then, follows

$$f^*(u,s) = p(u,s). \tag{3.15}$$

The calculation of $p(u,s)$ can be imagined through the expressions (3.13) and (3.1.2) as the effect of passing a local flow from the clique US to SZ through the separator S. Additionally, the quantity $h^*(s)/h(s)$ is the *update ratio.*

We have, for the marginal density $p(u,s,w)$:

$$\begin{aligned} p(u,s,w) &= f(u,s)\frac{1}{h(s)}k(s,w) \\ &= f(u,s)\frac{h^*(s)}{h(s)}\frac{1}{h^*(s)}k(s,w) \\ &= f^*(u,s)\tfrac{1}{h^*(s)}k(s,w), \end{aligned} \tag{3.16}$$

by (3.1.2). Thus, the effect of the passage of the flow is a new representation for $p(u,s,w)$ as function of the marginal densities.

Now, the flow has to pass in the other direction, from SZ to US. Parallel to (3.13) it is defined

$$h'(s) = \sum_u f^*(u,s),$$

which is equal to $p(s)$ by . Similarly, parallel to we have

$$k'(s,w) = k(s,w)\frac{h'(s)}{h^*(s)} = p(s,w).$$

Finally, parallel to the overall representation involves only marginal densities.

$$p(u,s,w) = f^*(u,s)\frac{1}{h'(s)}k'(s,w),$$

that is

$$p(u,s,w) = p(u,s)\frac{1}{p(s)}p(s,w).$$

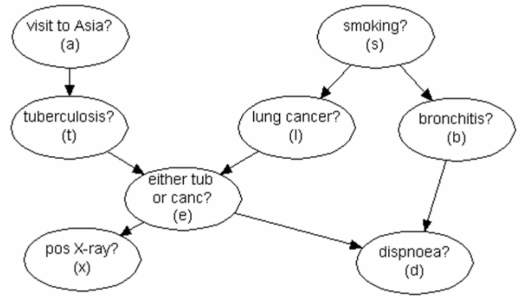


Figure 3.1: The Bayesian network representing the Asia example.

3.1.3 Propagation algorithm applied to the Asia example

In this section the propagation algorithm will be explained through a fictitious example, the Asia network, already treated by Lauritzen and Spiegelhalter (1988).

Dyspnoea is a disease that produce shortness-of-breath and can be caused by tuberculosis, or lung cancer or bronchitis, or none of them, or a combination of them. A recent visit in Asia increases the change of tuberculosis. Additionally, smoking is a risk factor for lung cancer and bronchitis. A single chest X-ray test does not discriminate between lung cancer and tuberculosis and it does not provide information about presence or absence of dyspnoea.

A casual network of this medical problem is shown in Figure 3.1. The model is a directed acyclic graph with binary variables and directed edges representing casual influences. Assuming that a patient has been recently in Asia, it is of interest in evaluating the chance that the patient has to contract any of these diseases.

The joint distribution $p(a, t, x, e, d, l, b, s)$ can be factorized as the product of the conditional distributions of each node given the parents (see Section 2.3.2), *i.e.*

$$p(a)p(s)p(t \mid a)p(e \mid t, l)p(l \mid s)p(b \mid s)p(d \mid e, b)p(x \mid e). \tag{3.17}$$

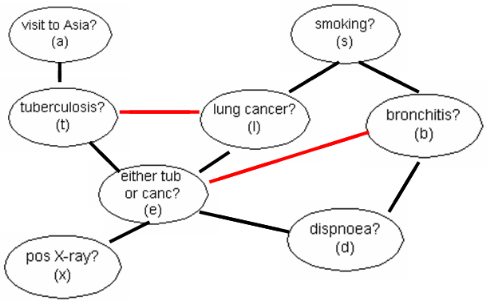


Figure 3.2: Moralization of the Asia example. Parents are married through the red edges.

The factorization is now expressed as function of *potentials* $\psi(\cdot)$:

$$\psi(a)\psi(s)\psi(t,a)\psi(e,t,l)\psi(l,s)\psi(b,s)\psi(d,e,b)\psi(x,e), \qquad (3.18)$$

where these potentials were, initially, the conditional probabilities in (3.17), i.e $\psi(a) = p(a)$, $\psi(e,t,l) = p(e \mid t,l)$ etc. for all variables.

The undirected form of the graph is now considered in order to keep track of the groups of variables entering into the potentials ψ. Thus, the corresponding moral graph of the Asia example obtained dropping the directions and marrying parents is shown in Figure 3.2.

Note that the factorization in (3.18) involves several expressions that are function of the cliques in the moral graph.

Such moral graph is not triangulated since there are cycles of length 4 or more without a chord, e.g. the cycle involving the nodes (s,l,e,b). Therefore, it need to be made chordal in order to construct a junction tree. Thus, a chord is added between the nodes l and b, as shown in Figure 3.3.

If the potentials ψ defined on the cliques of the filled-in graph are considered, the joint distribution becomes

$$\psi(a,t)\psi(e,t,l)\psi(s,l,b)\psi(e,l,b)\psi(d,e,b)\psi(x,e), \qquad (3.19)$$

where the functions ψ are obtained by matching the terms in (3.17). For example, $\psi(a,t) = p(a)p(t \mid a)$, $\psi(e,t,l) = p(e \mid t,l)$, $\psi(s,l,b) = p(l \mid s)p(b \mid s)p(s)$, etc. Thus, this expression can be reduced to the product of the

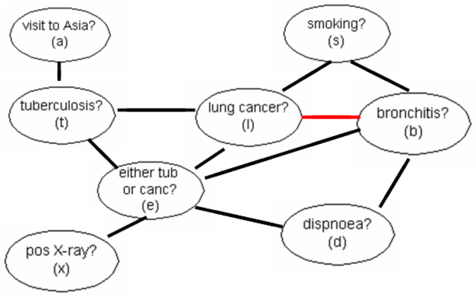


Figure 3.3: Triangulated version of the Asia example. In this case we could either add an edge between the nodes s and e, or between the nodes l and b to obtain a triangulated graph.

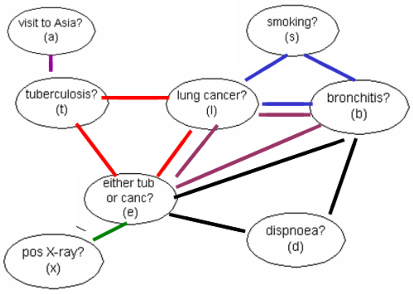


Figure 3.4: Cliques in the triangulated graph of the Asia example.

potentials defined on the cliques of the graph shown in Figure 3.4:

$$p(x) = \prod_{cliquesC} \psi_C.$$

The maximum cardinality search applied to the Asia example gives the initial ordering shown in Figure 3.5, which corresponds to a junction tree involving cliques and separators as reported in Figure 3.6.

For each network many different junction trees can be obtained, depending on the choice of the elimination order. There are $N!$ possible elimination sequences, where N is the total number of variables in the network. An efficient junction tree has small clique tables and few cliques in order to have the minimum total clique size table. The clique size table depends on both the number of variables in the clique and the number of states for each variable in the table. Thus, the size table t of an individual clique C is given by the product of the number of the states in each variable, *i.e.* $t_C = n_1 n_2 ... n_{N_C}$,

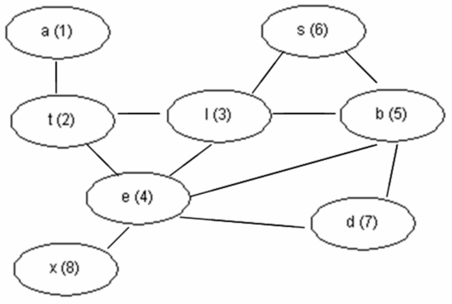


Figure 3.5: A possible initial ordering of the Asia example.

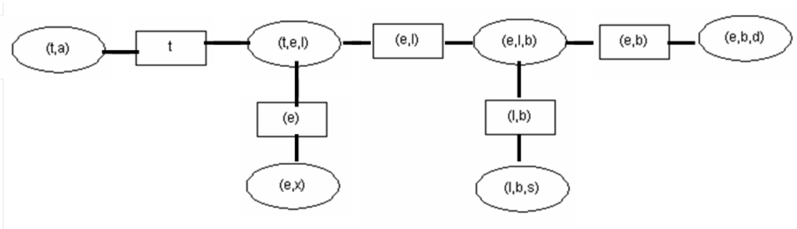


Figure 3.6: Asia net junction tree.

where t_C is the size table of the clique C, n_i, for $i = 1, 2, ..., N_C$, is the number of states in each variable in C, and N_C is the total number of variables in C.

A factorization over $\mathcal{T}$ of the joint distribution P allows to express P as a function of the individual marginal distributions of the cliques and separators, *i.e.*

$$\frac{p(a,t)p(t,l,e)p(l,e,b)p(l,b,s)p(e,b,d)p(e,x)}{p(t)p(l,e)p(l,b)p(e,b)p(e)} \tag{3.20}$$

The running intersection property ensures that the joint probability can also be expressed as

$$p(a,t)p(l,e \mid t)p(b \mid l,e)p(s \mid l,b)p(d \mid e,b)p(x \mid e). \tag{3.21}$$

This expression also can be obtained from (3.20), being $p(l,e \mid t) =p(t,l,e)/p(t)$, $p(b \mid l,e) = p(l,e,b)/p(l,e)$ etc. Here, p is simply a product of functions on cliques and hence (3.21) is yet another potential representation.

Generally, the equation (3.21) can be written as

i	Cliques	Residuals	Separators
	C_i	R_i	S_i
1	a, t	a, t	$\emptyset$
2	t, l, e	l, e	t
3	l, e, b	b	l, e
4	l, b, s	s	l, b
5	e, b, d	d	e, b
6	e, x	x	e

Table 3.1: Cliques, residuals and separators of the graph in Figure 3.1

$$\prod_{i=1}^{6} p(R_i \mid S_i)$$

where R_i are the residuals C_i/S_i, S_i are the separators and C_i are the cliques. Table 3.1 shows the cliques, the residuals and the separators. Each term in (3.21) can be obtained as

$$p(R_i \mid S_i) = \psi(C_i)/\sum_{R_i} \psi(C_i)$$

For example, the final term $p(R_6 \mid S_6)$ is equal to

$$\psi(x, e)/\sum_{x} \psi(x, e).$$

Then, given the representation in (3.21), the marginal cliques can be derived multiplying $p(R_i \mid S_i)p(S_i)$, where $p(S_i)$ is defined by marginalization from the previous calculated clique marginal. For example, from $p(C_1) = p(a, t) = p(t \mid a)p(a)$ we calculate $p(S_2) = p(t)$ by marginalization; from $p(C_2) = p(R_2 \mid S_2)p(S_2) = p(l, e \mid t)p(t)$ we obtain $p(S_3) = p(l, e)$ by marginalization, etc. A condition required for the separators is to be consistent, *i.e.* if C_1 and C_2 are two cliques separated by S, the marginal distributions for S is the same independently from the clique (either C_1 or C_2) performing the marginalization. The process to find the marginal distributions is called initialization of the junction tree.

Suppose now to observe that a patient visited Asia. The evidence on node

a is propagated throughout the junction tree until all cliques are updated. Firstly, the clique (a, t) is updated in the following way

$$p^*(a,t) = p(a,t)\frac{p^*(a)}{p(a)},$$

where p^* is the revised distribution after observing evidence; then, the message passes to the children clique (t, l, e) through the separator t as follows

$$p^*(t,l,e) = p(t,l,e)\frac{p^*(t)}{p(t)}.$$

Following a similar argument the evidence is propagated throughout the junction tree. The factors $p^*(a)/p(a)$ and $p^*(t)/p(t)$ are the update ratios.

Thus, each parent clique in the network passes its message to its children multiplying each term in the marginal distribution of the child by the update ratio between the new and the old probability. This passage requires the identification of a root-clique that initially collects evidence and then distributes it back toward the periphery.

Chapter 4

Genetic Background

In this section we give a general background on Genetics. In particular, in § 4.1 we give some notions on DNA biology. In particular, we see basic DNA principles, some details on the DNA structure, we introduced the definitions of chromosome and gene and we introduce the DNA markers nomenclature. In § 4.2 we threat the amplification process that is a technic to amplify and replicate a piece of DNA in order to analyse it. We also introduce the short tandem repeat markers which are the most common markers used in literature. Finally, in the last section 4.3 we explain the definitions of drop-out alleles and stutter which are artefacts that can occur during the amplification process.

4.1 DNA Biology

4.1.1 Basic DNA principles

The structural and functional unit of all living organisms is the *cell*. It is the smallest unit of an organism classified as living. Organisms can be *unicellular* if they consist of a single cell, *e.g.* bacteria, or *multicellular*, such as humans (an average human being is composed of approximately 100 trillion cells). The cell can be compared to a factory that produces energy using as resource thousands of different proteins called *enzymes*. All cells come from preexisting cells. The nucleus of the cell contains a chemical substance, the *deoxyribonucleic acid* (*DNA*), that include the instructions which are a code for replicating the cell and constructing the needed enzymes. The DNA located in the cell nucleus of the organisms is called *nuclear DNA*, but some minor DNA can house in human *mitochondria*, termed *mitochondrial DNA*, where a mitochondrion is a membrane-enclosed organelle that is found in most eukaryotic cells [1].

Furthermore, DNA provides hereditary information that specifies physical characteristics and other genetic attributes of the organism. Thus, DNA is material that governs inheritance since it carries on information from generation to generation. The whole hereditary information, *i.e.* the entire DNA in a cell, is referred to as the *genome* of the organism.

Thus, two are the main aims of DNA: (i) to pass instruction for replicating the cell and make enzymes; and (ii) to make copies of itself in order to pass down the organism's genetic information to future generations with one-half of a person's DNA information coming from their mother and on-half coming from their father.

4.1.2 DNA structure

DNA is a *nucleic acid* which is located and produced in the nucleus of the cell and need to preserve and to pass genetic information. Nucleic acids are composed of nucleotide units that are made up of a *nucleobase* (or base), a *sugar* and a *phosphate* (see Figure 4.1). Nucleobases represent the alphabet of the cell's genetic information and they are four: *A* (*adenine*), *T* (*thymine*), *C* (*cytosine*) and *G* (*guanine*). The combination of the nucleobases forms a *nucleotide* and defines a specific biological feature. Thus, nucleotides produce the diverse biological differences among living creatures. There are approximately three billion nucleotide positions in the human genomic DNA.

Phospate and sugar form the backbone structure of the DNA molecule, whilst nucleobase discerns nucleotide unit. The sugar in DNA is *2-deoxyribose*,

[1] Eukaryotics are organisms whose cells are Eukaryotic, *i.e.* they have a nucleus isolated by a nuclear envelop.

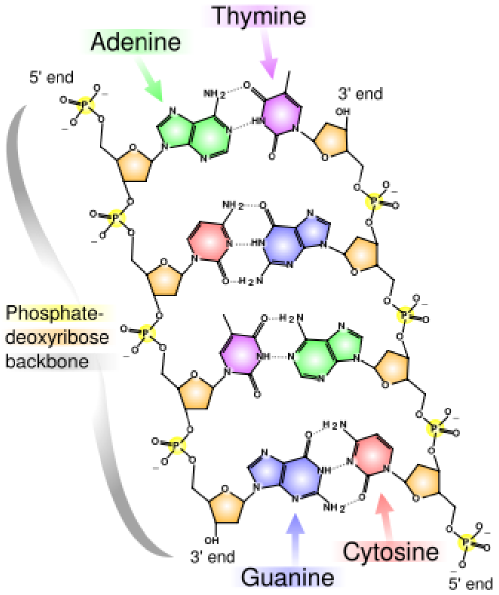


Figure 4.1: DNA chemical structure. Image from http://en.wikipedia.org/wiki/Dna.

which is a pentose (five carbon) sugar. Sugars are joined to phosphate groups through the third and fifth carbon atoms of adjacent sugar rings, referred to as the 5' (*five prime*) and 3' (*three prime*) ends.

In the cell, DNA is composed of two *strands* linked together through a *hybridization* process. Thus, each individual nucleotide matches up with a *complementary base* through a *hydrogen bound* between the bases and following a specific pairing rule such that adenine can only hybridize to thymine and cytosine can only pairs up with guanine (see Figure 4.2). Actually, since guanine and cytosine are paired up each other through three hydrogen bounds, whilst there are two hydrogen bounds between adenine and thymine, CG bound is stronger than AT base pair. Thus, knowing the sequence of one DNA strand, it is straigthforward to determine the complementary sequence. As shown in Figure 4.2, the two DNA strands are connected in the shape of a double helix structure that is a right-handed spiral. The two strands of DNA are *anti-parallel*, *i.e.* direction of the nucleotides in one strand is opposite to orentation in the other strand.

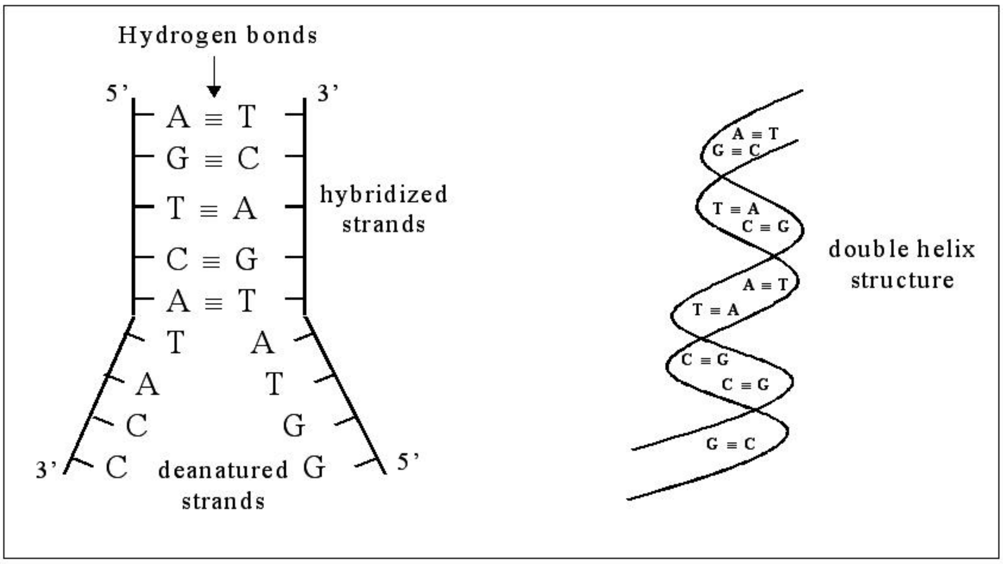


Figure 4.2: Representation of DNA strands forming a double helix structure. Image from Butler, 2005.

4.1.3 Chromosomes

Nuclear DNA is packaged with *proteins* into automosomal *chromosemes.* In the human genome, there are 46 different chromosemes in 23 pairs where the 23nd pair are the chromosemes X and Y indicating the sex of the individual. Females are identified by the couple XX containing two copies of the X chromosemes, whilst males are identified by the pair XY since they contain a single copy of both X and Y types of chromosemes. In each pair one choromosome is inherited from mother and one from father, but it is not possible to distinguish which is which, with the exception that a Y chromosome must have come from a male individual, hence the father, and the X of a male must have come from his mother (*Mendelian segregation*).

Each chromosome contains a *centromere* which is a specific region that holds together the two similar halves of the chromosome, termed the sister *chromatids.* It is the strongest and thinnest region in the middle of the chromosome. Since centromere is always off center it yields the *short arm* and the *long arm* of the chromosome (see Figure 4.3).

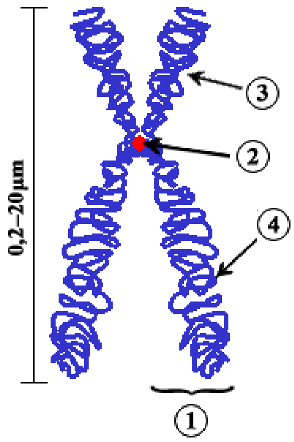


Figure 4.3: Representation of a chromosome. (1) Chromatid: one of the two identical parts of the chromosome. (2) Centromere: the point where the two chromatids touch. (3) Short arm. (4) Long arm. Image from http://en.wikipedia.org/wiki/chromosome.

Choromosemes with same size and that have same genetic structure are said *homologous*. Cells that contain a pair of homologous chromosomes are called *diploids*, *haploid cells* have a single copy (*e.g.* the sex cell sperm and ova), whilst *polyploid cells* have more copies, such as liver cells. The sequence of DNA in the homologous pair is the same except if mutations occur. A chromosomale pair is derived by each parent at the time of conception, when an egg cell combines with a sperm cell giving life a zygote that is a diploid cell.

4.1.4 Genes

The DNA is divided into *coding* and *non-coding* regions. The coding regions are referred to as *genes*, whilst a non-coding region is referred to as a *locus*. They code proteins. Thus, genes are composed of *exons*, *i.e.* protein-coding portions, and *introns*, *i.e.* the intervening sequences. A size gene ranges from a few thousand to tens thousands of base pairs. Each gene has a copy of its in the homologous chromosome at same locus, or position.

Gene expression is termed *allele*. For example, a gene that represents the genetic information ”eyes colour” has two alleles that definde light or dark colour. A pair of alleles on homologous chromosomes forms a *genotype*. For example, it is supposed that the alleles at a locus are `A` and `a`, then the possible genotypes are: `AA`, `Aa`, `aa`. The `AA` genotype is termed *homozygous*, since the two alleles are identical at a specific genetic locus on a homologous chromosomal pair, whilst the `Aa` genotype is termed *heterozygous*, since they are different. The capital letter indicates the dominant allele, whilst the small letter indicates the recessive allele. Alleles are generally represented as positive integers indicating the times that a certain word, given by a particular sequence of the four bases represented by the letters A,T,C and G, is repeated. It is defined *marker* a specific locus where alleles are amplified.

The combination of genotypes for multiple loci forms a DNA *profile*. Thus, an individual's DNA profile consists of measurement on a number of markers, each comprising a genotype represented by an unordered pair of alleles. In human identity tests or mixture tests multiple loci are examined in order to reduce errors in identification deriving by random matches between individuals which actually are unrelated.

4.1.5 DNA markers nomenclature

The nomenclature for DNA markers is straightforward to use. Now, we distinguish between DNA markers that fall within a gene and those that fall outside. Markers that are part of a gene use the gene name for their designation. For example, consider the short tandem repeat (STR) marker `TH01`. The letters TH are the initial letters of the gene name **t**yrosine **h**ydroxylase, whilst the number sequence '01' is the number of the intron of the gene where the repeat region is located. It is possible to add the prefix HUM- at the beginning of the marker name if we are interested in indicating that the marker is from the human genome.

For DNA markers falling outside of gene regions, the chromosomal position characterizes the name. For example, consider the STR loci `D7S820` and `DYS393` (see § 4.2.2). The letter 'D' at the beginning of the name stands for DNA; the next character is referred to respectively the chromosome number and the Y chromosome. The letter 'S' indicates that the DNA is a single copy sequence. Finally, the last numbers are the order in which the markers has been discovered and categorized for a particular chromosome.

4.2 DNA amplification and STR markers

4.2.1 Polymerase chain reaction (PCR) process

Techniques regarding DNA amplification, such as polymerase chain reaction (PCR), has been developed in 1983 by Kary Mullis and members of Human Genetics group at the Cetus Corporation (now Roche Molecular Systems). Such techniques revolutioned molecular biology so that Kary Mullis received the Nobel Prize in 1993.

PCR derives its name from one of its key components, a *DNA polymerase* that is an enzyme used to amplify (*i.e.*, replicate) a piece of DNA. This process allows to make millions of copies of a specific sequence of DNA that is replicated over and over again. The ability of PCR to make copies of DNA sequence is important especially for forensic science where DNA samples are often limited in both quantity and quality and otherwise, without this new technology, samples would be impossible to analyse. In effect, PCR can be used to analyze extremely small amounts of sample amplifying a single or few copies of a piece of DNA across a number of orders of magnitude, generating millions or more copies of the DNA piece. When polymerase chain reaction permits simultaneous amplification of more than one regions of DNA, PCR is said *multiplex*.

Polymerase chain reaction process involves heating and cooling samples that are subject to over 30 thermal cycles. During each cycle, a copy of the target DNA sequence is generated for every molecule. Thus, a billion copies are generated after 30 cycles.

A DNA signature is represented as an *electropherogram* (EPG) that measures responses in *relative fluorescence units* (RFU). The alleles in the electropherogram are represented with peaks that have a specific height and area around each allele. An example of electropherogram is shown in Figure 4.4 where the alleles with *repeat number* 11 and 12 for marker D5 of a DNA

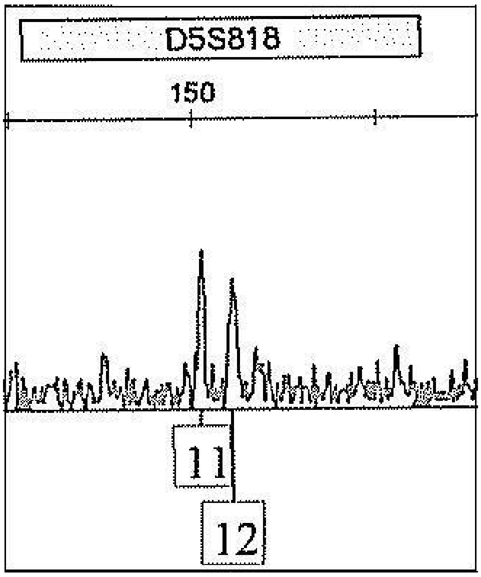


Figure 4.4: Electropherogram for a DNA profile for marker D5. The alleles have *repeat numbers* 11 and 12.

profile are amplified.

4.2.2 Short tandem repeat analysis

Eukaryotic genomes have a great number of repeated DNA sequences and they differentiate for the length of the core repeat unit and the number of contiguous repeat units or the overall length of the repeat region. Regions with this high number of repeated DNA sequences are said *satellite* DNA. Repetitions of a short DNA sequence tend to produce a different frequency of the nucleotides adenine, cytosine, guanine and thymine, and thus have a different density from bulk DNA, such that they form a second (or satellite) band when genomic DNA is separated on a density gradient. Regions with a medium lenght repeat, approximately 10-100 bases (bp) in length, are termed *minisatellite* or a VNRT (variant number of tandem repeats). The shortest DNA regions (2-6 basees in length) are those called *microsatellites*, simple sequence repeats (SSRs), or short tandem repeats (STRs) (see Figure 4.5).

STRs are the most common DNA repeat markers used in forensic science

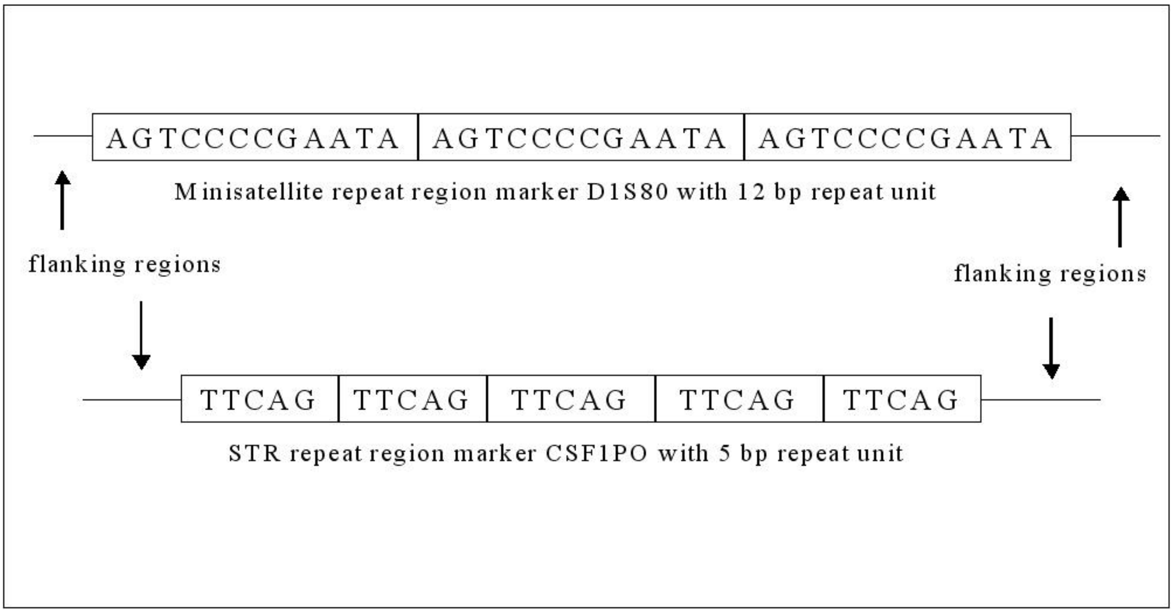

AGTCCCCGAATA
AGTCCCCGAATA
AGTCCCCGAATA
Minisatellite repeat region marker D1S80 with 12 bp repeat unit
flanking regions
flanking regions
TTCAG
TTCAG
TTCAG
TTCAG
TTCAG
STR repeat region marker CSF1PO with 5 bp repeat unit

Figure 4.5: Repeat unit structure of minisatellite and microsatellite DNA markers.

due to the fact that they can be easily amplified through the PCR tecnique since the repeat size of both alleles from an hetetozygous individual are small and so similar.

The analysis is performed by extracting nuclear DNA from the cells of a forensic sample of interest, then flanking regions, *i.e.* the regions that surround the repeats, are determined and specific polymorphic regions of the extracted DNA are amplified by means of the polymerase chain reaction.

STR sequences differentiate each other for more factors. An element is the length of the repeat unit that gives the name to the repeat sequence. Thus, if a sequence is composed of two nucleotides repeated, then this sequence is said dinucleotide; if it is composed of three nucleotides repeated, then it is said trinucleotide; if four, tetranucleotide; if five, pentanucleotide; and if six, hexanucleotide. Now, for mono-, di-, tetra-, penta-, and hexanucleotide repeats the possible motifs are respectively 4, 16, 64, 256, 1024, 4096. For example, for mononucleotide repeats they are: A, C, G, T; for dinucleotide repeats the possible motifs are: AC, AG, AT, CG, CT, GT, AA, CC, GG, TT, CA, GA, GC, TA, TC, and TG. Actually, microsatellites are tandemly repeated, thus some motifs are equivalent to others. As a consequence, the

AGGAG	AGGAG	AGGAG	AGGAG	AGGAG	AGGAG	AGGAG	AGG

Figure 4.6: Microvariant allele having repeat number 7.4. The allele contains seven pentanucleotide repeats and one incomplete tetranucleotide with the third missing repeat at a single guanine of the normal AGGAG repeat unit

possible motifs become 2, 4, 10, 33, 102, and 350 for mono-, di-, tetra-, penta-, and hexanucleotide repeats. Thus, for example, for dinucleotide repeats became AC, AG, AT, CG.

Another element to discriminate a STR sequence is the repeat pattern. For this reason, they are divided into a number of categories based on the repeat pattern: *simple repeats* containing units of identical length and sequence; *compound repeats* containig two or more adjacent simple repeats; *complex repeats* where unit length and intervening sequences are variable.

In a number of cases alleles in STR locus can contain incomplete repeat units. These are called *microvariants*; for example, an allele 7.4 at a certain STR locus contains seven pentanucleotide repeats and one incomplete tetranucleotide with the third missing repeat at a single guanine of the normal AGGAG repeat unit (see Figure 4.6).

STR markers are the most popular for forensic DNA typing (particularly, among different types of STRs, tetranucleotides repeats are the most used comparing with di- or trinucleotides; whilst, in the human genome, penta- and hexanicleotides are less common but however examined in a number of laboratories) since they are robust enough to survive in conditions of low-quantity or degraded DNA. In effect PCR amplification of degraded DNA work better with smaller product sizes. A forensic DNA laboratory often has to deal with DNA samples that have been found in critical conditions. For example, think of a crime where the biological material has been left exposed to environmental factors for days, or the retrieved biological sample has been found in limited quantity. DNA molecules are degraded by environmental exposure that breaks the molecules randomly into smaller pieces. Particularly, the materials that damage DNA are water and enzymes called *nucleases.*

There is an inverse relation between the size of the locus and successful PCR amplification from degraded DNA. This is due to the fact that, since STR loci are amplified with small product sizes, intact DNA strands are easier to be found. Furthermore, the narrow size range of STR alleles decreases the chances of drop-out (see § 4.3).

4.3 Alleles drop-out and stutter

During PCR amplification of STR alleles a number of artefacts can occur interfering with interpretation and genotyping of the alleles in the amplified DNA. In this chapter we investigate artefacts represented by *allele drop-out* and *stutter*.

Allelic drop-out are due to equipment failure when the low DNA level is insufficiently amplified to give a detectable signal. This is often due to reduced quantities of DNA, so that they are not detectable. In particular, they occur especially in presence of extremely unbalanced contributions to the mixture. For example, suppose that the genotype of an individual is represented by the alleles with repeat number $\{10, 11\}$ for a certain marker, whilst suppose to observe in the amplification process the allele 10 only. In this scenario, the allele 11, present in the genotype of the individual, is not observed since it is a drop-out allele. Similarly, suppose to observe a 2-person mixture where the DNA proportions are $15 : 1$, *i.e.* 15 parts of DNA come from a contributor and 1 part comes from the other. Moreover, suppose to observe in the mixture the alleles with repeat number $\{8, 9, 10\}$ and that profiles of the contributors are $\{8, 9\}$ and $\{10, 13\}$. In this scenario, the allele 13, present in the genotype of the second contributor, is not observed in the mixture since it is a drop-out allele.

Other frequent artifacts are stutters. These are due to a slippage of the DNA during the replication process. They are spurious products with extremely small peaks and they contain one repeat unit less than the corresponding main allele peak.

Chapter 5

The experimental design procedure

In this chapter we describe the DNA mixtures that have been analysed. We used two different groups of data. The first one, analysed in chapters 7 and 9, are DNA mixtures produced in the laboratory and provided by *Capitano Gianpietro Lago* and *Tenente Elena Salata* from *Ra.C.I.S.* (Raggruppamento Carabinieri Investigazioni Scientifiche). The second one, analysed in chapter 8, are DNA mixtures produced in the laboratory provided by *FSS* (Forensic Science Service) in London.

The data provided from *Ra.C.I.S.* are mixed blood samples prepared in known proportions and made up of two or three individuals termed X, Y and Z. X and Y are two male individuals, whilst Z is a female. Mixtures are termed *mix-A*, *mix-B*,..., *mix-Q*. For each DNA extraction two kind of amplifications have been performed: the first one employs the kit Identifiler$^{\mathrm{TM}}$ of Applera, the other one employs the kit PowerPlex16$^{\mathrm{TM}}$ of Promega. Table 5.1 shows, for each mixture, the contributors to the mixture and the DNA proportions of each contributor. The mixed traces *A*-*F* are 2-person mixtures, and the contributors are the individuals X and Y, whilst the mixed traces *G*-*Q* are 3-person mixtures.

The data provided from *FSS* are mixed blood samples prepared in known proportions and made up of two individuals. They are six mixtures termed *mix-1*, *mix-2*, ..., *mix-6*.

The contributors to the mixtures have been called using the alphabetic

Mixture	Contributors		
	X	Y	Z
A	1	1	0
B	1	2	0
C	1	5	0
D	1	10	0
E	1	20	0
F	1	40	0
G	1	1	1
H	1	2	1
I	1	2	2
L	1	5	1
M	1	5	2
N	1	5	5
O	1	10	1
P	1	10	5
Q	1	10	10

Table 5.1: *Lago* data, DNA proportions for each contributor.

Mixture	Contributors	Common contributor
1	A-B	A B
2	C-D	D
3	E-A	A E
4	F-D	D
5	E-G	E
6	B-H	B

Table 5.2: *FSS* data, contributors to the mixtures.

letters. Table 5.2 in the second shows the contributors for each mixture. Thus, for example, *mix-1* is made up from the contributors A and B, *mix-2* from the contributors C and D, etc. Some of these mixtures have an individual in common, *i.e.* an individual is present in both mixtures. The third column of the table displays the common contributor between the mixtures. Thus, for example, the individual A is present in both *mix-1* and *mix-3*, the individual B is present in both *mix-1* and *mix-6*, the individual D is present in both *mix-2* and *mix-4*, etc. Each mixture has been realized in 7 ways, each one with different ratios of DNA from contributors shown in Table 5.3. Thus, for example, *mix-1* was available with proportions 1 : 2,

Proportions for each mixture						
1:1	1:2	1:5	1:10	2:1	5:1	10:1

Table 5.3: *FSS* data, DNA proportions for each contributor.

i.e. the DNA proportion of contributor A is 1 and the DNA proportion of contributor B is 2, but also with proportions 2 : 1, *i.e.* the DNA proportion of contributor A is 2 and the DNA proportion of contributor B is 1. Similarly for the other five mixtures.

A DNA signature is represented as an *electropherogram* (EPG) that measures responses in *relative fluorescence units* (RFU). The alleles in the mixture are represented with peaks that have a specific height and area around each allele. An example of an electropherogram is shown in Figure 5.1 where the alleles for marker D8 of *mix-D* are amplified. The alleles have *repeat number* 10, 12 and 14, and *peak area*, respectively, 19481, 2118 and 16979. It is worth noting that, since there are three alleles, it is a mixture made up of at least two contributors. In effect, since each individual has at most two alleles, the amplification of three alleles in marker D8 allows to conclude that there must have been at least two contributors to the trace.

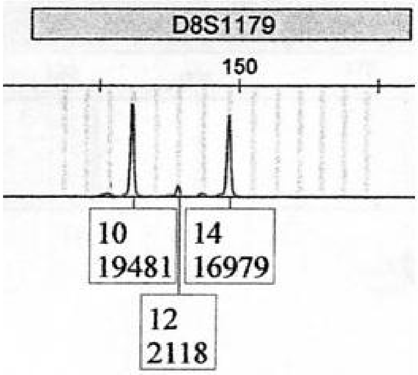


Figure 5.1: Mix-D, marker D8.

5.1 Data for 2-person mixtures

In chapter 7 we analyse mixed DNA traces involving two contributors only. Table 5.4 shows the data used in the analysis in chapter 7. They correspond to *mix-D* amplified employing the kit Identifiler$^{\text{TM}}$of Applera.

In Table 5.4 column "Mixture" shows the alleles observed in the mixture; columns "Peak Area" and "Rel. Weight" show, respectively, the measured peak areas and the relative weights[1]; finally, columns "Suspect" and "Victim" show the genotypes of two identified individuals, termed victim v, and suspect s. Here victim and suspect correspond respectively to individuals X and Y. For the analysis in chapter 7 we used 7 markers, which are Amelogenin, D5, D7, D8, D16, D18 and D21 (see Figure 5.2).

We note that in Table 5.4 for marker vWA the allele with repeat number 18 possessed by the suspect is not observed in the mixture. This is due to the

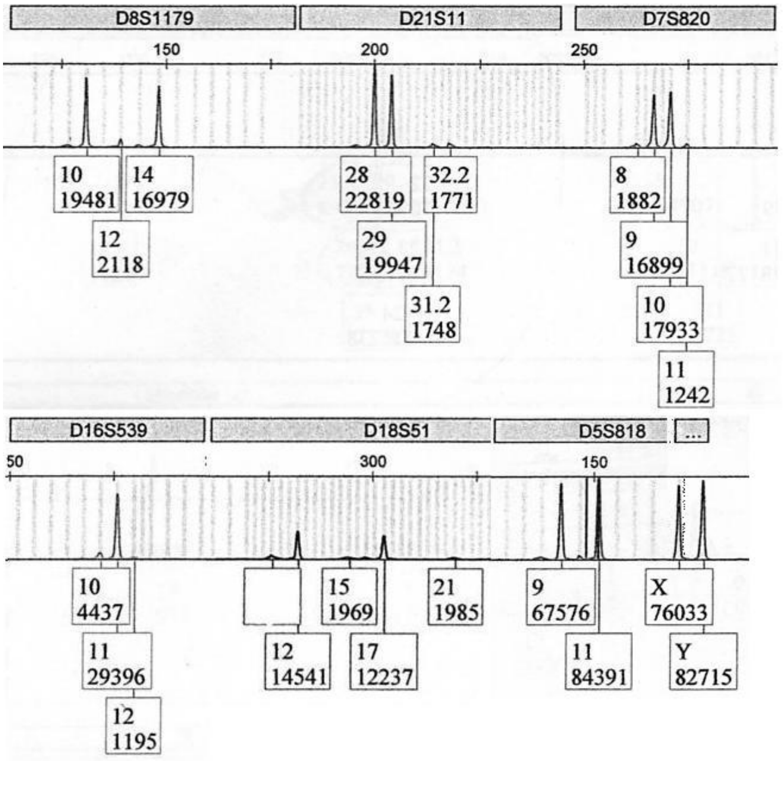


Figure 5.2: Mix-E, markers Amelogenin, D5, D7, D8, D16, D18 and D21.

[1] Details on relative weights are given in chapter 7

Marker	Mixture	Peak Area	Rel. Weight	Suspect	Victim
Amelogenin	X	22328	0.5092	X	X
	Y	21520	0.4908	Y	Y
D2	19	1021	0.0311		19
	20	970	0.0311		20
	24	24390	0.9378	24	
D3	15	21075	0.4350	15	15
	17	1662	0.0389		17
	18	2176	0.0569		
	19	17951	0.0493	19	
D5	9	23749	0.4217	9	
	11	28177	0.5783	11	11
D7	8	1882	0.0418		8
	9	16899	0.4223	9	
	10	17933	0.4979	10	
	11	1242	0.0379		11
D8	10	19481	0.4254	10	
	12	2118	0.0555		12
	14	16979	0.5191	14	14
D13	8	33963	0.4002	8	
	11	29484	0.4777	11	
	12	2552	0.0451		12
	14	3734	0.0770		14
D16	10	4437	0.1161		10
	11	29396	0.8463	11	
	12	1195	0.0375		12
D18	12	14541	0.3846	12	
	15	1969	0.0651		15
	17	12237	0.4585	17	
	21	1985	0.0919		21
D21	28	22819	4808	28	
	29	19947	0.4353	29	
	31.2	1748	0.0410		31.2
	32.2	1771	0.0429		32.2

fact that this is a drop-out allele (see § 4.3). Additionally, the alleles with repeat number 18 in marker D3 and 15 in marker vWA are observed in the mixture but they are not possessed by the identified individual. In effect, if

Marker	Mixture	Peak Area	Rel. Weight	Suspect	Victim
CSF	10	1438	0.3970		10
	11	2065	0.0627		11
	12	12093	0.4004	12	
	13	13866	0.4973	13	
FGA	22	3072	0.0794		22
	23	17131	0.4628	23	
	24	16238	0.4578	24	24
THO1	6	23512	0.8747	6	
	8	2525	0.1253		8
TPOX	8	11224	0.5049	8	8
	10	8806	0.4951	10	
vWA	14	1720	0.0308		14
	15	3074	0.0590		
	16	4878	0.0998		16
	17	37291	0.8105	17	
				18	

Table 5.4: *Lago* data, 2-person mixture - a two individuals mixture composition with relative peak areas, relative peak weights, suspect's and victim's genotypes.

we look at Figure 5.3 that shows markers D3 and vWA, we note that these alleles are stutters (see § 4.3).

Table 5.5 shows the population gene frequencies[2] referred to alleles in the mixtures in Table 5.4.

5.2 Data for two traces 2-person mixtures

In section 8 we analyse two mixed DNA traces involving two contributors only. Table 5.6 shows the data used in the analysis in chapter 8. They correspond to *mix-1* (with A and B as contributors) and *mix-6* (with B and H as contributors), therefore sharing the contributor B. Furthermore, we analysed *mix-1* made up of 1 part of DNA coming from A and 5 parts from B, whilst *mix-6* has been chosen as made up of 10 parts of DNA from and 1 from H.

[2] Population gene frequencies used in this thesis have been provided from *Ra.C.I.S.*

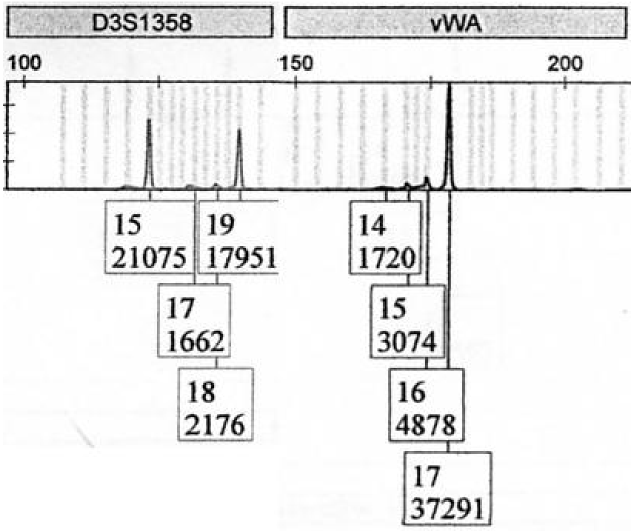


Figure 5.3: Mix-E, markers D3 and vWA with stutters 18 in D3 and 15 in vWA.

Marker	Allele	Frequencies
D5	9	0.041
	11	0.393
D7	8	0.164
	9	0.176
	10	0.272
	11	0.180
D8	10	0.097
	12	0.1404
	14	0.2135
D16	10	0.056
	11	0.319
	12	0.302
D18	12	0.139
	15	0.136
	17	0.123
	21	0.012

Table 5.5: *Lago* data, 2-person mixture - Population alleles frequencies.

In Table 5.6 In Table 5.6 column "Trace1" and "Trace2" show the alleles observed in the first and second mixture, respectively; columns "Rel. Weight" show, the measured relative weights in both traces; finally, columns

“Suspect1” and “Suspect2” show the genotypes of two identified individuals, suppose two suspects, *s1* and *s2*. Here suspect1 and suspect2 correspond respectively to the individuals B and H. For the analysis in chapter 8 we use 6 markers, which are Amelogenin, D2, D21, FGA, THO1 and vWA.

Table 5.7 shows the population gene frequencies referred to alleles in the mixtures in Table 5.6.

5.3 Data for 3-person mixtures

In chapter 9 we analyse mixed DNA traces involving three contributors. Table 5.8 shows the data used in the analyses in chapter 9. There we analyse *mix-M* amplified employing the kit PowerPlex16$^{\mathrm{TM}}$of Promega.

In Table 5.8 column “Mixture” shows the alleles observed in the mixture; columns “Peak Area” and “Rel. Weight” show, respectively, the measured peak areas and the relative weights for all markers; columns “Suspect1”, “Suspect2” and “Victim” show the genotypes of three identified individuals, for example two suspects, *s1* and *s2*, and one victim v. Here suspects and victim correspond respectively to the individuals Y, X and Z. It is worth noting that, since there are five alleles for marker Penta E and vWA, this is a mixture made up of at least three contributors. For the analysis in chapter 9 we use 4 markers only, which are Amelogenin, D7, D8 and D21 (see Figure 5.4).

In this table we note that in the markers D3 and vWA, shown in Figure 5.5, the alleles with repeat number 14 for D3 and 15 for vWA are observed in the mixture but not in the genotypes of the three identified individuals. This is due to the fact that they are stutters (see § 4.3). Table 5.9 shows the population gene frequencies of the alleles in the mixtures for a subset of markers in Table 5.8.

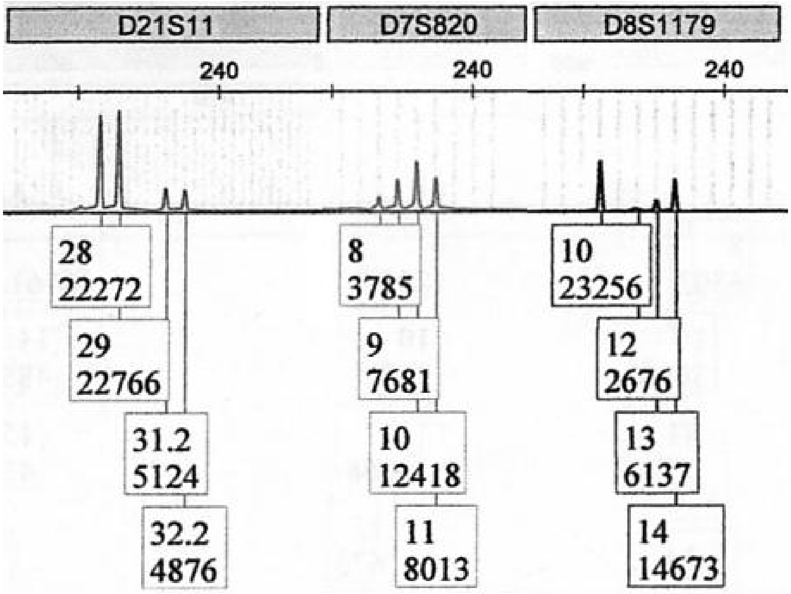


Figure 5.4: Mix-M, markers Amelogenin, D7, D8, and D21.

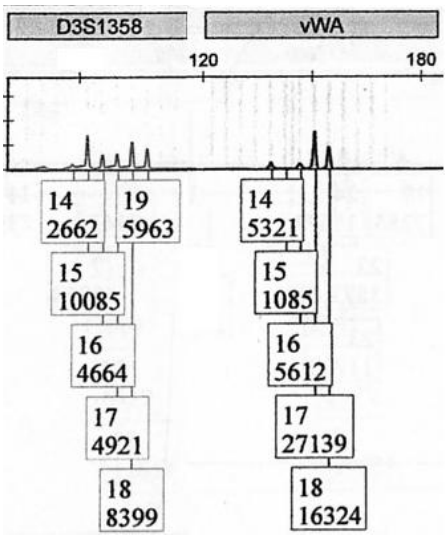


Figure 5.5: Mix-M, markers D3 and vWA with stutters 14 in D3 and 15 in vWA.

Marker	Trace1	Rel. Weight1	Trace2	Rel. Weight2	Suspect1	Suspect2
Amelogenin	X	0.6147	X	0.4950	X	X
	Y	0.3853	Y	0.5050	Y	Y
D2	19	0.5112	19	0.4338	19	
	20	0.3792	20	0.4949	20	20
	21	0.0486				
	23	0.0610	23	0.0712		23
D3	14	0.0802	14	0.1226		14
			15	0.1168		15
	18	0.9198	18	0.7607	18	
D8			11	0.1185		11
	12	0.4305	12	0.3526	12	
	15	0.5695	15	0.5289	15	15
D16	9	0.4320	9	0.4479	9	9
	12	0.5680	12	0.4250	12	
			13	0.1271		13
D18	11	0.3278	11	0.3840	11	
	12	0.1066	12	0.1107		12
	14	0.4312	14	0.4287	14	
	15	0.1343				
			21	0.0766		21
D19			12	0.1057		12
	13	0.1629	13	0.3995	13	
	14	0.1629				
	15	0.4216	15	0.3934	15	
			16	0.1014		16
D21	28	0.5017	28	0.5163	28	28
	30	0.4983	30	0.4152	30	
			32.2	0.0685		32.2
FGA	22	0.3963	22	0.5791	22	22
	23	0.6037	23	0.4209	23	
THO1	9.3	1	9.3	1	9.3	9.3
vWA	14	0.4918	14	0.3801	14	
	18	0.0885	18	0.1164		18
	19	0.4197	19	0.5035	19	19

Table 5.6: *Lago* data, two traces 2-person mixtures - two 2-individuals mixture compositions with relative peak weights, suspect1's and suspect2's genotypes.

Marker	**Allele**	**Frequencies**
D2	19	0.1375
	20	0.1461
	21	0.0258
	23	0.1146
D21	28	0.167
	30	0.252
	32.2	0.072
FGA	22	0.1691
	23	0.1519
THO1	9.3	0.2908
vWA	14	0.0831
	18	0.2249
	19	0.0831

Table 5.7: *Lago* data, two traces for 2-person mixtures - Population alleles frequencies.

Marker	**Mixture**	**Rel.**	**Rel.**	**Suspect1**	**Suspect2**	**Victim**
		Area	**Weight**			
Amelogenin	X	44748	0.7760	X	X	X
	Y	33583	0.2240	Y	Y	
D3	14	2662	0.0610			
	15	10085	0.2475	15	15	
	16	4664	0.1221			16
	17	4921	0.1369		17	
	18	8399	0.2473			18
	19	5963	0.1853	19		
D5	9	5340	0.2084	9		
	11	7599	0.3624	11	11	
	12	5154	0.2682			12
	13	2856	0.1610			13
D7	8	3785	0.0971		8	
	9	7681	0.2218	9		
	10	12418	0.3984	10		10
	11	8013	0.2828		11	11
D8	10	23256	0.4229	10		10
	12	2676	0.0584		12	
	13	6137	0.1451			13
	14	14673	0.3736	14	14	
D13	8	6432	0.1414	8		
	11	20591	0.6225	11		11
	12	4276	0.1410		12	
	14	2472	0.0951		14	
D16	9	4995	0.1815			9
	10	5512	0.2225		10	10
	11	10175	0.4518	11		
	12	2978	0.1443		12	
D18	12	14071	0.2582	12		
	14	7781	0.1665			14
	15	4976	0.1141		15	
	17	14492	0.3767	17		17
	21	2632	0.0845		21	
D21	28	22272	0.3896	28		28
	29	22766	0.4125	29		29
	31.2	5124	0.0999		31.2	
	32.2	4876	0.0981		32.2	

Marker	**Mixture**	**Rel. Area**	**Rel. Weight**	**Suspect1**	**Suspect2**	**Victim**
CSF	10	1496	0.1195		10	
	11	1675	0.1472		11	
	12	5608	0.5376	12		12
	13	1885	0.1957	13		
FGA	19	12144	0.3517			19
	22	4119	0.1381		22	
	23	5032	0.1764	23		
	24	9120	0.3337	24	24	
Penta D	10	10463	0.3881	10		10
	11	6272	0.2559	11	11	
	13	7382	0.3560		13	13
Penta E	5	4740	0.1501			5
	7	1443	0.0640		7	
	11	6317	0.4400	11		
	13	2578	0.2122			13
	17	1242	0.1337		17	
THO1	6	22613	0.7786	6		6
	8	4822	0.2214	8	8	
TPOX	8	11840	0.6059	8	8	8
	10	3740	0.2392	10		
	11	2201	0.1549			11
vWA	14	5321	0.0796		14	
	15	1085	0.0174			
	16	5612	0.0960		16	
	17	27139	0.4930	17		
	18	16324	0.3140	18		18

Table 5.8: *Lago* data, 3-person mixture - a three individuals mixture composition with relative peak areas, relative peak weights, suspects' and victim's genotypes.

Marker	**Allele**	**Frequencies**
D7	8	0.164
	9	0.176
	10	0.272
	11	0.180
D8	10	0.097
	12	0.1404
	13	0.3852
	14	0.2135
D21	28	0.167
	29	0.205
	31.2	0.095
	32.2	0.072

Table 5.9: *Lago* data, 3-person mixture - Population alleles frequencies.

Chapter 6

Introduction to DNA mixtures

A mixed trace derives typically from an unidentified biological stain or in general from an admixture of biological material thought to be associated with a crime. They arise when two or more individuals contribute to the sample being tested. Think of, for example, a rape, or a robbery where an object has been handled by a number of individuals. Here we assume that a mixed DNA trace, of unknown origin and constitution and containing DNA from more than one contributor, has been obtained and profiled in connection with a specific crime (*e.g.* a murder). Furthermore, DNA profiles from identified individuals are measured. For example, if they belong to a victim and a suspect, our aim is to match them with those contained in the mixture to discriminate whether any of these have contributed DNA to the crime trace. It is worth noting that our intention is not to determine the innocence or guilt of a suspect, but whether the suspect and/or the victim can be assumed to be present in the mixture.

In a case at law, data can be represented by evidence involved in the hypotheses under test on which the court has to decide. Both hypotheses and evidence are characterized by uncertainty and the role of an expert statistician is to quantify this uncertainty. This can be done assigning a probability to the guilt of the suspect in the light of the presented evidence in order to define the weight of evidence. For this purpose we use the ratio between the probability of the evidence under the hypotheses of guilt and innocence (this ratio is the likelihood ratio). Such probabilities do not prove the guilt of the suspect, but in a number of cases, when the evidence is extremely significant, the court could acquire a benefit from them. In effect, mistakes are most likely to occur when deciding on the base of the collected evidence. For example, consider the trial of Sally Clark. Sally Clark was convicted of the murder of two of her new born babies and was declared innocence in 2003 after a number of appeals. The defence declared that her sons died for SIDS deaths (a particular type of unexplained natural death). However, an expert medical witness testified that natural double infant deaths are very rare, since the probability that a baby would have died from natural causes was one in 8543. Thus, the probability that both her babies would have died from natural causes was approximately one in 73 million[1]. The mistake was due to the fact that the courtroom misinterpreted the probability that Sally was guilt G given the evidence $\mathcal{E}$ as 1- the probability of the evidence given that Sally was innocent. Mathematically, $\Pr(G|\mathcal{E})$ was misinterpreted as one minus $\Pr(\mathcal{E}|\overline{G})$. This is clearly a mistake since, actually, $\Pr(G|\mathcal{E})$=1-$\Pr(\overline{G}|\mathcal{E})$. This error is known as "the prosecutors fallacy" or "transposing the conditional". Now, since the expert computed $\Pr(\mathcal{E}|\overline{G}) = 1/73\text{million}$, *the prosecutor's fallacy* gave $\Pr(G|\mathcal{E}) \simeq 1$ and Sally was convicted. However, the probability of a double infanticide, has been estimated in approximately one in 2 billion. Therefore, if we compare the probability of this event with the probability that the babies died for SIDS, we can obtain the following ratio called likelihood ratio:

$$\frac{\Pr(\mathcal{E}|G)}{\Pr(\mathcal{E}|\overline{G})} = \frac{1/2\text{billion}}{1/73\text{million}} \simeq 0.0365.$$

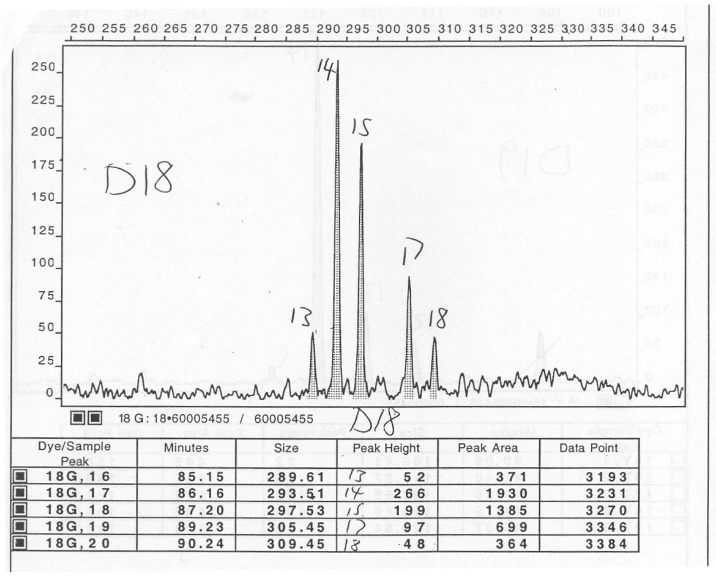
Thus, the weight of evidence is in favour of the Sally's innocence.

[1] For this calculation the hypothesis of independence of the two events has been supposed, but even without this assumption the probability would be low.

6.1 Representation of a DNA mixture

A DNA mixture is represented as an *electropherogram* which reports the alleles in the mixture as peaks having a specific height and area around each allele. Such area provides important information due to the fact that it is approximately proportional to the amount of DNA of that specific allele. Information about the composition of the mixture are given by the band density around each allele in the relative fluorescence units. An example of electropherogram is shown in Figure 6.1 where the alleles for marker vWA of a DNA mixture sample are amplified. Since there are three alleles it is a mixture made up of at least two contributors. The alleles have repeat number 15, 17 and 18, whilst peak area is reported in its appropriate column in the table under the picture.

As shown in Figure 6.2, a typical mixture may consist of major/minor



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
■ 18G,16	85.15	289.61	13 52	371	3193
■ 18G,17	86.16	293.51	14 266	1930	3231
■ 18G,18	87.20	297.53	15 199	1385	3270
■ 18G,19	89.23	305.45	17 97	699	3346
■ 18G,20	90.24	309.45	18 48	364	3384

Figure 6.1: Electropherogram for marker vWA in a DNA mixture sample. The alleles have repeat numbers 15, 17, 18, whilst peak area is reported in its appropriate column in the table under the picture.

components. If a sufficient difference in peak areas between the two pairs of alleles exists, the major contributor is sufficiently represented and therefore it can be separated according to its area. Hence, assuming a 2-person mixture

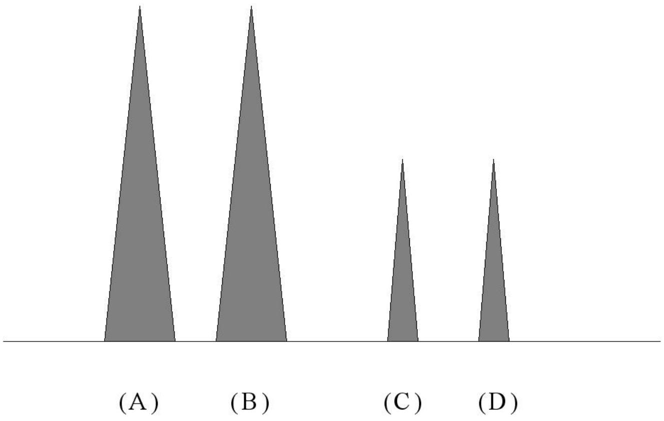


Figure 6.2: A four-alleles mixture showing major contributor's profile $\{A, B\}$, and minor $\{C, D\}$.

in the example in Figure 6.2, a possible combination should be the profile $\{AB\}$ for the major contributor and $\{CD\}$ for the minor. If we took into account the repeat number of the alleles only, then also other combinations would be admitted, *e.g.* $\{AC\}$ for the major and $\{BD\}$ for the minor. On the contrary, such combinations of profiles would be excluded if peak area information were added. It is worth highlighting that, when the mixture consists of the same amount of DNA generated by the two contributors, *i.e.* 50:50, the repeat number of alleles and peak area give the same information, so both pairs of the considered profiles, *i.e.* $\{AB, CD\}$ and $\{AC, BD\}$ are accepted.

We suppose now that the contributors share an allele at a certain marker. For example, the genotype of two persons are $\{AB\}$ and $\{BC\}$ where allele B is in both profiles at that marker. This phenomenon is called *masking* because shared alleles result in "masking" causing asymmetry in the allelic bands. Since the contributions are additive in the mixture, for a crime trace with ratio 2:1, the proportions for the alleles are A:B:C=2:3:1, and this ratio is approximately the same across markers (see Figure 6.3). In a similar scenario the interpretation is more informative but also more difficult, since the profile is no balanced. If we consider all of the possible combinations, one could be $\{AB\}$ for the major component and $\{BC\}$ for the minor, but also other reasonable combinations exist, such as $\{BB, AC\}$. Thus, we need to

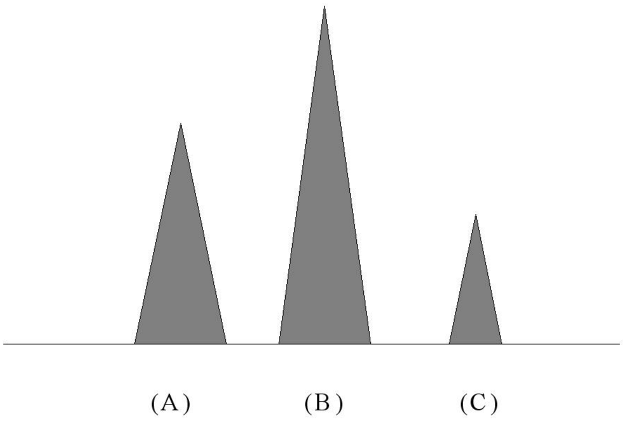


Figure 6.3: A three-alleles mixture showing major contributor's profile $\{A, B\}$, and minor $\{B, C\}$.

find that one which best fits to the peak areas rejecting all of the alternatives that give low probabilities for the areas.

6.2 Basic framework

In this section and in chapter 7 we investigate the case of a DNA mixture from exactly two contributors, which is apparently the most common scenario in forensic casework. For the sake of simplicity, complications, such as two traces analyzed at the same time and DNA mixtures involving more than two contributors, are studied in chapters 8 and 9.

In a courtroom context we need to formulate hypotheses H about suspect and victim. A typical analysis of a crime sample compares the prosecution hypothesis H_p with the defence hypothesis H_d. For example the prosecution may hypothesise that both victim and suspect contributed to the mixture, *i.e.* $H_p : v\&s$, whilst the defence may hypothesise that the suspect did not contribute to the mixture but that only the victim and an unknown individual u contributed, *i.e.* $H_d : v\&u$. Henceforth we refer to H_p as the null hypothesis H_0 and H_d as the alternative hypothesis H_1. Furthermore, we denote $\mathcal{E}$ the elements of evidence.

In this context the adjudicator needs to estimate the conditional probability for either hypotheses given the evidence, $\text{pr}(H_0|\mathcal{E})$ and $\text{pr}(H_1|\mathcal{E})$. Since it is not often possible to assess such probabilities directly we need to calculate them applying the *Bayes theorem*. As it is well known, we can express

$$\frac{\mathrm{pr}(H_0|\mathcal{E})}{\mathrm{pr}(H_1|\mathcal{E})} = \frac{\mathrm{pr}(H_0)}{\mathrm{pr}(H_1)} \frac{\mathrm{pr}(\mathcal{E}|H_0)}{\mathrm{pr}(\mathcal{E}|H_1)}, \tag{6.1}$$

where the left-hand side is the *posterior odds* for comparing H_0 versus H_1 given the evidence, whilst the first term in the right-hand side is the *prior odds* which represent prior knowledge on the hypotheses, and the second term in the right-hand side is the *likelihood ratio* (LR).

We now consider the joint probability of observing the entire DNA evidence, *i.e.* the mixed trace and the profiles of identified individuals. This is the likelihood associated to the specific hypothesis that the observed DNA profiles in the mixture match those in the set of the examined individuals. Such likelihoods can be used to compare more hypotheses. Particularly, in the case of just two hypotheses, this comparison is represented by the likelihood ratio.

In a courtroom, statistician needs to give the weight of evidence which is given in the form of the likelihood ratio. For this reason, forensic experts are often induced to formulate the reasonable assumption that the prior probabilities for each hypothesis H are equal, assessing that there is no evidence to discriminate the suspect from any other potential suspect (in law this is called *presumption of innocence*). Actually, it is preferred to leave to adjudicators, judges or juries to formulate the prior assessments and update the likelihood ratio to get $\Pr(H|\mathcal{E})$. As a consequence of the *Bayes theorem*, the likelihood ratio becomes the conditional probabilities, under H, of obtaining the crime trace evidence. However, it is worth noting that the ratio of the likelihoods only enters in the analysis, whilst their single values are not needed. Additionally, if the likelihood ratio is greater than one, then the evidence favours H_0, but if it is less than one, then the evidence favours the alternative hypothesis H_1.

In a single-contributor case, *i.e.* not in a mixture but in a DNA stain made up of the DNA of a single individual, the probability of observing the evidence, *i.e.* the stain profile and the suspect's profile, under the hypothesis

H_0:s that the stain profile comes from the suspect only, is one if the evidence and the suspect's profile are compatible. Thus, the likelihood ratio reduces to the reciprocal of the posterior probability that the suspect is not the contributor, $\mathrm{pr}(\mathcal{E}|H_1)$. If we ignore complications, such as drop-out alleles or stutters, this ratio becomes the reciprocal of the profile's population frequency.

We now assume that the evidence consists of DNA profiles extracted from a suspect s and a victim v and a mixed trace. We further suppose that $v \cup s = \xi$, *i.e.* the mixture ξ is given by the union of the two suspect's and victim's profiles. We test the hypotheses $H_0 : v\&s$ versus $H_1 : v\&u$, then the likelihood ratio LR can be expressed as

$$LR = \frac{\mathrm{pr}(\mathcal{E}|H_0)}{\mathrm{pr}(\mathcal{E}|H_1)} = \frac{1}{\sum_y \mathrm{pr}(u = y)}, \tag{6.2}$$

where y are all the profiles, except that of the victim, compatible with the mixture, *i.e.* y is such that $v \cup y = \xi$. If we additionally assume that all individuals belong independently to a common population with known allele frequencies, we obtain

$$LR = \frac{1}{\sum_y \mathrm{p}_y}, \tag{6.3}$$

where p_y is the allele frequency of the profile y. Some examples are shown.

Example 6.1 Assume that a DNA mixture $\xi = \{A, B, C\}$ from two contributors is observed and that the following profiles for the suspect and the victim are examined: $s = \{B, C\}$, $v = \{A, C\}$. We are interested in testing the hypotheses $H_0 : v\&s$ versus $H_1 : v\&u$. Hence, equation (6.3) becomes

$$LR = \frac{1}{\mathrm{p}_B^2 + 2\mathrm{p}_A\mathrm{p}_B + 2\mathrm{p}_B\mathrm{p}_C}$$

where p_i is the frequency of allele i in the population. This result is due to the fact that, since it must be $v \cup y = \xi$, the profile y will be represented by one of the following: $\{B, B\}, \{A, B\}, \{B, A\}, \{B, C\}, \{C, B\}$.

□

Example 6.2 Suppose that a DNA mixed trace $\xi = \{A, B, C, D\}$ is observed and that only the following suspect's profile is examined: $s = \{A, B\}$. We are interested in testing the hypotheses $H_0 : s\&u$ versus $H_1 : 2u$. Assuming that this is a 2-person mixture, Table 6.1 shows all the possible genotype combinations. We further calculate the probabilities for each combination. Thus, the probability of genotype $\{A, B\}$ is $2\text{p}_A\text{p}_B$, and the probability of genotype $\{C, D\}$ is $2\text{p}_C\text{p}_D$. Multiplying them together we obtain the probability of $\{A, B\} \cup \{C, D\}$ as $4\text{p}_A\text{p}_B\text{p}_C\text{p}_D$. This is repeated for each combination, and the sum of all the probabilities gives $\sum_y \text{p}_y = 24\text{p}_A\text{p}_B\text{p}_C\text{p}_D$.

Thus,

$$LR = \frac{2\text{p}_C\text{p}_D}{24\text{p}_A\text{p}_B\text{p}_C\text{p}_D} = \frac{1}{12\text{p}_A\text{p}_B}.$$

□

Genotype p1	Genotype p2	Probability
AB	CD	$4\text{p}_A\text{p}_B\text{p}_C\text{p}_D$
AC	BD	$4\text{p}_A\text{p}_B\text{p}_C\text{p}_D$
AD	BC	$4\text{p}_A\text{p}_B\text{p}_C\text{p}_D$
CD	AB	$4\text{p}_A\text{p}_B\text{p}_C\text{p}_D$
BD	AC	$4\text{p}_A\text{p}_B\text{p}_C\text{p}_D$
BC	AD	$4\text{p}_A\text{p}_B\text{p}_C\text{p}_D$
Total		$24\text{p}_A\text{p}_B\text{p}_C\text{p}_D$

Table 6.1: Example 5.2 - all the possible genotype combinations with associated probability for an observed mixture $\xi = \{A, B, C, D\}$.

6.3 OOBN for analysing DNA mixtures

The statistical tools used to analyse the DNA mixtures in this thesis are the object-oriented Bayesian networks (OOBNs). OOBNs are a recent extension of the BNs. They are blocks of Bayesian networks combined in a hierarchical form where any node itself can represent a (object-oriented) network containing several *instances* of other generic *classes* of networks. Instances can have ordinary nodes as well as interface *input* and *output* nodes. An input node can have at most one incoming arrow from an output node of another network. Input and output nodes must have identical probabilistic structure, *i.e.* must be of the same type and have the same states, since the arrows connecting output and input nodes represent identity links, whilst arrows between ordinary nodes represent probabilistic dependence. Each node has at least two states that can be Boolean (defined as *true* or *false*), numerical (discrete or continuous), or a range. Furthermore, each node can have assigned a function which defines how the probability distribution over states of the node depends on the parents of the node.

Henceforth, we indicate in **bold** a network class, whilst in `teletype` a single node. In figures, we represent instances of a class with a rounded rectangle, discrete nodes with a single outline, and continuous nodes with a double outline. Output nodes are always drawn with a grey outer ring and a solid line, rather than input nodes that are represented by a dotted line. Observation nodes, *i.e.* where the evidence is entered, are coloured in grey, whilst target nodes, *i.e.* where hypotheses are formulated and the network returns an output, are coloured in dark.
Figure 6.4 (b) represents an example of two instances connected through their output and input nodes. These instances reproduce the Bayesian network in Figure 6.4 (a).

In this thesis we show how object-oriented Bayesian networks are an useful tool for evaluating DNA mixtures. Dawid *et al.* (2002) introduced probabilistic expert systems (PES) for analysing DNA evidence and, in particular, they used a Bayesian network (BN) to solve forensic identification problems. Since this network includes a number of repeated structures (for example, the structure of the suspect's and victim's genotype are the same) it can be synthesized with an OOBN. Thus, we modified the BN representation of Mortera *et al.* (2003) to obtain an OOBN. Figure 6.5 shows the Bayesian network, for the **marker** class, used by Mortera *et al.* (2003), whilst Figure 6.6 shows how this network has been modified obtaining an OOBN structure. We note that our OOBN includes less nodes than the network of the authors. In effect, for example, in Figure 6.5, the nodes, `A_in_v`, `B_in_v`, `Av`, `Bv`, etc. which are referred to the victim's genotype, have

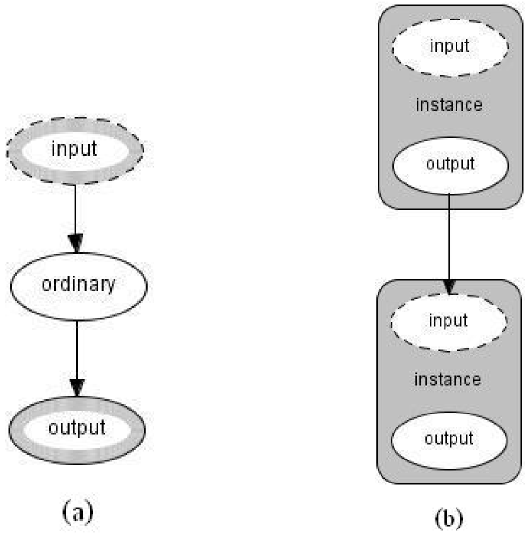


Figure 6.4: (a) A Bayesian network an output, input and ordinary node. (b) Instances of the network in (a) and they are connected through their output and input node.

been included in the object vgt in Figure 6.6. Similarly for the nodes referred to the suspect's and unknown individual's genotypes. Details of our network are given in Appendix A.1. Cowell *et al.* (2007b) introduced how include peak area information in the network. Thus, the authors modelled peak weights through an OOBN representation that allow to solve both identification and separation problems. Details of the network are given in Appendix A.2. Furthermore, we extended both networks in order to include a second trace (see Appendix B) and a third contributor to the mixture (see Appendix C).

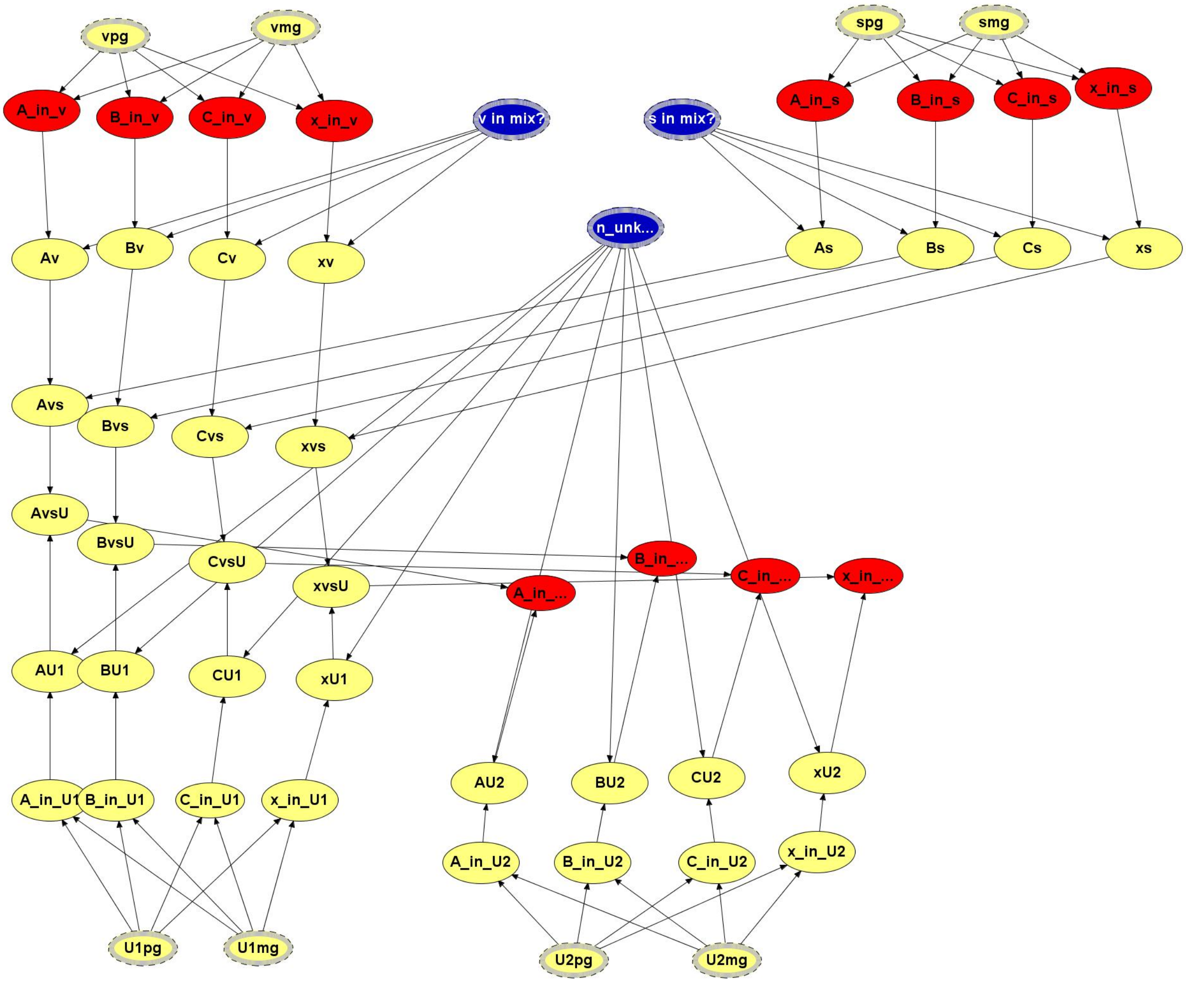


Figure 6.5: Bayesian network used by Mortera *et al.* (2003). Marker class.

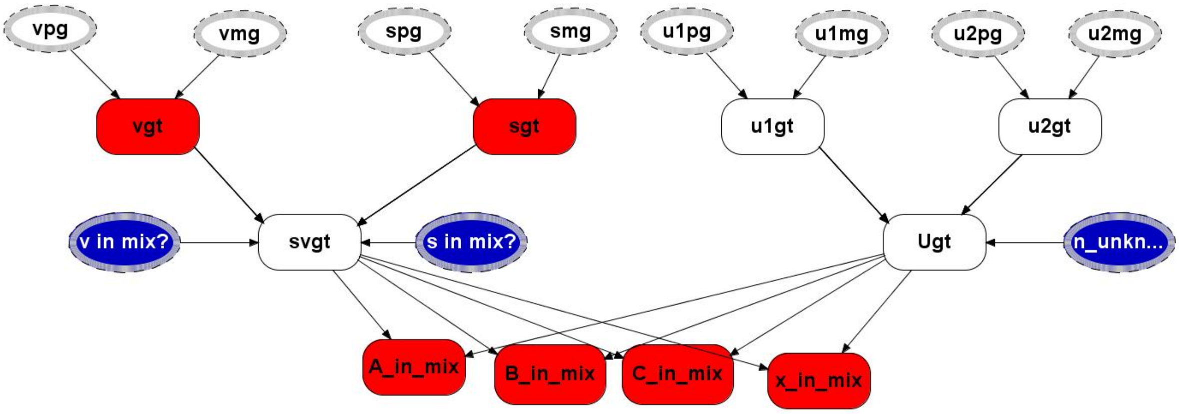


Figure 6.6: The Bayesian network used by Mortera *et al.* (2003) and modified as an OOBN. Marker class.

Chapter 7

Identification and separation for 2-person DNA mixtures

This chapter is concerned with the analysis of mixed DNA traces involving exactly two contributors. In section 7.1 a forensic identification analysis has been performed using the PES constructed by (Mortera *et al.* 2003) which employs information about which alleles were present in the mixture. Actually, this PES, having the form of a Bayesian network, has been changed assuming the structure of an OOBN. Details on the network are given in Appendix A.1. (Cowell *et al.* 2007b) introduced a method, based upon object-oriented Bayesian networks, for analysing DNA mixtures using peak area information in addition to allele's repeat numbers. After introducing in § 7.2.1 the conditional-Gamma and conditional-Gaussian models for peak weights, we illustrate in § 7.2.2 and 7.2.3 how to use the OOBN described in Appendix A.2 to solve both identification and genotypes' separation problems in mixtures of two DNA samples.

7.1 Identifying the genotype each of the possible contributors to the mixture

In this section the network described in Appendix A.1 is applied to a specific case, suppose for example a murder. Data are given in Table 5.4 § 5. Such data are DNA mixtures realized in laboratory, thus a-priori the profiles of contributors are known. Additionally, the profiles of two identified individuals, it is supposed a victim v and a suspect s, are examined and a-priori it is known they match with those of the contributors.

In such analysis the evidence comprises DNA profiles extracted from the mixed trace, from a suspect, and a victim and the hypotheses of interest are reported in Table 7.1. These hypotheses mimics different cases, for example a murder case where a DNA mixed stain coming from the victim and a suspect is found, or a robbery case where a DNA mixture coming from two suspects is found, etc. However, in a courtroom only two hypotheses will be compared.

It is worth noting that all these hypotheses involve two contributors.

Hypotheses under test	
s&v	both suspect and victim contributed to the mixture
s&u	both suspect and an unknown individual contributed to the mixture
v&u	both victim and an unknown individual contributed to the mixture
2u	two unknown individuals contributed to the mixture

Table 7.1: Hypotheses under test.

In effect, since two is the maximum number of alleles that can be observed for each individual, the presence of more than two alleles in the mixture, *i.e.* three observed alleles for markers D5, D16 and D18, and four for the remaining markers, allows to conclude that there must have been at least two contributors to the crime trace. Conversely, we can say nothing about the upper bound of contributors.

The main investigation is whether suspect and victim contributed to the mixture. A variant could be represented by the introduction of an *unknown contaminator u* instead of the victim. Firstly, the total number of contributors to the crime trace is assumed to be known and it is supposed to be exactly 2. Secondly this assumption is relaxed to be proved with an appropriate analysis. Thus, the evidence is entered in the appropriate nodes and propagated throughout the network so that, the target node, shown in Figure 7.1, contains the updated probabilities. In particular, the evidence on the suspect's genotype is entered in the nodes contained in the instance `sgt`

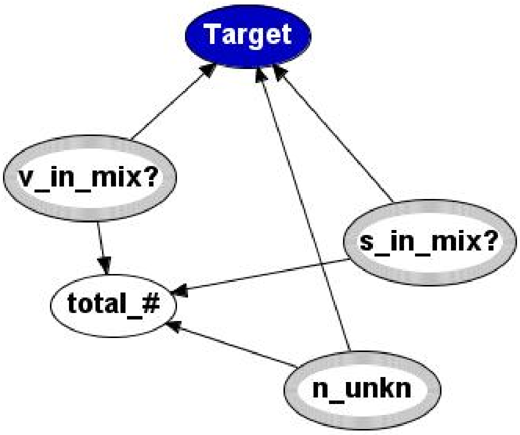


Figure 7.1: Two person mixture. Target class.

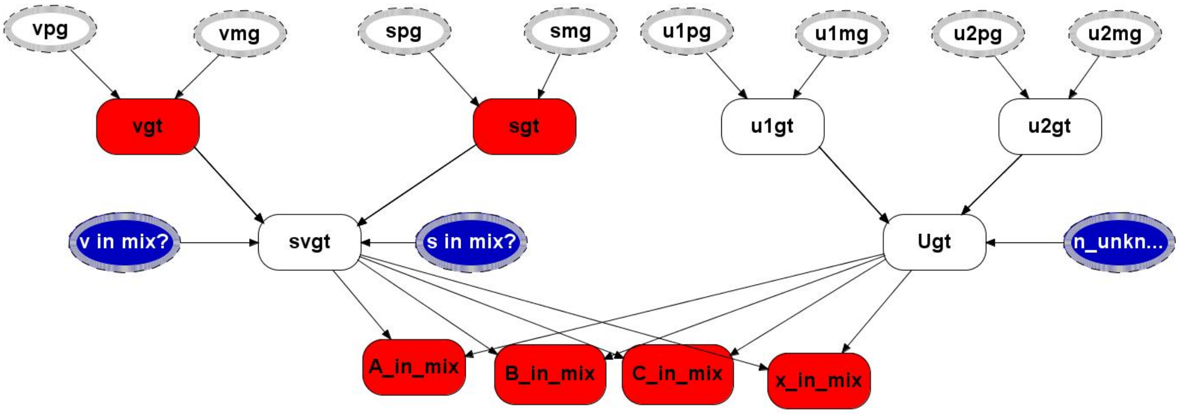


Figure 7.2: Two person mixture. Marker class.

in Figure 7.2; the evidence on the victim's genotype is entered in the nodes contained in the instance `vgt`; the evidence on the observed alleles in the mixture is entered in the node contained in the instance `A_in_mix`, `B_in_mix`, `C_in_mix` and `x_in_mix`.

Thus, the ratio of the updated probabilities is taken to obtain the likelihood ratios. Table 7.2 gives the logarithm of the likelihood ratio among the pairwise comparisons in the first column.

Such ratios show strong evidence against both suspect and victim whereas the highest value $10^{12.11} \simeq 1.3 \times 10^{10}$, is for the comparison $H_0 : s\&v$

vs. $H_1 : 2u$. However, in a case at law just two hypotheses are compared, *i.e.* the prosecution and the defence hypotheses. Since for this analysis we supposed a murder, in this scenario we would be interested in investigating whether the suspect contributed to the mixture. Thus, we would consider just the comparison involving the hypotheses $H_0 : s\&v$ and $H_1 : v\&u$. In the second row of Table 7.2, this comparison shows strong evidence against the suspect.

It is assumed now that only the genotype of a suspect is available. In this case we are interested in comparing the hypotheses $H_0 : s\&u$ versus $H_1 : 2u$. The final row in Table 7.2 shows the respective logarithm of the likelihood ratio that indicates evidence against the suspect since the likelihood ratio is $10^{3.40} \simeq 2,500$.

Hypotheses	$Log_{10}LR$
s&v vs. 2u	12.11
s&v vs. v&u	8.74
s&v vs. s&u	6.71
s&u vs. 2u	3.40

Table 7.2: *Lago* data, 2-person mixture - logarithms of the likelihood ratio in favour of suspect and victim and in favour of suspect and an unknown individual.

If the node `total_#` in Figure 7.1 is not constrained to be two as above, cases, where the total number of contributors is unknown, can be handled and hypotheses about it can be made. Table 7.3 displays the posterior probabilities for the total number of contributors. As expected, the posterior probability for the hypothesis that the mixture comprises less than two profiles is zero. However, almost the entire evidence is against two contributors and the greater the number of hypothetical contributors the lower the probability. This is due to the fact that the probability of the evidence under H_0 and H_1 is maximised when the total number of contributors is minimised (Gill *et al.* 2006).

Number contributors	Probability
0	0
1	0
2	0.9995
3	0.0005
4	0.0000289

Table 7.3: *Lago* data, 2-person mixture - probability of the total number of contributors.

7.2 Analysis of DNA mixtures using peak area information

In this section a method for analysing forensic identification problems using peak area information is introduced. The network applied is the one from Cowell *et al.* (2007b) which details are given in Appendix A.2. The aim is not only to investigate whether individuals, whose profiles have been measured, have contributed to the mixture, but also to discriminate the genotypes of the unknown individuals contributing to the mixture and to predict their DNA profiles.

Both analyses have been performed using a single probabilistic model. Thus, the entire OOBN network can be used to solve both problems of suspect's and victim's identification and prediction of the contributor's profiles. In the first case the likelihood ratios are read in the `target` node, whilst in the latter case separated profiles are indicated in the **jointgt** class.

It is worth noting that, comparing this OOBN with the one described in Appendix A.1 to solve identification problems in § 7.1, here node concerning the total number of contributors to the mixture is not present. This is due to the fact that in this model the total number of contributors is assumed to be known since the lower bound is always defined by the evidence. Thus, inference on it is not made. However, a check on the total number of contributors can be made using the discrete network described in Appendix A.1.

7.2.1 Basic model assumptions

It is assumed as the basic model for the allele composition of the mixture sample the Mendelian genetic model explained in § 4, and we further assume to known the gene frequencies of single STR alleles. Such gene frequencies are those reported in Table 5.5 § 5. For a mixture made up of two contributors (`p1, p2`), when the mixture sample is amplified it consists of an unknown number of cells from `p1`, and a further unknown number of cells from `p2`, where every cell contains exactly two alleles for each marker. The fraction (or proportion) of cells from the first contributor `p1` measures the amount of DNA originated from `p1` across the markers. This quantity is denoted as θ.

Peak area of alleles provides information about its post-amplification proportions for each marker. Their information is included in the analysis through the *relative peak weight*. The absolute *peak weight* w_a of an allele with *repeat number* a is defined as the product between the peak area α_a of the allele a and its repeat number, *i.e.*

$$w_a = a\alpha_a.$$

This product has been introduced to correct the peak weight whereas alleles with a high repeat number tend to be less amplified than alleles with a low repeat number. Now, the observed *relative peak weight* r_a are defined as the following ratio

$$r_a = w_a \backslash w_+,$$

where $w_+ = \sum_a w_a$, so that the constraint $\sum_a r_a = 1$ holds.

The conditional-Gamma Model

Here, it is assumed that:

(i) there are I potential contributors to the mixture;

(ii) the analysis of the mixture is based on M markers;

(iii) the general marker $m = 1, 2, ..., M$ has A_m allelic type.

Furthermore, we define $\theta = \theta_i$, for all $i = 1, 2, ..., I$, $\theta_i \geq 0$ and $\sum_i \theta_i = 1$, where θ_i is the proportion of DNA in the sample from individual i. Let $\gamma\theta_i$ be the contribution of the i-th individual to the mixture, where γ is the total amount of DNA in the sample.

Let W_{ia}^m denote the contribution of individual i to the peak weight at allele a of marker m. Then,

$$W_{ia}^m \sim \Gamma(\rho_m \gamma \theta_i n_{ia}^m, \eta_m)$$

where $\Gamma(\alpha, \beta)$ denotes the gamma distribution with density

$$f(w) = \frac{\beta^\alpha}{\Gamma(\alpha)} w^{\alpha-1} e^{-\beta w},$$

where ρ_m is an amplification factor, and η_m a scale parameter. It is reasonable to suppose that (i) the pre-amplification mixture proportion vector of DNA in the sample θ is constant across markers, (ii) the peak weight for an allele is approximately proportional to the amount of DNA of that specific allelic type and (iii) the peak weight for that allele is the sum of the respective weights for each contributor, when 2 or more contributors have the same allelic type. Although the total weights W_{+a}^m of each single allele in the mixture can be measured, the individual weights W_{ia}^m are *unobservable.* Thus, being W_{+a}^m the sum of the individual contributions, it has Gamma distribution as follows

$$W_{+a}^m = \sum_i W_{ia}^m \sim \Gamma(\rho_m \sum_i \gamma \theta_i n_{ia}^m, \eta_m),$$

with $i = 1, ..., I; m = 1, ..., M; a = 1, ..., Am.$
Now, let

$$B_a^m = \sum_i \gamma \theta_i n_{ia}^m,$$

be the *weighted allele number*, then their sum B_+ has the property

$$B_+ = \sum_a B_a^m = \sum_a \sum_i \gamma \theta_i n_{ia}^m = \sum_i \gamma \theta_i \sum_a n_{ia}^m = \sum_i 2\gamma \theta_i = 2\gamma$$

to be twice the total amount of DNA and to be independent of m. Letting

$$\mu_a^m = \frac{B_a^m}{B_+} = \frac{\sum_i \theta_i n_{ia}^m}{2}, \tag{7.1}$$

then

$$W_{+a}^m \sim \Gamma(2\rho_m \mu_a^m \gamma, \eta_m), \tag{7.2}$$

and

$$W_{++}^m = \sum_a W_{+a}^m \sim \Gamma(2\rho_m \gamma, \eta_m), \tag{7.3}$$

where $\sum_a \mu_a^m = 1$. The peak weight is here reported in terms of relative values in order to avoid arbitrariness in its scaling, thus

$$R_a^m = \frac{W_{+a}^m}{W_{++}^m}.$$

Now, the set of relative peak weights on each marker has a Dirichlet distribution because it is the ratio between a Gamma and the sum of Gamma distributions

$$R^m = \{R_a^m\} \sim Dir(2\rho_m \mu_a^m \gamma). \tag{7.4}$$

It is worth noting that R^m is independent on the scale parameter η_m and has

$$E[R_a^m] = \mu_a^m \tag{7.5}$$

and

$$V[R_a^m] = \frac{\mu_a^m(1-\mu_a^m)}{2\rho_m\gamma + 1} = \sigma_m^2 \mu_a^m (1-\mu_a^m), \tag{7.6}$$

where

$$\sigma_m^2 = \frac{1}{2\rho_m \gamma + 1}.$$

The total and the relative peak weights are independent from one another and from any other variable, conditional on the vector μ^m. Thus, the relative peak weights R_a^m contains information on mixture composition from the peak areas about μ^m. The respective likelihood factorizes as

$$L(\mu \mid W) = f(W \mid \mu) = L(\mu \mid R, W_{++})$$

$$\propto L(\mu \mid R) == \prod_a \frac{r_a^{2\rho\gamma\mu_a - 1}}{\Gamma\left(\mu_a\left(\frac{1}{\sigma^2}-1\right)\right)} = \prod_a \frac{r_a^{\mu_a\left(\frac{1}{\sigma^2}-1\right)}}{\Gamma\left(\mu_a\left(\frac{1}{\sigma^2}-1\right)\right)}, \tag{7.7}$$

where the dependence on marker m has been dropped and where the likelihood depends on the *observed relative peak weights* r_a, the mean of the relative peak weights μ_a, and the variance σ^2 Cowell *et al.* (2006); Cowell *et al.* (2007a).

Conditional-Gaussian approximation

Here, a conditional-Gaussian (CG) model is assumed for the peak areas. This is an approximation of the above more appropriate quantitative model based on the conditional-Gamma distribution. (Cowell *et al.* 2007b) specified the following distributional approximation:

$$R_a^m \sim N(\mu_a^m, \tau_a^2), \tag{7.8}$$

where μ_a^m are defined as in equation (7.1) and n_{ia}^m is the number of alleles with repeat number a for marker m possessed by person i. The error variance τ_a^2 is defined as

$$\tau_a^2 = \sigma^2 \mu_a^m (1 - \mu_a^m) + \omega^2, \tag{7.9}$$

where σ^2 and ω^2 are variance factors deriving by the variation due to amplification and measurement processes. Particularly, if $I = 2$, *i.e.* the mixture is made up of two contributors, the mean in (7.1) becomes

$$\mu_a^m = \frac{\theta n_{1a}^m + (1-\theta) n_{2a}^m}{2}. \tag{7.10}$$

Here θ is the proportion, or fraction of DNA in the mixture generating from the first contributor.

This mean for each allele represents its pre-amplification proportion, whilst the amplification variance σ_a^2 is zero in the following two extreme scenarios:

(1) if the pre-amplification proportion is zero, and the mixture does not contain alleles of a certain type, so there is not even post-amplification;

(2) if the pre-amplification proportion is unity, and the mixture pre-amplification comprises only one allele of a specific type of a certain marker, so also the post-amplification mixture contains only one of that allelic type.

In a perfect heterozygote, the proportions of the allele peak areas would be the same, *i.e.* (1 : 1). However, during the PCR amplification process, measurement errors or slight differences can occur causing variation between the observed peak areas of alleles in a given heterozygote. Gill *et al.* (1997) showed that the relative peak area differences, denoted by ϕ[1], increases as the mean peak area of the two STR fragments increases, and they found that calculated standard deviation of ϕ between 0.06 and 0.08. Now, (Cowell *et al.* 2007b) used for the analysis $\sigma^2 = 0.01$ and $\omega^2 = 0.001$. In effect, this chosen produces, for $\mu_a = 0.5$ produces a standard deviation equal to $\sqrt{0.01/4 + 0.001} = 0.06$ which is consistent with the result of Gill *et al.* (1997). Thus, the model provide a correct forensic analysis, even if this issue needs further consideration; since preliminary investigations indicate that the variance factor may depend on the total amount of DNA available in the mixture. This suggests that such variance varies from case to case.

To avoid arbitrariness in scaling the relative peak weights have been considered. These relative peak weights must sum to unity, thus their correlation must be taken into account. If we assume $\omega = 0$, the variance in (7.9) ignores this correlation, but (Cowell *et al.* 2007b) proved that using instead the distribution with a variance factor

$$R_a^m \sim N(\mu_a^m, \sigma^2 \mu_a^m) \tag{7.11}$$

this problem is overcome.

In general, let $X = (X_1, ..., X_A)$ be a vector of independent and normally distributed random variables such that their sum S is one. Then, $X_a \sim$

[1] Gill *et al.* (1997) defined ϕ as the ratio between the area of smallest peak and the area of largest peak. This quantity is 1 if peak areas are the same.

$N(\mu_a, \tau_a^2)$ and has joint density

$$f(x_1, ..., x_d|\mu, T) = \left(\frac{1}{2\pi}\right)^{d/2} \prod_{a=1}^{A} \frac{exp\{-\sum_a \frac{(x_a-\mu_a)^2}{2\tau_a^2}\}}{\tau_a}.$$

Here, $\mu = (\mu_1, ..., \mu_A)$ is a vector and τ^2 are the diagonal elements of the matrix T=diag$(\tau_a^2, ..., \tau_A^2)$. It is worth highlighting that since T is a diagonal matrix, the covariance elements are zero since the random variables are independent. Thus, this distribution does not take into account the correlation between the variables and this correlation is due to the fact that they must sum to unity. If the distribution of the sum $S = \sum_a X_a$ is considered, this is still Normal with unit mean (since $\sum_a \mu_a = 1$) and variance $\tau^2 = \sum_a \tau_a^2$. The conditional distribution $X|S = 1$ is multivariate Normal with the same mean vector $\mu = (\mu_1, ..., \mu_A)$ and covariance matrix T^* which has elements

$$\tau_{aa}^* = \tfrac{\tau_a^2(\tau^2-\tau_a^2)}{\tau^2}, \qquad \tau_{ab}^* = \tfrac{-\tau_a^2\tau_b^2}{\tau^2}.$$

If the variance τ_a^2 has the form in (9.6), *i.e.* $\tau_a^2 = \sigma^2\mu_a$, then $\tau^2 = \sum_a \tau_a^2 = \sigma^2 \sum_a \mu_a = \sigma^2$, where τ^2 is independent of μ. Additionally, the elements of the covariance matrix T^* become

$$\tau_{aa}^* = \sigma^2\mu_a(1 - \mu_a), \qquad \tau_{ab}^* = -\sigma^2\mu_a\mu_b.$$

Here the variance τ_{aa}^* is exactly the same as in the above conditional-Gaussian approximation model in (7.8), except for the factor ω^2. This justifies in using the *correct relative peak weights* in (9.6) since they take into account their correlation.

7.2.2 Identifying the genotype each of the possible contributors to the mixture

In this section forensic identification problems are analysed using quantitative peak area information in addition to alleles repeat number information. The main aim is to show that the introduction as evidence of peak area of the allele increases the likelihood ratios in favour of the suspect and victim providing more accurate probabilities. Thus, the following comparisons are of interest:

(i) H_0: $s\&v$ vs. H_1: $2u$,

(ii) H_0: $s\&v$ vs. H_1: $v\&u$,

(iii) H_0: $s\&v$ vs. H_1: $s\&u$.

Then, the logarithms of such ratios are compared with those obtained in the previous section. A variant could be represented by the introduction of an *unknown contaminator* u instead of the victim and therefore, the hypotheses H_0: $s\&u$ (both suspect and victim contributed to the mixture) versus H_1: $2u$ (two unknown individuals contributed to the mixture) are considered. Here peak areas are modeled with a conditional-Gamma distribution. However, it is intention to carrying out also an analysis with the conditional-Gaussian approximation in order to show that these two models provide similar probabilities.

Here, the evidence consists of DNA profiles of a suspect s and a victim v, mixed trace, and relative peak weights. The evidence, represented by the data reported in Table 5.4 § 5, is entered in the appropriate nodes and propagated throughout the network described in Appendix A.2. If peak area information is used, the likelihoods are entered in the appropriate nodes. Thus, the posterior probabilities associated with the target node are produced and, taking their ratios, the likelihood ratios of interest are obtained. The logarithm on base 10 of such ratios are displayed in Table 7.4. In the second column of this table the logarithms are obtained under the assumption that only the evidence on the repeat number of the alleles is used. These ratios are equal those given in Table 7.2 in § 7.1. On the contrary, column "Areas" displays the logarithm of the likelihood ratio obtained adding peak area information. The inclusion of area information is indeed strengthening the evidence against the suspect and victim since the logarithm of the ratios increase for all the hypotheses considered in the table. For example, for the comparison H_0: $s\&v$ vs. H_1: $2u$ the ratio increases from $10^{12.11} \simeq 1.3 \times 10^{10}$ to $10^{16.77} \simeq 59 \times 10^{15}$ when the area information is included, which corresponds approximately to a factor $45,700$. However, in a case at law just the prosecution and the defence hypotheses are compared. If we consider, for example, a murder case, we would be interested in investigating whether the suspect contributed to the mixture. Thus, we would consider just the comparison involving the hypotheses $H_0 : s\&v$ and $H_1 : v\&u$.

The variable θ describes the proportion (or fraction) of DNA contributed by the first contributor. Table 7.5 shows the logarithm of the likelihood ratios in favour of H_0: $s\&v$ versus H_1: $2u$ as function of this parameter. Such logarithm assumes negative values for $\theta \leq 0.4$ highlighting that the evidence favours $H_1 : 2u$, whilst the maximum occurs for $\theta = 0.7$. Figure 7.3 shows the posterior density of the mixture proportion. The analysed mixture has a proportion $10 : 1$, which, if we scale θ in a range $[0, 0.1, 0.2, ..., 1]$, corresponds to a value of θ equal to 0.9 that is consistent with the maximum value displayed in the figure.

Hypotheses	Alleles	Areas
s&v vs. 2u	12.11	16.77
s&v vs. v&u	8.74	9.30
s&v vs. s&u	6.71	7.74

Table 7.4: *Lago* data, 2-person mixture - logarithms of the likelihood ratio in favour of suspect and victim when only evidence on repeat number of the alleles are available, and when peak are information is added.

Suppose now that DNA profiles are extracted only from a suspect and from the mixed trace and the other contributor is called a *contaminator*. In this scenario the evidence is given by the mixture composition and the suspect's profile. Table 7.7 displays the logarithm of the likelihood ratios in favour of H_0: $s\&u$ versus H_1: $2u$. Also in this case the inclusion of area has a dramatic effect on the likelihood ratio. In effect, it changes from $10^{3.40} \simeq 2,499$ to $10^{7.44} \simeq 27,500,000$ corresponding approximately to a factor $11,000$. Table 7.7 shows the logarithm of the likelihood ratios in

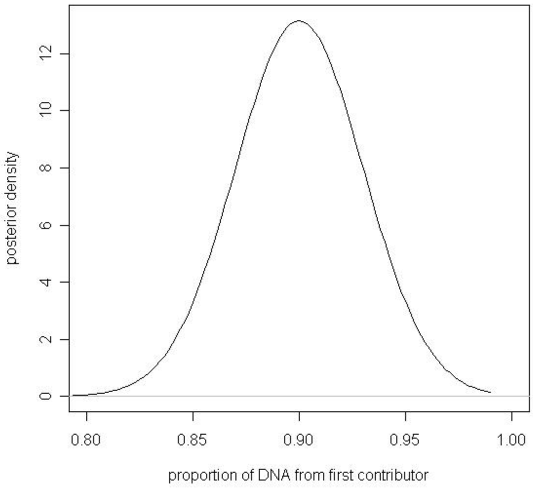


Figure 7.3: Lago data - posterior density of the proportion of DNA from the major contributor for 10:1 2-person mixture.

relation with the proportion of DNA originating from the first contributor. This assumes negative values for $\theta \leq 0.4$ indicating that the evidence favours two unknown individuals as contributors to the mixture. Figure 7.4 shows the posterior density of the mixture proportion. The real DNA proportion of the mixture is $10 : 1$, which, if we scale θ in a range $[0, 0.1, 0.2, ..., 1]$, corresponds to a value of θ equal to 0.9 and the posterior density concentrates around this value.

Now, the same analysis developed so far is repeated, but the peak areas are modeled with a conditional-Gaussian distribution. Table 7.8 shows the logarithm of the likelihood ratios in favour of $H_0 : s\&v$ and in favour of $H_: s\&u$ (last row) obtained applying both conditional-Gamma model and conditional-Gaussian approximation. As it is shown, these values are almost the same with the maximum difference given by a factor 1.07 for the comparison $H_0 : s\&v$ versus $H_1 : v\&u$ and therefore, it can be inferred that the conditional-Gaussian model approximates the conditional-Gaussian model in a consistent way.

θ	$Log_{10}LR$
0	0
0.1	-inf
0.2	-236.24
0.3	-151.45
0.4	-67.57
0.5	11.81
0.6	16.63
0.7	17.11
0.8	17.07
0.9	16.84
1	0

Table 7.5: *Lago* data, 2-person mixture - logarithm of the likelihood ratio in favour of $H_0 : s\&v$ vs. $H_1 : 2u$ as function of θ for a 10:1 mixture.

	$Log_{10}LR$
Alleles	3.40
Areas	7.44

Table 7.6: *Lago* data, 2-person mixture - logarithms of the likelihood ratio in favour of $H_0 : s\&u$ vs. $H_1 : 2u$ when only evidence on repeat number of the alleles are available, and when peak are information is added.

7.2.3 Separation of genotypes

In this section the mixed DNA profile only is assumed to be available and the main aim is to predict the genotypes of the unknown individuals who contributed to the mixture. This might be of interest, for example, in order to compare the separated profile to those in a given database of DNA profiles. Here, a way for separating mixtures based on the same network applied to profile-identification is suggested. For the sake of simplicity, it is assumed that it is a 2-person mixture and it is discriminated between the processing for the separation of one profile only and the processing for separating both profiles. As it could be expected, the first situation is clearly the easiest case to deal with, because the genotype of the other contributor to the mixture is known.

θ	$Log_{10}LR$
0	0
0.1	-inf
0.2	-235.61
0.3	-154.57
0.4	-73.32
0.5	4.05
0.6	7.65
0.7	7.81
0.8	7.74
0.9	7.58
1	0

Table 7.7: *Lago* data, 2-person mixture - logarithm of the likelihood ratio in favour of H_0 : $s\&u$ vs. H_1 : $2u$ assuming known mixture proportion for a 10:1 mixture.

Hypotheses	Gamma model	CG model
s&v vs. 2u	16.77	16.78
s&v vs. s&u	7.74	7.74
s&v vs. v&u	9.30	9.33
s&u vs. 2u	7.44	7.44

Table 7.8: *Lago* data, 2-person mixture - comparison of the logarithms of the likelihood ratio in favour of suspect and victim and in favour of suspect and an unknown individual when peak area are modeled with a conditional-Gamma model and when are modeled with a Normal approximation.

The problem of separating a mixture into two components is now approached. The evidence is represented by peak area and repeat number information on the mixture, whilst no profiles from identified individuals are examined. In this scenario, if information on pre-amplification proportion of DNA in the sample is not available, identifiability problems in assigning genotype combinations to each person occur. This is due to the fact that there is symmetry between the two individuals `p1` and `p2` and the separated genotypes could be assigned indifferently to both persons. This issue is overcome as breaking the symmetry, *i.e.* entering evidence that the pre-

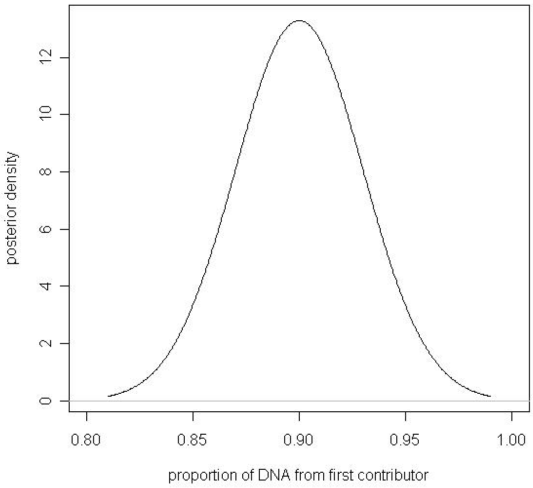
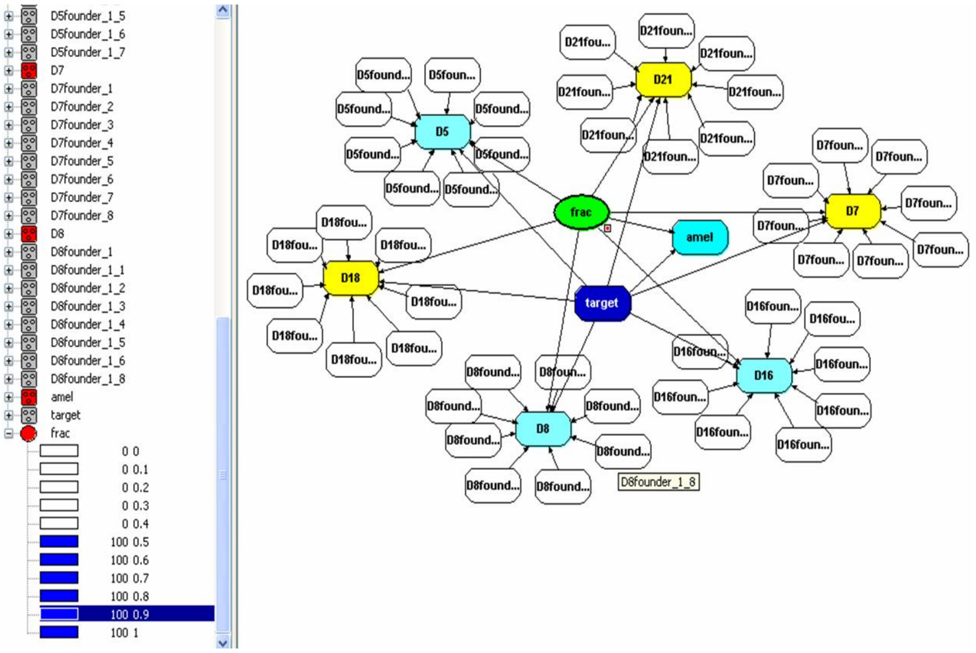


Figure 7.4: Lago data - posterior density of the proportion of DNA from the major contributor for 10:1 2-person mixture.

amplification proportion of DNA generated by p1 is at least one half of the total DNA in the sample. In the network this is done by entering likelihood vector in the node frac which represents the DNA proportion θ contributed by p1. In particular, the likelihood vector is set zero when θ is in the range $[0, 0.1, 0.2, ..., 0.4]$, and one when θ is in the range $[0.5, 0.6, ..., 1]$ (see Figure 7.5). Viceversa, indifferently the hypothesis that p1 contributed at most half of the DNA to the mixture sample might be used.

Table 7.9 shows the predicted genotypes for both individuals. Here, the predictive posterior probabilities associated to each genotype shows that both profiles are predicted with extremely high probability. Furthermore, since a-priori the true profiles of the contributors are known, it is concluded that all of the markers are correctly identified.

Supposed that the genotype of one of the contributors (*e.g.* the victim) is measured; the attention is now focused on the prediction of the genotype of the other contributor. Thus, the evidence is given by the composition of the crime trace and from victim's profile and it is entered into the appropriate nodes contained in the classes **Amean**, **Bmean**, **Cmean**, and **xmean**



```
D5founder_1_5
D5founder_1_6
D5founder_1_7
D7
D7founder_1
D7founder_2
D7founder_3
D7founder_4
D7founder_5
D7founder_6
D7founder_7
D7founder_8
D8
D8founder_1
D8founder_1_1
D8founder_1_2
D8founder_1_3
D8founder_1_4
D8founder_1_5
D8founder_1_6
D8founder_1_8
amel
target
frac
      0 0
      0 0.1
      0 0.2
      0 0.3
      0 0.4
    100 0.5
    100 0.6
    100 0.7
    100 0.8
    100 0.9
    100 1
```

Figure 7.5: Lago data - master network representing the node `frac` which indicates the DNA proportion θ contributed by `p1`. The symmetry is broken by setting likelihood in the node `frac`.

referred to the mixture composition, the class **vgt** referred to the victim's profile (see Figure 7.6). Furthermore, both peak area and repeat number information are used. Table 7.10 displays the predicted genotypes that are read in the node `jointgt` which states are the aggregation of the genotypes of the two contributors `p1` and `p2`. (Cowell *et al.* 2007b) showed that, when one of the genotypes is available, the prediction of the other profile is more accurate than in the event that they are both unknown. Lago data does not show this result since the probabilities obtained to predict the genotype of the contributor `p1` are the same when the victim's profile is known and when no profiles are available. This is due to the fact that in this mixture peak areas are extremely informative and the knowledge of the victim's profile does not improve the prediction.

If both the unknown profiles or a single unknown profile are predicted using a CG approximation for the peak areas, then tables 7.11 and 7.12 show the

Marker	Genotype p1	Genotype p2	Probability
Amelogenin	X Y	X Y	0.8388
D5	9 11	11 11	0.9357
D7	9 10	8 11	≈ 1
D8	10 14	12 14	0.9404
D16	11 11	10 12	≈ 1
D18	12 17	15 21	≈ 1
D21	28 29	31.2 32.2	≈ 1

Table 7.9: *Lago* data, 2-person mixture - predicted genotypes of both contributors. Evidence consists of the mixed trace.

Marker	Genotype p1	Probability
Amelogenin	X Y	0.8388
D5	11 11	0.9357
D7	8 11	≈ 1
D8	12 14	0.9404
D16	10 12	≈ 1
D18	15 21	≈ 1
D21	31.2 32.2	≈ 1

Table 7.10: *Lago* data, 2-person mixture- predicted genotype of the suspect knowing victim's profile. Evidence consists of the mixed trace and DNA profile extracted from the victim, v.

same predicted profiles which are obtained applying a conditional-Gamma distribution. The posterior probabilities for each profile, obtained applying both two models, are similar, thus it is concluded that both conditional-Gamma and conditional-Gaussian models appear to perform well.

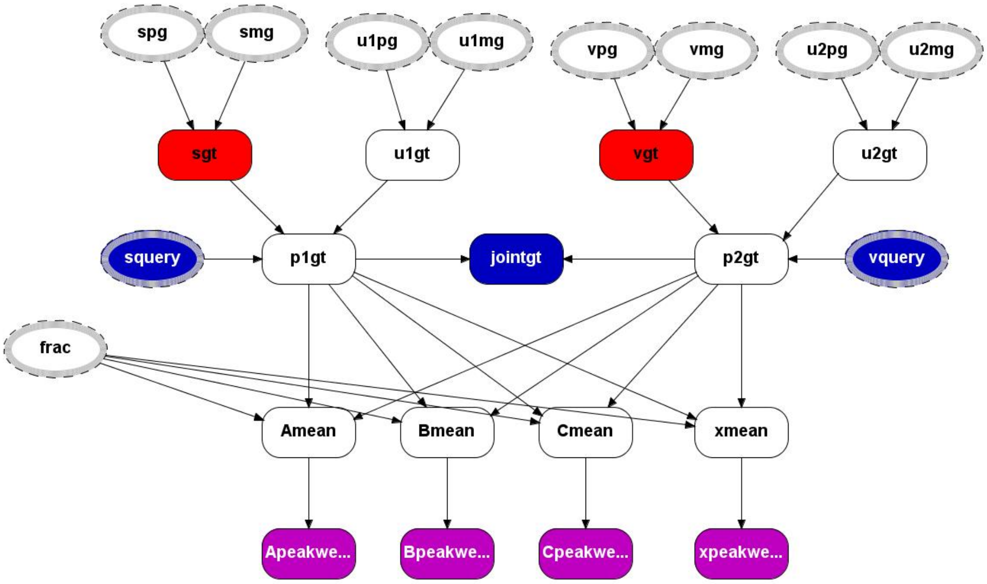


Figure 7.6: Two person mixture. Marker class.

Marker	Genotype p1	Genotype p2	Gamma model	CG model
Amelogenin	X Y	X Y	0.8388	0.8429
D5	9 11	11 11	0.9357	0.9379
D7	9 10	8 11	≈ 1	≈ 1
D8	10 14	12 14	0.9404	0.9520
D16	11 11	10 12	≈ 1	≈ 1
D18	12 17	15 21	≈ 1	≈ 1
D21	28 29	31.2 32.2	≈ 1	≈ 1

Table 7.11: *Lago* data, 2-person mixture - predicted genotype of both contributors when peak areas are modeled with a conditional-Gamma distribution and when are modeled with a CG approximation. Evidence consists of the mixed trace.

Marker	Genotype	Gamma model	CG model
Amelogenin	X Y	0.8388	0.8429
D5	11 11	0.9357	0.9379
D7	8 11	≈ 1	≈ 1
D8	12 14	0.9404	0.9520
D16	10 12	≈ 1	≈ 1
D18	15 21	≈ 1	≈ 1
D21	31.2 32.2	≈ 1	≈ 1

Table 7.12: *Lago* data, 2-person mixture - predicted genotype of of the suspect knowing victim's profile when peak areas are modeled with a conditional-Gamma distribution and when are modeled with a CG approximation. Evidence consists of the mixed trace and DNA profile extracted from the victim, v.

Chapter 8

Analysis of two DNA mixed traces

Relative simple modifications of the networks described in Appendix A can allow the simultaneous analysis of a couple of DNA mixed traces. Although this joint analysis may look like purely speculative, it can have important applications, such as for the analysis of multiple samples which occur when a DNA sample is amplified a number of times providing different results because, for example, the sample is degraded or the DNA proportion of one of the contributors is too low. In this chapter we solve an identification and separation problem for the genotype of two suspects, termed *s1* and *s2*. In particular, we consider a robbery case where some tools, used for breaking into an apartment, have been handled by more than one individual. Furthermore, we suppose to be interested into two specific traces that we analyse simultaneously. In particular in § 8.1.1 we solve a problem of identification when alleles' repeat number only is available, whilst in § 8.2.1 we add peak areas as evidence. Thus, we show the contribution to the peak areas, since sometimes an investigation based on alleles' repeat number only can lead to erroneous inference, whereas the inclusion of the peak area information in the analysis gives the correct result. Finally, in § 8.2.2 we provide a prediction of the genotypes of the contributors to the mixtures.

8.1 Two mixed traces analysis using allele's repeat number information

8.1.1 Identification of the suspects' genotypes

In this section we show how to use the statistical model exploiting the alleles' repeat number information only, that has been implemented through the OOBN described in Appendix B.1. The network is an extension of that introduced by Mortera *et al.* (2003), which is applied to solve identification problems for 2-person mixtures in one trace only. This network allows to enter as evidence alleles repeat number information only and, furthermore, to make inference on the total number of contributors. Actually, although this network has the form of a Bayesian network, it has been modified to an OOBN in Appendix A.1.

We analyse simultaneously two mixed traces for a robbery case as the one described above. We suppose to examine two mixed traces each one containing biological material from two individuals. In this scenario, the evidence is represented by two mixtures and DNA profiles from two suspects, *s1* and *s2*.

Data are reported in Table 8.1. This table shows the alleles observed in both traces, called *Trace1* and *Trace2*, the measured relative peak weights[1] and the genotypes of the two suspects. We take into account markers Amelogenin, D2, D21, FGA, THO1 and vWA.

If we observe carefully the composition of the mixtures, by a preliminary investigation, we note that in the marker D2 the allele with repeat number 21, present in the first trace, is not observed in the second trace. Similarly, in the marker D21 the allele with repeat number 32.2, present in the second trace, is not observed in the first one. Thus, considering the alleles' repeat number information only, from markers D2 and D21 we note that the two mixtures are different. Furthermore, observing also the genotypes of the suspects, in the marker D2 the allele with repeat number 21, observed in the

[1] We denote by *Rel.Weights1* the relative peak weights referred to the first trace, and *Rel.Weights2* the relative peak weights measured in the second trace.

Marker	**Trace1**	**Rel. Weight1**	**Trace2**	**Rel. Weight2**	**Suspect1**	**Suspect2**
Amelogenin	X	0.6147	X	0.4950	X	X
	Y	0.3853	Y	0.5050	Y	Y
D2	19	0.5112	19	0.4338	19	
	20	0.3792	20	0.4949	20	20
	21	0.0486				
	23	0.0610	23	0.0712		23
D21	28	0.5017	28	0.5163	28	28
	30	0.4983	30	0.4152	30	
			32.2	0.0685		32.2
FGA	22	0.3963	22	0.5791	22	22
	23	0.6037	23	0.4209	23	
THO1	9.3	1	9.3	1	9.3	9.3
vWA	14	0.4918	14	0.3801	14	
	18	0.0885	18	0.1164		18
	19	0.4197	19	0.5035	19	19

Table 8.1: *Lago* data, two traces 2-person mixtures - two 2-individuals mixture compositions with relative peak weights, suspect1's and suspect2's genotypes.

first mixture, is not contained in any genotype of the suspects, whilst in the marker D21 the allele with repeat number 32.2, contained in the genotype of the second suspect, is not observed in the first trace. Thus, it can be assumed that an unknown individual contributed to the first trace and that the second suspect *s2* did not contribute to the first trace. On the contrary, since *s1* seems to be compatible with both traces, we can assume that the first suspect contributed to both.

Now, we develop a preliminary analysis on the total number of contributors to the mixtures. The evidence on allelic repeat number of both mixtures is entered in the appropriate nodes contained in the classes `A_in_T1`, `B_in_T1`, `A_in_T2`, `B_in_T2`, etc. of the **marker** network. On the contrary, the evidence on allelic repeat number of the two suspects' genotypes is entered in the appropriate nodes contained in the classes `s1gt` and `s2gt` of the **marker** network (see Figure 8.1). The normalized likelihoods reported in Table 8.2

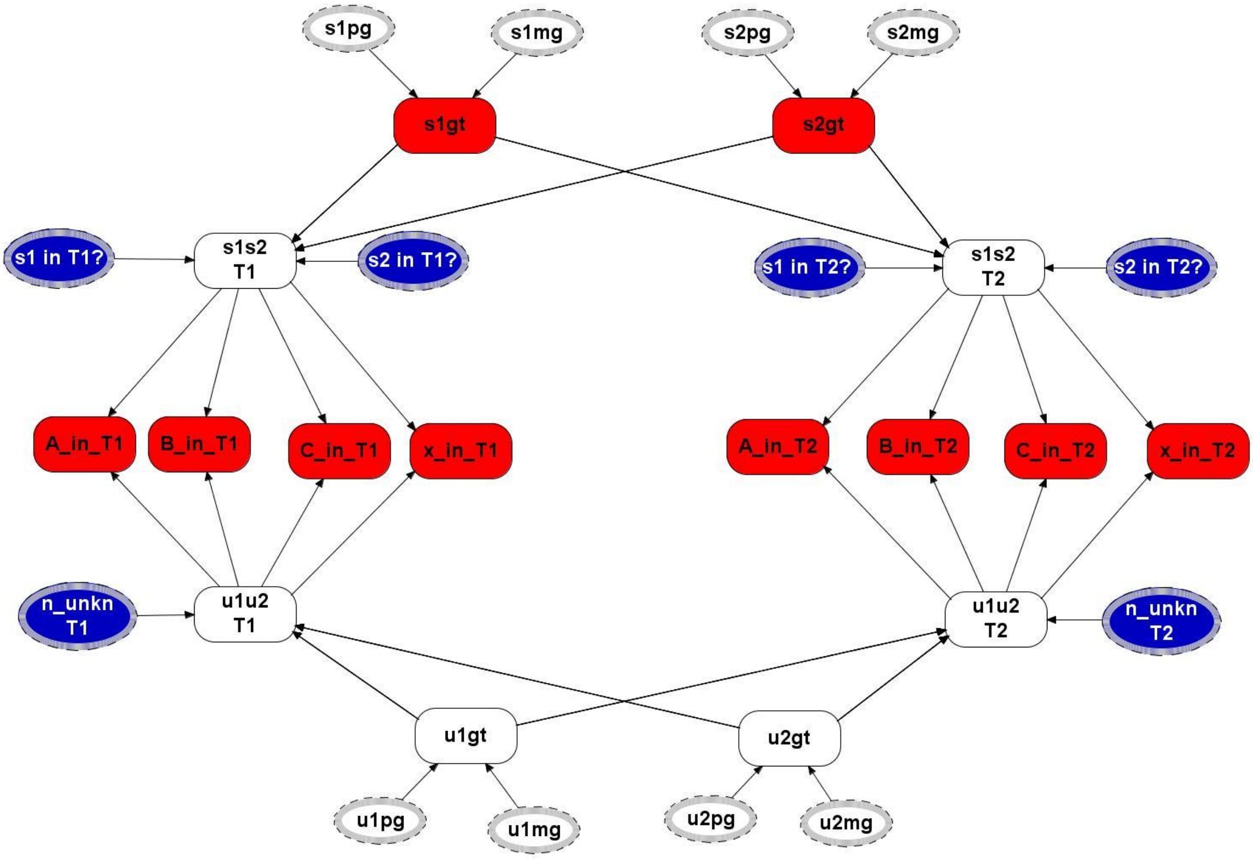


Figure 8.1: Two traces. Marker class.

are read in the nodes total_#_T1 and total_#_T2 in the **target** network (see Figure 8.2). Details of the networks are given in Appendix B.

The normalized likelihoods, associated to the hypotheses that 2 is the total number of contributors, is 0.9997 for the first trace, and 0.9999 for the second one. Negligible likelihoods are associated to a number of three contributors. Furthermore, for the hypotheses of a total number of contributors less than 2 the normalized likelihoods are zero. In effect, whereas in the marker D2 we observe four alleles in the first trace and three in the second one, and each individual can possesses at most two alleles, we can assume that there are at least two contributors in both mixed traces.

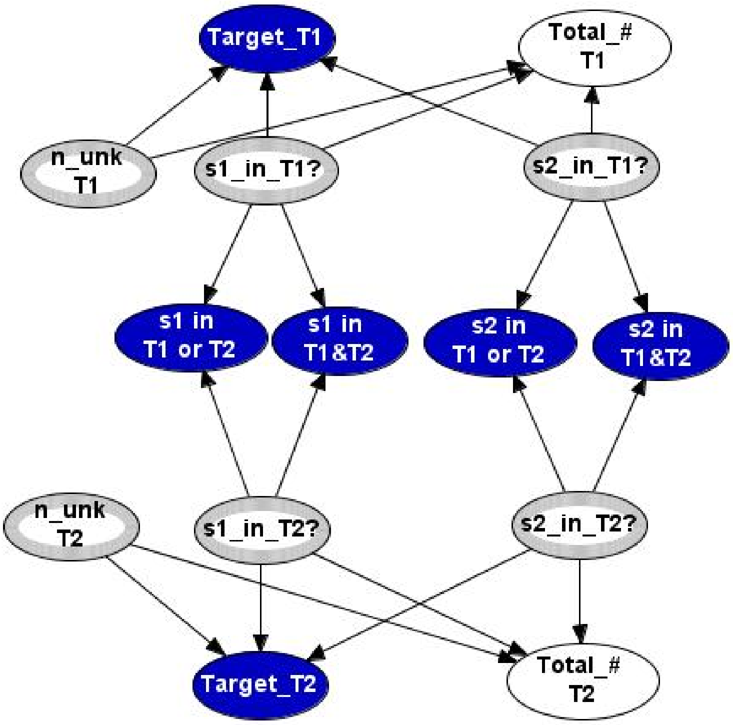


Figure 8.2: Two traces. Target class.

It is worth noting that, since we assume uniform priors (see § 6.2), the equality between the normalized likelihoods and the posterior probabilities holds.

Now, we can carry on our analysis by assuming that the total number of contributors is two in both traces and therefore by constraining both the

Number contributors	Normalized Likelihoods	
	Trace 1	Trace 2
0	0	0
1	0	0
2	0.9997	0.9999
3	0.0003	0.0001
4	0	0

Table 8.2: *Lago* data, two traces - normalized likelihoods of the total number of contributors for both traces.

nodes `total_#_T1` and `total_#_T2` to be equal two.

Table 8.3 shows the normalized likelihoods for the hypotheses displayed in the first column obtained entering as evidence alleles' repeat number only. Such hypotheses form the states of the nodes `target_T1` and `target_T2` in the **target** class. In the first column we observe a high normalized likelihood associated to the hypothesis that the first suspect and an unknown individual contributed to the first mixed trace. On the contrary, the weight of evidence is against both suspects in the second trace.

	Normalized Likelihoods	
Hypotheses	Trace1	Trace2
s1&s2	≈ 0	0.9999
s1&u	0.9999	≈ 0
s2&u	≈ 0	0.000292
2u	0.001477	≈ 0

Table 8.3: *Lago* data, two traces - normalized Likelihoods for the tested hypotheses in both traces `T1` and `T2` and the second one for identification problems when only allele's repeat number information is used. Evidence consists of the mixed traces and DNA profiles extracted from two suspects, *s1* and *s2*.

However, in a courtroom we could be mainly interested in investigating whether each suspect contributed to both or at least one crime trace. Table 8.4 shows the normalized likelihoods for the hypotheses in the first column when only alleles' repeat numbers information is used. Generally, the normalized likelihoods are high for almost all the hypotheses with the exception that (i) `s2` contributed to the first trace (`s2 in T1`) and that (ii) `s2` contributed to both traces (`s2 in T1&T2`). In effect, in the previous Table 8.3, column "Trace1" shows a high likelihood associated to the hypothesis $s1\&u$, where the second suspect `s2` does not appear.

Hypotheses	Normalized Likelihoods
s1 in T1	0.9999
s1 in T2	0.9999
s2 in T1	≈ 0
s2 in T2	≈ 1
s1 in T1 or T2	≈ 1
s1 in T1&T2	0.9998
s2 in T1 or T2	≈ 1
s2 in T1&T2	≈ 0

Table 8.4: *Lago* data, two traces - normalized likelihoods against the suspects for identification problems when only allele's repeat number information is used. Evidence consists of the mixed trace and DNA profiles extracted from two suspects, *s1* and *s2*.

8.2 Two mixed traces analysis adding peak ares information

In this section we carry on our analysis introducing peak area information. We discriminate between an analysis for identifying the genotypes of the suspects and for predicting the genotypes of the contributors to the two mixtures. We built an OOBN for two mixed traces based on both conditional-Gamma and conditional-Gaussian distributions for the peak areas (see § 7.2.1). The exact same network, which details are given in Appendix B.2, can be used both for identification and as well as for separation problems, without any further modification. Furthermore, whereas both traces are 2-person mixtures, the conditional-Gamma and the conditional-Gaussian model for the peak weights are the same as those introduced in § 7.2.1.

8.2.1 Identification of the suspects' genotypes

In this section the evidence consists of DNA extracted from the suspects *s1* and *s2*, the allelic repeat numbers and relative peak weights in both traces. After converting peak areas in normalized weights and calculating the likelihood vectors for the conditional-Gamma model (see equation (7.7) in § 7.2.1), we enter them as evidence in the relevant nodes in the classes **Amean_T1**, **Bmean_T1**, **Amean_T2**, **Bmean_T2**, etc. of the **marker** network. On the contrary, evidence on the suspects' genotypes is entered in the nodes `gt` of the classes **s1gt** and **s2gt** in the **marker** network (see Figure 8.3). This is described in detail in Appendix B.2.

Table 8.5 shows the normalized likelihoods for the hypotheses displayed in the first column. Such hypotheses form the states of the nodes `target_T1` and `target_T2` in Figure 8.4 The normalized likelihoods in the "Alleles" columns are obtained entering as evidence alleles' repeat numbers only, whilst the normalized likelihoods displayed in the "Areas" columns are obtained adding peak area information. It is worth noting that the values displayed in the columns denoted "Alleles" are the same as those reported in the Table 8.3 in the previous section. For the first trace we note high likelihoods, equal to 0.9999, associated to the hypothesis that the first suspect and an unknown contaminator contributed to the mixture. This value is obtained

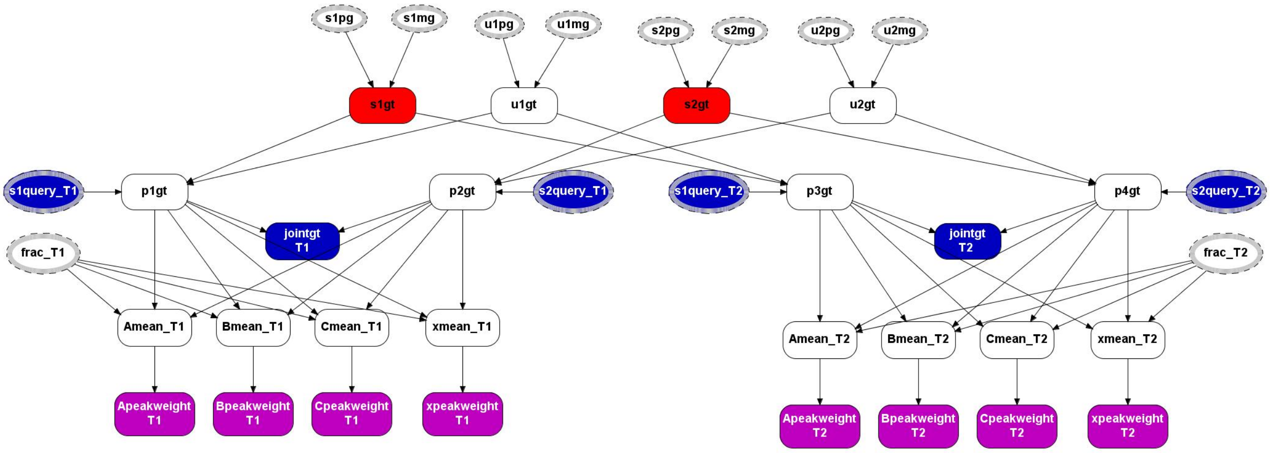


Figure 8.3: Two traces. Marker class.

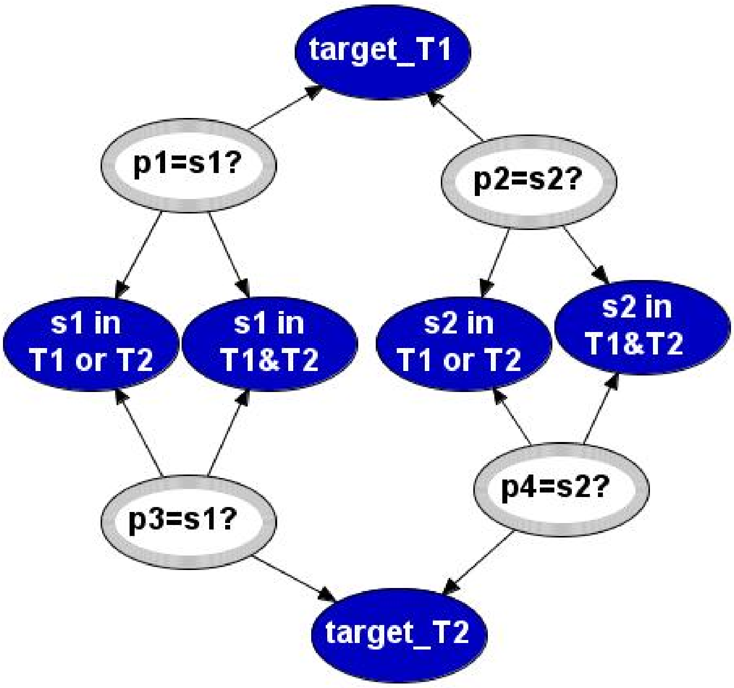


Figure 8.4: Two traces. Target class.

when (i) the evidence is given by alleles' repeat numbers only and (ii) is given by peak area as well. On the contrary, for the second trace we observe high likelihoods, equal to 0.9999, associated to the hypothesis that both suspects contributed to the mixture. Also in the second trace, this value is obtained when considering the alleles' repeat numbers alone and thereafter in combination with the peak area information.

We now investigate whether each suspect contributed to both or at least one crime trace. Table 8.6 shows the normalized likelihoods for the hypotheses in the first column. The normalized likelihoods for the "Alleles" columns are the same as those displayed in Table 8.4 in the previous section, when the analysis is based on alleles' repeat numbers only. As well, those displayed in the column denoted "Areas" are obtained by adding peak area information. In Table 8.6 we note high likelihoods with the exception of the hypotheses `s2 in T1` and `s2 in T1&T2`, where these values approach zero.

Furthermore, it is worth noting that, in the Tables 8.5 and 8.6, the

	Trace1		Trace2	
Hypotheses	Alleles	Areas	Alleles	Areas
s1&s2	≈ 0	≈ 0	0.9999	0.9999
s1&u	0.9999	0.9999	≈ 0	≈ 0
s2&u	≈ 0	≈ 0	0.000292	0.0000202
2u	0.001477	0.0000304	≈ 0	≈ 0

Table 8.5: *Lago* data, two traces - normalized likelihoods for the tested hypotheses in both traces `T1` and `T2` for identification problems when only allele's repeat number information is used (Alleles) and when peak area information is added (Areas). Evidence consists of the mixed trace and DNA profiles extracted from two suspects, *s1* and *s2*.

normalized likelihoods, obtained when alleles' repeat number only are entered as evidence, are very similar to those obtained when peak area information is added. This result is due to the fact that the likelihoods reported in the columns "Alleles" are extremely high for some hypotheses, *i.e.* close to unity. As a consequence, the additional information introduced by the peak ares can overall be considered negligible, compared to the one delivered by the repeat numbers. In particular, in this case the markers D2 and D21 in the mixture are extremely meaningful, whereas the two traces for these markers have one different allele[2] and, therefore we can conclude that an unknown individual contributed to the first trace and that the first trace itself is incompatible with the genotype of the second suspect.

Excluding from the analysis the markers D2 and D21, *i.e.* disregarding the two most meaningful sources of information, and taking into account the alleles' repeat number only, the two traces seems to be equal. Table 8.7 shows the logLR on base 10 in favour of the two suspects for both traces `T1` and `T2`, discriminating the two cases when alleles' repeat number information only is used ("Alleles") and when peak area information is added ("Areas"). Firstly, it is worth noting that the values displayed in both the "Alleles" columns are the same, since, observing the allelic repeat numbers only, the two mixed traces seem to be equal. Thus we obtain the same results. Furthermore, for

[2] In the marker D2 the second trace does not contain the allele with repeat number 21, whilst in the marker D21 the allele 32.2 is not contained in the first trace

	Normalized likelihoods	
Hypotheses	Alleles	Areas
s1 in T1	0.9999	0.9999
s1 in T2	0.9999	0.9999
s2 in T1	≈ 0	≈ 0
s2 in T2	≈ 1	≈ 1
s1 in T1 or T2	≈ 1	≈ 1
s1 in T1&T2	0.9998	0.9998
s2 in T1 or T2	≈ 1	≈ 1
s2 in T1&T2	≈ 0	≈ 0

Table 8.6: *Lago* data, two traces - posterior probabilities against the suspects for identification problems when only allele's repeat number information is used (Alleles) and when peak area information is added (Areas). Evidence consists of the mixed trace and DNA profiles extracted from two suspects, *s1* and *s2*.

	Trace1		Trace2	
Hypotheses	Alleles	Areas	Alleles	Areas
s1&s2 vs. 2u	5.95	2.83	5.95	6.64
s1&s2 vs. s1&u	2.62	-1.21	2.62	2.61
s1&s2 vs. s2&u	3.02	4.08	3.02	4.22

Table 8.7: *Lago* data, two traces - logarithms of the LR in favour of suspects in both traces `T1` and `T2` for identification problems when alleles repeat number information only is used (Alleles) and when peak area information is added (Areas). Evidence consists of the mixed trace and DNA profiles extracted from two suspects, *s1* and *s2*. Markers D2 and D21 have been excluded from the analysis.

the first trace, the inclusion of peak area information changes the likelihood ratios. If the evidence consists of the alleles' repeat number only, since the likelihood ratio is equal to $10^{5.95} \simeq 882,000$ the weight of evidence is against both suspects, whilst, when peak area information is added, we note a negative logLR associated to the comparison $H_0 : s1\&s2$ versus $H_1 : s1\&u$. A negative logLR has meaning that the alternative hypothesis $H_1 : s1\&u$ is greater than the null hypothesis $H_0 : s1\&s2$. Thus, the weight of evidence is now against the first suspect and an unknown individual. In effect, if we consider the Table 8.8, where the normalized likelihoods for the hypotheses in the first column are obtained by excluding markers D2 and D21 from the analysis, when peak area information is included, the normalized likelihood that the second suspect contributed to the first trace, `s2 in T1`, decreases from 0.9976 to 0.0578, corresponding approximately to a factor 17.25. As a consequence, also the probability that the second suspect contributed to both traces decreases, since it is the logical conjunction $\{$*s2 in T1*$\} \cap \{$*s2 in T2*$\}$, and it changes from ≈ 1 to 0.06.

Table 8.9 shows the logLR in favour of the first suspect and an unknown individual for the first trace and when peak area information is used. In effect, the logLR in the table are all positive showing a weight of evidence against the first suspect and an unknown individual. Furthermore, since *s2* is not a contributor to the mixture, the highest ratio, $10^{5.3} \simeq 197,500$, is associated to the hypotheses $H_0 : s1\&u$ versus $H_1 : s2\&u$. Actually, whereas the DNA mixture has been produced in laboratory, we know that the contributors are the first suspect and an unknown individual (see § 5). Thus, using the information provided by alleles' repeat number only, the resulting likelihood ratios lead to erroneous inference as the weight of evidence is in favour also of the second suspect. However, it has here been confirmed that the inclusion of the peak area allows to recover this loss in performance.

	Normalized Likelihoods	
Hypotheses	Alleles	Areas
s1 in T1	0.9999	0.9999
s1 in T2	0.9999	0.9999
s2 in T1	0.9976	0.0578
s2 in T2	0.9976	0.9976
s1 in T1 or T2	≈ 1	≈ 1
s1 in T1&T2	≈ 1	≈ 1
s2 in T1 or T2	≈ 1	≈ 1
s2 in T1&T2	≈ 1	0.06

Table 8.8: *Lago* data, two traces - normalized likelihoods against the suspects for identification problems when only allele's repeat number information is used (Alleles) and when peak area information is added (Areas). Evidence consists of the mixed trace and DNA profiles extracted from two suspects, *s1* and *s2*. Markers D2 and D21 have been excluded from the analysis.

Now, consider the "Areas" column for the second trace in Table 8.7.

Hypotheses	Trace1
s1&u vs. 2u	4.05
s1&u vs. s2&u	5.30
s1&u vs. s1&s2	1.21

Table 8.9: *Lago* data, two traces - logarithms of the LR in favour of the first suspect `s1` and an unknown individual in the first trace `T1` for identification problems when peak area information is added (Areas). Evidence consists of the mixed trace and DNA profiles extracted from two suspects, *s1* and *s2*. Markers D2 and D21 have been excluded from the analysis.

When peak area information is added as element of evidence the logLR increases for the comparisons H_0 : $s1\&s2$ versus H_1 : $2u$ and H_0 : $s1\&s2$ versus H_1 : $s2\&u$, and the increase corresponds respectively to a factor approximated to 5 and 16. On the contrary, for the hypotheses H_0 : $s1\&s2$ versus H_1 : $s1\&u$ the logLR has a small decrease corresponding to a factor 0.99. This is due to the fact that peak areas do not add information on the second suspect.

8.2.2 Separation of mixtures

In this section the same networks as in the identification task (with peak areas) are here applied for predicting the genotypes of contributors to the mixtures. We suppose to observe two mixed traces containing DNA genotypes from two individuals. In particular, we suppose to observe the alleles' repeat numbers and the peak areas associated to each allele. Furthermore, we consider the case of separation of both unknown profiles. Thus, no information concerning the two contributors to the mixtures is available. As discussed in § 7.2.3, it is recalled that separation of mixtures, when both contributors are unknown, is only possible when the contributions to the DNA mixtures has taken place in quite different proportions. As a consequence, we need to break the symmetry between the individuals p1 and p2 in the first trace and p3 and p4 in the second trace. Thus, we enter the evidence that the pre-amplification proportion of DNA in the sample from individual p1 (and p3) is at least one half of the total DNA in the sample[3]. Using the network described in Appendix B.2, breaking symmetry is obtained by setting likelihood evidence to be more (or less) than 0.5 in the nodes `frac_T1` and `frac_T2` in Figure 8.5, *i.e.* the posterior distribution of both nodes is zero for values lower (or grater) than 0.5. We note that, when we enter the evidence that the proportion of DNA θ originating from, p1 in the first trace and from p3 in the second trace, is more (or less) than 0.5, automatically the proportion of DNA originating from, p2 in the first trace and from p4 in the second trace, is set to be less (or more) than 0.5, since this is defined to be equal to$1 - \theta$.

Figure 8.6 shows the posterior distribution of the mixture proportion θ

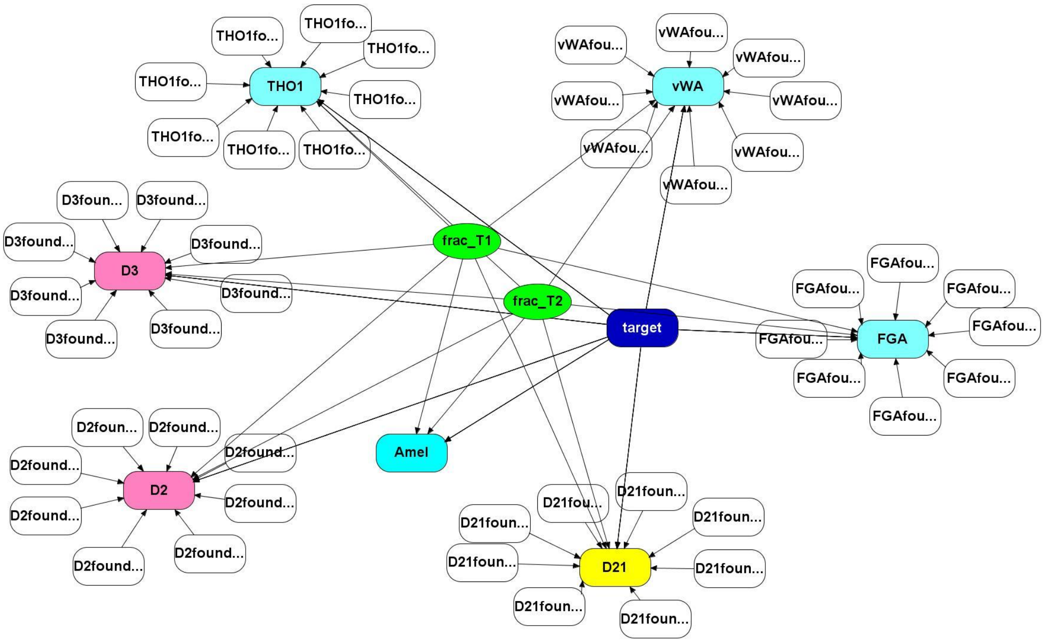


Figure 8.5: Two traces. Target class.

for the contributors p1 and p3. The posterior distribution for the proportion originated from p1 is represented as a solid curve, whilst the posterior distribution for the proportion originated from p3 is represented as a broken line.

[3] Equally, the symmetry breaking could be achieved assuming that p1 (and p3) contributed at most half of the DNA to the mixture sample.

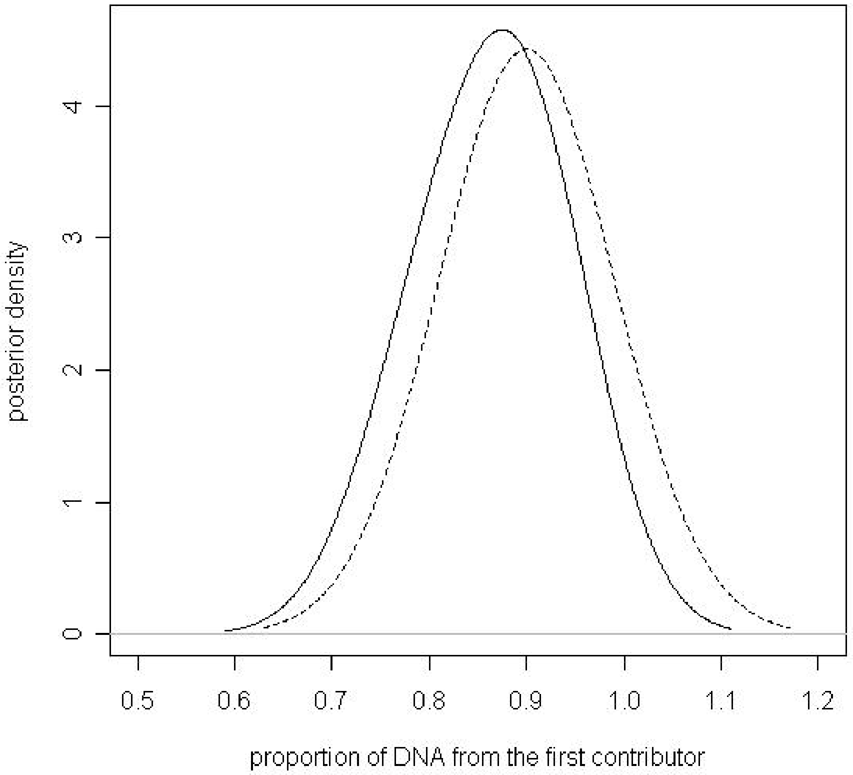


Figure 8.6: Lago data - posterior density of the proportion of DNA from the major contributor for the first trace (solid curve) and the second trace (broken line). The exact proportions of the mixtures are 5 : 1 for the first trace and 10 : 1 for the second one which, in a scale ranging [0 : 1] correspond to 0.83 and 0.91.

They both have a maximum around 0.9, although this is a little bit lower for p1, and they both are zero for $\theta < 0.6$. Since the distribution of θ for p1 is closer zero than the distribution of θ for p3, we can conclude that θ for p1 is smaller than θ for p3. In effect, the exact proportions of the mixtures are $5 : 1$ for the first trace and $10 : 1$ for the second one, which in a scale ranging $[0 : 1]$ corresponds to 0.83 and 0.91.

The predicted genotypes of the four contributors (p1 and p2 for the first trace, and p3 and p4 for the second trace) are shown in Table 8.10. The predicted profiles are correct for all contributors with high posterior probabilities. In particular, we note an extremely high probability, approximated to unity, for the genotypes of individuals `p1` and `p2` at marker D2. This is due to the fact that the prediction ability increases for loci with a high incidence of heterozygotes. In effect, for this marker the contributors `p1` and `p2` are two heterozygous individual who do not share any allele. On the contrary, the lowest posterior probability, 0.6356, is associated to the profiles of the contributors `p3` and `p4` at marker FGA since `p4` is a homozygote. Additionally, also at marker THO1 the prediction has an extremely high posterior probability since at this marker one allele only is observed and therefore the predicted profiles are immediate.

Marker	Genotype p1	Genotype p2	Prob.	Genotype p3	Genotype p4	Prob.
Amelogenin	X X	X Y	0.9315	X Y	X Y	0.9013
D2	19 20	21 23	≈ 1	19 20	20 23	0.7965
D21	28 30	28 30	0.8760	28 30	28 32.2	0.8514
FGA	22 23	23 23	0.7944	22 23	22 22	0.6356
THO1	9.3 9.3	9.3 9.3	≈ 1	9.3 9.3	9.3 9.3	≈ 1
vWA	14 19	14 18	0.8246	14 19	18 19	0.8959

Table 8.10: *Lago* data, two traces - predicted genotypes of all contributors.

We model now peak areas with a conditional-Gaussian approximation. Tables 8.11 and 8.12 show the predicted genotypes for the contributors p1 and p2 for the first trace, and p3 and p4 for the second trace. The predicted genotypes are equal to those obtained modelling peak areas with a conditional-Gamma distribution. The last two columns of the two tables report the posterior probabilities for the predicted genotypes obtained using the conditional-Gamma model and the conditional-Gaussian model. Such probabilities, computed applying both models, are similar, thus it is concluded that both conditional-Gamma and conditional-Gaussian models appear to perform well.

Trace1				
Marker	Genotype p1	Genotype p2	Gamma model	CG model
Amelogenin	X X	X Y	0.9315	0.9312
D2	19 20	21 23	≈ 1	≈ 1
D21	28 30	28 30	0.8760	0.8814
FGA	22 23	23 23	0.7944	0.7968
THO1	9.3 9.3	9.3 9.3	≈ 1	≈ 1
vWA	14 19	14 18	0.8246	0.8255

Table 8.11: *Lago* data, two traces - predicted genotypes of all contributors in the first trace `T1` when peak areas are modeled with a Gamma distribution and when are modeled with a Normal approximation.

Trace2				
Marker	Genotype p3	Genotype p4	Gamma model	CG model
Amelogenin	X Y	X Y	0.9013	90.64
D2	19 20	20 23	0.7965	0.8128
D21	28 30	28 32.2	0.8514	0.8741
FGA	22 23	22 22	0.6356	0.6440
THO1	9.3 9.3	9.3 9.3	≈ 1	≈ 1
vWA	14 19	18 19	0.8959	0.8917

Table 8.12: *Lago* data, two traces - predicted genotypes of all contributors in the second trace `T2` when peak areas are modeled with a Gamma distribution and when are modeled with a Normal approximation.

Chapter 9

Identification of DNA mixtures involving more than two contributors

This chapter is concerned with the analysis of mixed traces where more than two individuals may have contributed to a DNA sample left at a crime scene. In particular, for the sake of simplicity, we consider a mixed trace comprising DNA from three individuals only.

In § 9.1 we describe the 3-person mixture model and compare the conditional Gamma model and the conditional-Gaussian model. In particular, we show the differences with the 2-person mixture model. In § 9.2 we discuss the advantages and disadvantages of each network developing a comparison of their efficiency. In § 9.3 we calculate an upper bound limit for the total number of unknown contributors to be included in the example analysed. In § 9.4 we discriminate among three different situations. In the first case we consider, for example, a rape case where a sample contains biological material from the victim and two perpetrators. Thus, the evidence consists of a mixed trace and DNA profiles extracted from a victim, v, and two suspects, *s1* and *s2*. The main aim is to compare the available genotypes and the mixture in order to determine whether the individuals, whose genotypes are observed, contributed to the mixture. In the second case we investigate the event when an *unknown contaminator* is present. For example, we consider the case of a scuffle (or a brawl) during which a person is killed. In this case, the evidence consists of a mixed trace and DNA profiles extracted from *s1* and *s2* only. Finally, we analyze the case of DNA profiles extracted from one of the suspects, *e.g. s1*.

Our analysis is concerned with forensic identification problems using both information provided by allele's repeat number and quantitative peak areas. Our aims are: (i) to show the efficiency of both networks, (ii) to solve identification problems and (iii) to show that peak weights need to be taken into account since they increase the likelihoods. Furthermore, although the analysis is performed using a conditional-Gamma model for the relative peak weights, we provide the results when we apply a conditional-Gaussian model in order to show that the latter is a good approximation of the conditional-Gamma model. Finally, in § 9.5 we explain the reasons why the analysis for separating the genotypes of the contributors could not been performed.

9.1 Model assumptions

We present a description of the 3-person model before analysing our data. We assume that the mixture is made up of DNA from three persons, who we refer to as `p1`, `p2` and `p3`. The sample before the amplification consists of an unknown number of cells from `p1`, an unknown number of cells from `p2` and a further unknown number of cells from `p3`, where every cell contains exactly two alleles[1] for each marker. Now, let θ_1 be the proportion of cells from `p1`, θ_2 be the proportion of cells from `p2` and θ_3 be the proportion of cells from `p3`. Thus, these quantities, θ_1, θ_2 and θ_3, represent the pre-amplification proportions of DNA from each contributor, and we assume them to be constant across markers.

Details on the post-amplification proportions of alleles for each markers are given in § 7.2.1.

[1] which are different for heterozygote and the same for homozygote

The conditional-Gamma model for a 3-person mixture

The conditional-Gamma model for the peak weights has been widely discussed in § 7.2.1. In this section we extend the model for a 3-person mixture providing the main definitions.

Assume that: (i) there are 3 potential contributors to the mixture; (ii) the analysis of the mixture is based on M markers with generic marker $m = 1, 2, ..., M$ having a_m allelic type. As shown in detail in § 7.2.1, W^m_{ia} denotes the contribution of individual i to the peak weight at allele a of marker m. This has a Gamma distribution as follows:

$$W^m_{ia} \sim \Gamma(\rho_m \gamma \theta_i n^m_{ia}, \eta_m).$$

Also the total weights W_{+a} of a single allele a at marker m in the mixture have a Gamma distribution

$$W^m_{+a} \sim \Gamma(2\rho_m \mu^m_a \gamma, \eta_m),$$

where, in general,

$$\mu^m_a = \frac{\sum_i \theta_i n^m_{ia}}{2}, \tag{9.1}$$

and where θ_i is the DNA proportion from individual i, n^m_{ia} is the number of alleles with repeat number a possessed by person i at marker m, ρ_m is an amplification factor and η_m is a scale parameter. Since we assumed that there are 3 potential contributors to the mixture, *i.e.* $i = 1, 2, 3$, the mean in (9.1) becomes

$$\mu^m_a = \frac{\theta_1 n^m_{1a} + \theta_2 n^m_{2a} + \theta_3 n^m_{3a}}{2}. \tag{9.2}$$

Whereas the quantities θ_i represent DNA proportions, their sum must be 1, *i.e.*

$$\theta_1 + \theta_2 + \theta_3 = 1. \tag{9.3}$$

This allows to express the DNA proportion θ_3 from the third contributor `p3` as difference from the sum of the other proportions, *i.e.* $\theta_3 = 1 - \theta_1 - \theta_2$.

Thus, the mean in (9.2) becomes

$$\mu_a = \frac{\theta_1 n_a^{(1)} + \theta_2 n_a^{(2)} + (1 - \theta_1 - \theta_2) n_a^{(3)}}{2}. \tag{9.4}$$

Furthermore, the condition that

$$\theta_1 \geq \theta_2 \geq \theta_3 \tag{9.5}$$

must be verified where the labeling 1, 2, 3 is exchangeable. Thus, if this condition is verified, θ_1 represents the DNA proportion originated by the first major contributor, θ_2 the DNA proportion originated by the second major contributor, and θ_3 the DNA proportion originated by the minor contributor.

The peak weight is here reported in terms of relative values in order to avoid arbitrariness in its scaling. Thus,

$$R_a = \frac{W_{+a}}{W_{++}},$$

where

$$R_a \sim Dir(2\rho\mu_a\gamma),$$

with mean and variance as follow

$$E[R_a] = \mu_a,$$

where μ_a is defined in equation (9.4), and

$$V[R_a] = \sigma^2\mu_a(1 - \mu_a),$$

and where the dependence on marker m has been dropped.
See § 7.2.1 for more details.

The conditional-Gaussian model for a 3-person mixture

This model, based on conditional-Gamma distribution for the absolute scaled peak weights, can be approximate with a conditional-Gaussian (CG) model.

Thus, the model assumes a Gaussian distribution for the relative peak weights R_a

$$R_a \sim \mathcal{N}(\mu_a, \tau_a^2),$$

where μ_a is the same as in equation (9.4) and represents the pre-amplification proportion for allele a of the marker it belongs to, and τ_a^2 has the form in equation (7.9) in § 7.2.1

In the network model we consider the relative peak weights in order to avoid arbitrariness in scaling. But these relative peak weights must sum to unity, thus their correlation must be taken into account. The conditional-Gaussian distribution ignores this correlation, but Cowell *et al.* (2007b) proved that using instead the distribution

$$R_a \sim N(\mu_a, \sigma^2 \mu_a)$$

this problem can be overcome (proof is given in § 7.2.1).

9.2 Advantages and disadvantages of each network and their efficiency

In Appendix C we describe two networks in detail. The first network is an extension, for 3-person mixtures, of the network introduced by Mortera *et al.* (2003). It models DNA mixtures using alleles' repeat number information only and furthermore is used to make inference on the total number of contributors. The second network is an extension, for 3-person mixtures, of the network introduced by Cowell *et al.* (2007b). It models DNA mixtures using both alleles' repeat number and peak area information. We will not use it to make inference on the total number of contributors, even if this should be possible by adding appropriate nodes referred to the total number of contributors.

As a check, we must obtain the same posterior probabilities for the hypotheses under test when we apply both networks and include as evidence alleles' repeat number only.

The advantage of the second extended network is that it allows to introduce peak area information in addition to alleles' repeat number. This is an important result since we will show that the inclusion of peak areas strengths the weight of evidence. However, the disadvantage is that this network is computationally complex, in particular is much more complex in comparison to the other network.

Table 9.1 shows the total clique size tables for the main classes in both networks. Column "Alleles" displays the total clique size tables for the classes in the network that can employ alleles repeat number information only. Column "Areas" displays the total clique size tables for the classes in the network that can also include peak area information. For all the classes considered, except for the class **target**, the total clique size tables are greater in the second network than in the first one. For example, for the class **alleleinmix** the increase of the total clique size table in the second network is huge and corresponds approximately to a factor $3,800$.

Classes	Alleles	Areas
marker	184,264	3,020,787
amelogenin	960	20,516
target	80	80
alleleinmix	8	30,693
total	557,512	34,997,698

Table 9.1: Total clique size tables for the classes marker, amelogenin, target, alleleinmix and the total in the network extended, for 3-person mixtures, from the one introduced by Mortera *et al.* (2003) (column "Alleles") and in the network extended from the one introduced by Cowell *et al.* (2007b) (column "Areas").

In the last row we report the total clique size table for the entire network. This corresponds to the total clique size table of the **master** class since it is the top level and therefore contains instances of all the other classes. For the entire network the total clique size table in the second network increases approximately to a factor 62.

Therefore, the results obtained in Table 9.1 allow to conclude that the network extended for 3-person mixtures and introduced by Cowell *et al.*

(2007b) is computationally more complex but answering a much more complex problem and we will see in detail that this complexity represents a strong limitation of computer for the analysis of 3-person mixtures.

9.3 Bounding the total number of contributors

Consider a rape case where a sample contains biological material from the victim and multiple perpetrators. Assuming two perpetrators, a mixed DNA trace from three contributors is to be examined. Additionally, we measure the genotypes of three potential contributors, here named victim v and two suspects, *s1* and *s2*.

As it is impossible to evaluate the strength of evidence for all possible numbers of unknown contributors, it could be of interest to identify an upper bound limit on the unknown number of contributors. Lauritzen and Mortera (2002) derived an inequality for the probability of observing a given DNA profile when the bound of unknown individuals contributing to the mixture is assumed to be fixed. Then, they showed how to use this inequality to obtain an upper bound limit for the unknown number of contributors needed to be considered. We apply this rule to determine an upper bound on the unknown number of contributors for the forensic case analysed in this chapter.

9.3.1 Theoretical aspects

In a crime case where the evidence is given by a DNA profile for a mixed stain from two or more persons, the weight of evidence cannot be derived for all possible contributors. In general, although the evidence of the trace itself determines a lower bound limit observing the maximum number of alleles in any marker, we cannot be sure on the upper bound. However, Lauritzen and Mortera (2002) showed a way to identify an upper bound b on the unknown number of contributors.

For a hypothesis H involving x unknown individuals, the following likelihood is considered

$$\mathrm{P}_x(\mathcal{E}|H) = \mathrm{P}(\mathcal{E}_m = U_m \cup K_m|H) \qquad \forall \quad m = 1, ..., M, \tag{9.6}$$

where $\mathcal{E}_m$ is the observed evidence profile at marker m given by the observed set of alleles at a marker m on a total number of markers M, U_m is the observed set of alleles supplied by the unknown individuals and K_m is the alleles carried by the known individuals.

Thus, the likelihood ratio in favour of a hypothesis $H_0 : s1\&s2$ against an alternative hypothesis $H_1 : s1\&u$ is

$$\frac{\mathrm{P}_{x_0}(\mathcal{E}|H_0)}{\mathrm{P}_{x_1}(\mathcal{E}|H_1)}, \tag{9.7}$$

where x_i is the number of unknown individuals involved in the hypothesis H_i. In a court case, the defendant should be given the highest assumption of innocence. As a consequence, we should look for the minimum value of the likelihood ratio, which is equivalent to seeking an upper bound limit for the denominator of the LR in (9.7).

Thus, using the fact that $\mathrm{P}_x(\mathcal{E}|H)$ is smaller than the probability that all the alleles of the unknown contributors match those in $\mathcal{E}$, then

$$\mathrm{P}_x(\mathcal{E}|H) \leq \mathrm{P}(U_m \subseteq \mathcal{E}_m|H), \quad \forall \quad m = 1, ..., M. \tag{9.8}$$

Assuming that all the unknown individuals come from the same population and that the M markers are independent, the equation (9.8) becomes

$$\mathrm{P}_x(\mathcal{E}|H) \leq \prod_{m=1}^{M} \left(\sum_{a \in \mathcal{E}_m} p_a^m \right)^{2x}, \quad \forall \quad m = 1, ..., M, \tag{9.9}$$

where p_a^m is the frequency of allele a at marker m. When the evidence profile contains all the possible alleles for all markers, this product is one and therefore useless. On the contrary, if the evidence is represented by some alleles only, the product in equation (9.9) tends to zero at an exponential rate, representing a bound for the probability of observing the given evidence.

Now, there exists a generic specific number y such that

$$\prod_{m=1}^{M}\left(\sum_{a\in\mathcal{E}_m} p_a^m\right)^{2x} < y. \tag{9.10}$$

Inverting equation (9.10) and taking logs we can obtain x lower limit by the following function of y

$$x > b(y) = \frac{\ln y}{2\sum_{m=1}^{M}\ln(\sum_{a\in\mathcal{E}} p_a^m)}.$$

Thus,

$$x > b(y) \Rightarrow \mathrm{P}_x(\mathcal{E}|H) \leq y. \tag{9.11}$$

Now, if we assume that $y = \mathrm{P}_{x_1}(\mathcal{E}|H_1)$ and that the number of unknown contributors x_i, involved in a given hypothesis H_i, is greater than $b(y)$, then this hypothesis H_i is less likely than H_1 and therefore it does not need to be considered.

9.3.2 Bounding the number of contributors for the 3-person mixture analysed in § 9.4

In this section we apply this bound limit at the data in Table 9.3 in the next section.

It is supposed that the evidence under the hypothesis H_0:$s1\&s2\&v$ is available in the form of DNA profiles from the victim v and the two suspects *s1* and *s2* and that under this hypothesis the probability of the evidence is one. On the contrary, under the alternative hypothesis H_1:$s1\&v\&u$, it is assumed that the mixture consists of the profiles from the victim, one suspect, *e.g. s1*, and a single unknown contributor *u1*. Table 9.2 shows, for each marker, all the possible genotypes of *u1* given that the mixture consists of the profiles of *v*, *s1* and *u1*. Its associated probabilities are displayed in the row below the possible genotypes. The last column shows, for each marker, the probability of observing the given evidence under the hypothesis H_1. This is obtained as the sum of the probabilities associated to each profile. Multiplying over all markers the probability $P_{x_1}(\mathcal{E}|H_1) = 0.000637$ is obtained.

The denominator of the bound $b(y)$ is now computed as

Marker	Possible genotypes of *u1*							$P_{x_1}(\mathcal{E}\mid H_1)$
D7	(8,8)	(8,9)	(9,8)	(8,10)	(10,8)	(8,11)	(11,8)	
	0.027	0.029	0.029	0.045	0.045	0.029	0.029	0.233
D8	(12,12)	(10,12)	(12,10)	(12,13)	(13,12)	(12,14)	(14,12)	
	0.020	0.014	0.014	0.046	0.046	0.030	0.030	0.200
D21	(31.2,32.2)	(32.2,31.2)	-	-	-	-	-	
	0.007	0.007						0.014

Table 9.2: Bounding the number of contributors.

$$2\sum_{m=1}^{M} \ln\left(\sum_{a\in\mathcal{E}} p_a^m\right) = 2(\ln 0.07761 + \ln 0.792 + \ln 0.539) = -2.208,$$

where the terms of the logarithms are obtained as the sum of the frequencies for the alleles A, B, C and D in the mixture which are reported in Table 5.9 in § 5. Thus,

$$b(0.000637) = \frac{\ln(0.000637)}{-2.208} = 3.33.$$

Whereas 4 is the value of x greater than $b(y)$, we conclude that an alternative hypothesis, for example H^*:$s1\&s2\&v\&4u$, that involves more than 3 unknown individuals, produces a likelihood smaller than $H_1 : s1\&v\&u$ and therefore at most a hypothesis H':$s1\&s2\&v\&3u$ involving three unknown individuals ca be considered. As a consequence, 3 is the maximum number of unknown contributors that is admitted and 6 the maximum number for the total number of contributors. In fact, in the **target** class, the states of the nodes `n_unknown` and `total_#` have been set to a maximum value of, respectively, 3 and 6 (see Figure 9.1).

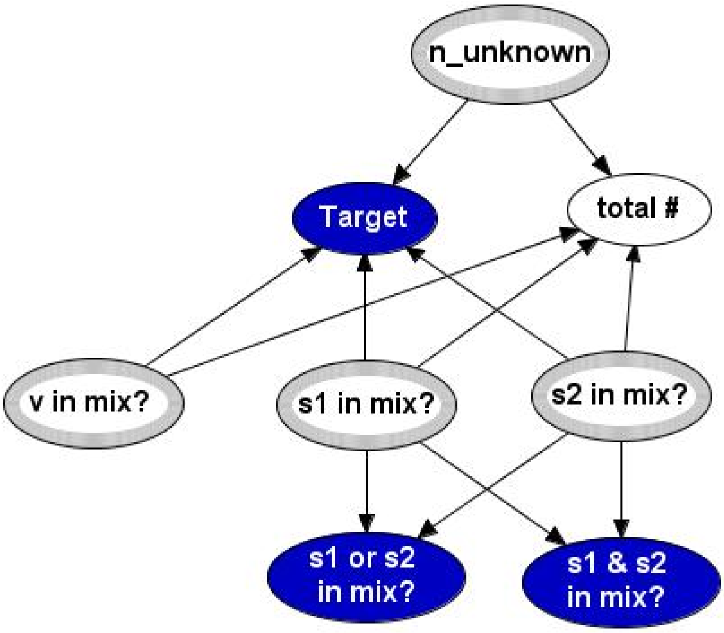


Figure 9.1: Two traces. Target class.

9.4 Forensic identification problems

In this section we solve a forensic identification problem applied to data in Table 9.3. As explained in the previous section we consider a rape case where a sample contains biological material from the victim and two perpetrators. Table 9.3 shows the alleles observed in the mixture, the measured peak areas, the relative weights on 4 markers (Amelogenin, D7, D8 and D21) and the genotypes of v, *s1* and *s2*.

As preliminary analysis, we make inference on the total number of contributors. Table 9.4 displays the normalized likelihoods[2]. Here the evidence is almost entirely in favour of a total number equal three, with a likelihood of 0.884. However, a low normalized likelihood of 0.04 is associated to a total number two. In fact, whereas two is the maximum number of alleles that can be observed for each individual, the presence of four alleles on each marker in the mixed stain suggests that there must have been at least two contributors

[2] Since we assume uniform priors (see § 6.2), the equality between the normalized likelihoods and the posterior probabilities holds.

Marker	**Mixture**	**Rel. Area**	**Rel. Weight**	**Suspect1**	**Suspect2**	**Victim**
Amelogenin	X	44748	0.7760	X	X	X
	Y	33583	0.2240	Y	Y	
D7	8	3785	0.0971		8	
	9	7681	0.2218	9		
	10	12418	0.3984	10		10
	11	8013	0.2828		11	11
D8	10	23256	0.4229	10		10
	12	2676	0.0584		12	
	13	6137	0.1451			13
	14	14673	0.3736	14	14	
D21	28	22272	0.3896	28		28
	29	22766	0.4125	29		29
	31.2	5124	0.0999		31.2	
	32.2	4876	0.0981		32.2	

Table 9.3: *Lago* data, 3-person mixture - a three individuals mixture composition with relative peak areas, relative peak weights, suspects' and victim's genotypes.

Number contributors	Normalized likelihood
0	0
1	0
2	0.040
3	0.884
4	0.102
5	0.011
6	0.001

Table 9.4: *Lago* data, 3-person mixture - normalized likelihoods of the total number of contributors.

to the crime trace.

In particular, if there were two contributors only, the admitted hypotheses would have been: $v\&s1$; $v\&s2$; $s1\&s2$; $v\&u$; $s1\&u$; $s2\&u$; $2u$. In this scenario, the first three hypotheses ($v\&s1$,v&s2, $s1\&s2$) are impossible events. For example, we assume the hypothesis $v\&s1$. Since the profiles for v and *s1* at marker D8 are respectively $(10, 13)$ and $(10, 14)$, the presence of the allele with repeat number 12 in the mixture is not justified. Similarly for the hypotheses $v\&s2$ and $s1\&s2$. Thus, if the total number of contributors were two, then the mixture would have to include at least an unknown individual.

We suppose now to observe the repeat number of the alleles only for the mixture and for the genotypes of the victim and the two suspects. If we apply the network described in Appendix C.1 and we constrain the node `total_#` to be equal three, the node `target` (see Figure 9.1) admits the hypotheses shown in Table 9.5.

However, in a court room, we would be interested in verifying whether the genotypes of both the two suspects only match those of the contributors since we are considering a rape case where the biological sample is taken from the victim and therefore we know that v is a contributor.

Hypotheses under test	
s1&s2&v	both suspects and victim contributed to the mixture
s1&s2&u	both suspects and an unknown individual contributed to the mixture
s1&v&u	the first suspect, the victim and an unknown individual contributed to the mixture
s2&v&u	the second suspect, the victim and an unknown individual contributed to the mixture
s1&2u	the first suspect and two unknown individuals contributed to the mixture
s2&2u	the second suspect and two unknown individuals contributed to the mixture
v&2u	the victim and two unknown individuals contributed to the mixture
3u	three unknown individuals contributed to the mixture

Table 9.5: Hypotheses under test.

Thus, we consider the comparisons in the first column of Table 9.6. This table displays the logarithm on base 10 of the likelihood ratios of the hypotheses in the first column. In the second column, denoted “Alleles”, the logLR are obtained when only the evidence on the repeat number of the alleles is used and when we apply the network described in Appendix C.1. Strong evidence against both the suspects is shown since the highest value $10^{3.47} \simeq 10,700$ is referred to the comparison H_0: $s1\&s2\&v$ vs. H_1: $v\&2u$.

In the third column “Areas” the logLR are obtained when we add peak

	$Log_{10}LR$	
Hypotheses	Alleles	Areas
s1&s2&v vs. v&2u	3.47	7.66
s1&s2&v vs. s1&v&u	3.17	4.00
s1&s2&v vs. s2&v&u	1.36	4.00

Table 9.6: *Lago* data, 3-person mixture - logarithms of the LR in favour of suspects and victim for identification problems when alleles repeat number information only is used (Alleles) and when peak area information is added (Areas). Evidence consists of the mixed trace and DNA profiles extracted from two suspects, *s1* and *s2*, and a victim *v*.

area information and we apply the network described in Appendix C.2. Additionally, we highlight that the relative peak weights have been modelled with the conditional-Gamma model described in § 9.1. Now, if we analyse the contribution of the relative peak weights for each allele, the column "Areas" in Table 9.6 displays that the inclusion of the area information is indeed strengthening the evidence against the suspects whereas the likelihood ratio increases dramatically for all the hypotheses considered. In particular, we note a strong increase in the likelihood ratio involving the hypotheses H_0 :$s1\&s2\&v$ versus H_1 :$v\&2u$ where it changes approximately by a factor $15,000$, indicating that the peak areas of the alleles of the genotypes of the two suspects are extremely informative.

However, in a courtroom context we could be mainly interested in investigating whether at least one or both profiles of the suspects match those contained in the mixture. Thus, as displayed in Table 9.7 column "Alleles", after introducing the evidence in the appropriate nodes and propagating it throughout the network, the node `s1_or_s2_in_mix` returns a high normalized likelihood (approximated to unity) that at least one suspect contributed to the mixture, whilst we obtain a normalized likelihood of 0.9587 that both suspects contributed to the mixture. In effect, if we investigate the single *Boolean nodes* referred to the presence of each single suspect in the mixture, we note a normalized likelihood equal to 0.9598 that *s1* is a contributor, and a normalized likelihood of 0.9989 for *s2*. Actually, these results are as expected since such data are DNA mixtures realized in laboratory, thus a-

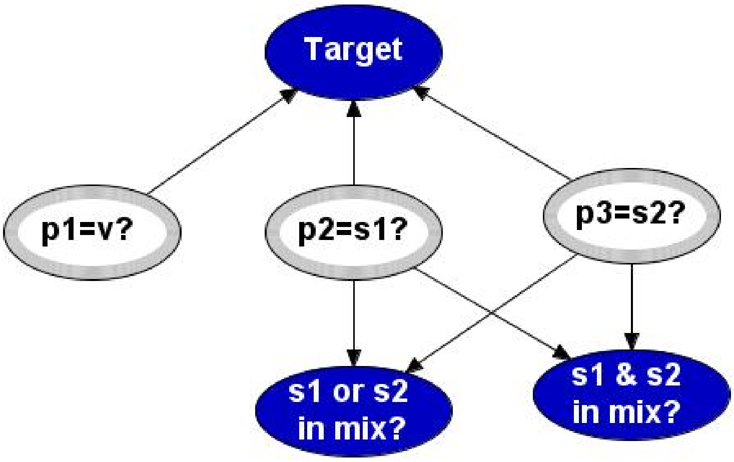


Figure 9.2: Two traces. Target class.

priori the profiles of contributors are known and we know that they match those of the identified individuals (see § 5). See Figure 9.1 for the nodes that represent the hypotheses in Table 9.7 and which are used when alleles' repeat number information only is entered as evidence; whilst, see Figure 9.2 for the same nodes but used when also peak area information is included.

Furthermore, if the peak area information is included in the evidence, all the normalized likelihoods increase, but without differing substantially from the normalized likelihoods displayed in the previous column. This is due to the fact that the likelihoods in column "Alleles" are high, close to unity, thus when peak area information is included they cannot increase significantly.

We suppose now that only genotypes from both suspects are available. This could be common, for example, in a case of a scuffle (or a brawl). Thus, we suppose to investigate a stain of biological material from three assailants and to measure the profiles of two suspects. In this scenario the third contributor is an *unknown contaminator*.

The hypotheses under test are shown in the first column of Table 9.8. Since the highest value $10^{3.39} \simeq 2,500$ is associated to the hypotheses H_0 : $s1\&s2\&u$ versus H_1 :$3u$, we can conclude that the strongest evidence is against both *s1* and *s2*. Additionally, we note stronger evidence against *s2* than *s1*, since the likelihood ratio associated to the comparison H_0 :

	Normalized Likelihoods	
Hypotheses	Alleles	Areas
s1 in mix	0.9598	0.9999
s2 in mix	0.9989	0.9999
s1 or s2 in mix	≈ 1	≈ 1
s1 & s2 in mix	0.9587	0.9998

Table 9.7: *Lago* data, 3-person mixture - normalized likelihoods against the suspects for identification problems when only allele's repeat number information is used (Alleles) and when peak area information is added (Areas). Evidence consists of the mixed trace and DNA profiles extracted from two suspects, *s1* and *s2*, and a victim *v*.

$s1\&s2\&u$ versus H_1 :$s1\&2u$ is greater than the likelihood ratio associated to the hypotheses H_0 :$s1\&s2\&u$ versus H_1 :$s2\&2u$. This result is supported by the normalized likelihoods reported in the first column of Table 9.9 and which are referred to the hypotheses *s1 in mix* and *s2 in mix*. In effect, the normalized likelihood that the second suspect is in the mixture is greater than the normalized likelihood that the first suspect is in the mixture.

The column "Areas" displays the results when peak area information is included to the analysis. Also in this case, the ratios increase dramatically, especially for the comparison H_0 :$s1\&s2\&u$ versus H_1 :$3u$, where the increase corresponds approximately to a factor of 52.

We consider now the normalized likelihoods associated to the hypotheses in Table 9.9. We note high likelihood for all the hypotheses in the first

	$Log_{10}LR$	
Hypotheses	Alleles	Areas
s1&s2&u vs. 3u	3.39	5.10
s1&s2&u vs. s1&2u	2.39	3.40
s1&s2&u vs. s2&2u	1.76	2.42

Table 9.8: *Lago* data, 3-person mixture - logarithms of the LR in favour of the suspects for identification problems when only allele's repeat number information is used (Alleles) and when peak area information is added (Areas). Evidence consists of the mixed trace and DNA profiles extracted from two suspects, *s1* and *s2*.

column. In particular, we note a normalized likelihood of 0.9999 that at least one suspect contributed to the mixture, and a normalized likelihood of 0.9782 that both suspects contributed to the mixture. Furthermore, these likelihoods increase when peak area information is included in the analysis. However, also in this case they increase very slightly compared to the normalized likelihoods displayed in the previous column for the same reason explained above.

On the contrary, we suppose that the evidence comprises the mixed trace and the genotype of one suspect only, for example *s1*. In this scenario we compare the hypotheses H_0: $s1\&2u$ versus H_1: $3u$. The likelihood ratio changes, when we include peak area information in addition to the allele's repeat number, from $10^{1.01} \simeq 10.3$ to $10^{1.77} \simeq 63$ corresponding approximately to a factor of 5.68. Thus, we conclude that peak weights need to be taken into account for identification analyses since they add important information that provide more strength to the weight of evidence.

	Normalized likelihoods	
Hypotheses	Alleles	Areas
s1 in mix	0.9825	0.9962
s2 in mix	0.9956	0.9995
s1 or s2 in mix	0.9999	≈1
s1 & s2 in mix	0.9782	0.9957

Table 9.9: *Lago* data, 3-person mixture - normalized likelihoods against the suspects for identification problems when only allele's repeat number information is used (Alleles) and when peak area information is added (Areas). Evidence consists of the mixed trace and DNA profiles extracted from two suspects, *s1* and *s2*.

So far we have considered three different hypotheses of evidence. In the first scenario the evidence consists of a DNA trace and the profiles of three potential contributors, in the second case we suppose to observe a DNA mixture and the profiles of two suspects only, finally we assume that the evidence comprises a mixture and DNA profile of one suspect only. We note that the weight of evidence decreases as the number of the identified individuals is lower, since we have less available information. Thus, for example, assuming to observe the genotypes of both victim and suspects, we find a weight of evidence against suspects and victim equal to $10^{7.66} \simeq 45,500,000$ (see Table 9.6), whilst, assuming to observe the genotypes of the suspects only, the weight of evidence against the suspects is smaller, $10^{5.1} \simeq 127,000$ (see Table 9.8).

Finally, we compare the results obtained by applying a conditional Gaussian model, described in § 9.1, to those obtained by modelling peak areas with a conditional-Gamma distribution described in the same section. Tables 9.10 and 9.11 show similar logLR and for the last two hypotheses in Table 9.10 they are roughly the same.

Furthermore, we perform an analysis involving the evidence in the form of the mixed trace and DNA profiles from the suspect *s1* only, with

Hypotheses	Gamma model	CG model
s1&s2&v vs. v&2u	7.66	7.63
s1&s2&v vs. s1&v&u	4.00	4.00
s1&s2&v vs. s2&v&u	4.00	4.00

Table 9.10: *Lago* data, 3-person mixture - comparison of the logarithms of the LR in favour of suspects and victim when peak area are modelled with a conditional-Gaussian approximation (CG model) and when are modelled with a conditional-Gamma model (Gamma model). Evidence consists of the mixed trace and DNA profiles extracted from two suspects, *s1* and *s2*, and a victim *v*.

Hypotheses	Gamma model	CG model
s1&s2&u vs. 3u	5.10	5.06
s1&s2&u vs. s1&2u	3.40	3.22
s1&s2&u vs. s2&2u	2.42	2.42

Table 9.11: *Lago* data, 3-person mixture - comparison of the logarithms of the LR in favour of suspects and victim when peak area are modelled with a conditional-Gaussian approximation (CG model) and when are modelled with a conditional-Gamma model (Gamma model). Evidence consists of the mixed trace and DNA profiles extracted from two suspects, *s1* and *s2*.

both conditional-Gaussian and conditional-Gamma models. In this case the likelihood ratio for the hypotheses $H_0 : s1\&2u$ versus $H_1 : 3u$ is $10^{1.80} \simeq 63$, if a conditional-Gaussian approximation is applied, and $10^{1.77} \simeq 58.5$ if peak areas are modelled with a conditional-Gamma distribution. Thus, these results allow to conclude that the conditional-Gaussian approximation is an extremely good approximation to the conditional-Gamma model.

9.5 Separation of genotypes

We did not perform an analysis in order to separate the genotypes of the unknown individuals who contributed to the mixture, since we met severe computational problems.

Table 9.12[3] shows the total clique size tables for the classes that have to be changed in order to introduce a third contributor in the network. The network considered is the one including the evidence on peak areas and described in Appendix C.2. In the second column termed "3mix" the total clique size tables for the classes in the network for 3-person mixtures are

Classes	3mix	2mix	factor
marker	14,411,412	93,180	154.66
joint	11,390,625	50,625	225.00
alleleinmix	30,693	702	43.72
amelogenin	21,356	830	25.73
alleleinmix_am	9,328	280	33.31
joint_am	64	16	4.00
target	80	16	5.00
total	23,491,503	97,214	241.65

Table 9.12: Total clique size tables for the classes in the first column of the network for 2-person mixtures (2-mix) and for 3-person mixtures (3mix), and factor of difference (factor). In the classes termed **alleleinmix_am** and **joint_am** the pedix *am* indicates that the class is used to construct the **Amelogenin marker**.

[3] This table has been constructed computing the total clique size tables of two networks including a single marker node (with alleles A, B, C, D and x) and the Amelogenin in the master class.

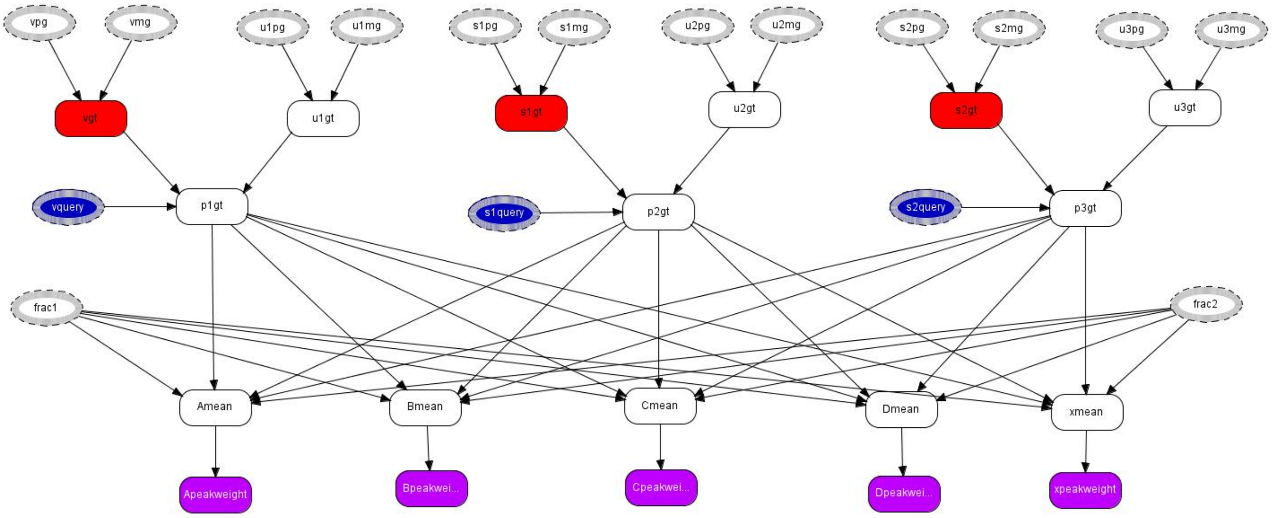


Figure 9.3: Two traces. Marker class.

represented. In the third column termed "2mix" we consider the total clique size tables for the classes in the network for 2-person mixtures. Finally, the last column termed "factor" shows the increase of the total clique size tables in the network for 3-person mixtures.

The clique size tables for the classes in the network for crime trace with three contributors are indeed higher than those in the network for a mixture with two only contributors. A class with a large size is the **marker** class since it is an upper level that contains instances of the other classes. In the network for 3person-mixtures the increase is huge and corresponds approximately to a factor 155. The **marker** class for a 3person-mixture is shown in Figure 9.3. However, the class with highest increase is the **joint** class, where the increase corresponds approximately to a factor 225.

Figure 9.4 shows the **joint** class for a 3-person mixture containing 5 alleles, A, B, C, D and x, where x represents all the other unobserved alleles. In a similar scenario, in the **joint** class each node `p1gt`, `p2gt` and `p3gt` has 15 states given by the aggregation of pair alleles, *i.e.* AA, AB, AC, AD, Ax, BB, BC, etc. Thus, their child node `p1gt&p2gt&p3gt` has a huge size state space equal to 15^3. In general, the clique size table depends both on the number of variables in the clique and the number of states of each variable.

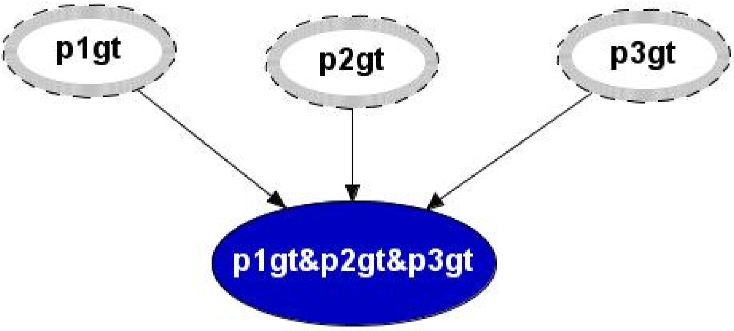


Figure 9.4: Two traces. Joint class.

As a consequence, it is given by the product of the number of the states of each variable in the clique. Therefore, the **joint** class has total clique size table equal to $15^6 \simeq 11,390,625$, since the node `p1gt&p2gt&p3gt` forms a clique with its parents. On the contrary, if we consider the same class in the network for 2-person mixtures, this has total clique size table equal to $15^4 \simeq 50,625$.

Additionally, consider the total in the last row of the table. This is the total clique size table of the entire network and is represented by the **master** class (Figure 9.5 shows the **master** class for a 3person-mixture) which is the top level and therefore contains instances of all the other classes. We note a huge increase corresponding approximately to a factor 240. Thus, we can conclude that the network for 3-person mixtures is computationally much more inefficient than the network for 2-person mixtures, and the class **joint** has the strongest weight.

Because of these severe computational problems we did not perform an analysis for separating the genotypes of the unknown individuals who contributed to the mixture. For the same reason, we performed an identification analysis including the alleles observed in 3 markers and in the Amelogenin as the only evidence, and it has not been possible to consider a higher number of markers. However, researching a method to predict the genotypes in a 3-person mixture and to extend the number of markers in the identification analysis without the risk to meet similar computational problems is an issue for future investigations.

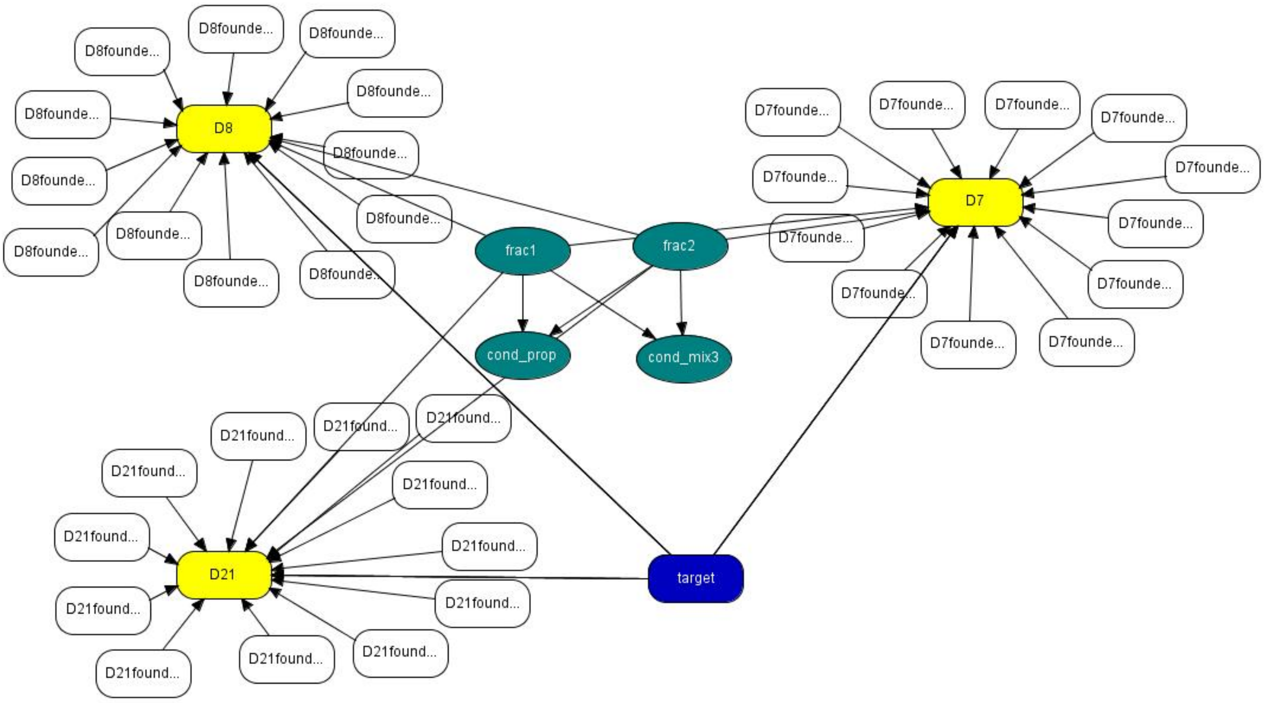


Figure 9.5: Two traces. Master class.

Chapter 10

Conclusions and further investigations

The main aim of this work was to introduce a powerful method for solving complex problems of identification and separation in DNA mixtures. In particular, we approached the issue using Bayesian networks (structured as object-oriented) that can compute numeric likelihood functions. As a consequence of their adaptability and natural flexibility, these networks can be modified to incorporate complications (such as multiple traces and more than two potential contributors) that can characterize the DNA scenario.

The analysis has been developed by considering three different examples in a crime scene. Firstly, we considered a murder where a DNA sample is observed in addition to DNA profiles from a victim and a suspect. Secondly, we took into account multiple traces obtained from a robbery. The power of the analysis carried out for this case is due to the fact that the two traces has been examined simultaneously and using the same network. This allows a mutual exchange of information, since we showed that the weight of evidence loses strength if each trace is analysed singularly. Finally, we considered a rape case where a biological sample from two perpetrators and a victim has been examined.

These examples showed that the predicted peak weights are useful to solve identification and separation DNA problems. In effect, when peak area information is included as evidence, the weight of evidence increases dramatically in all the considered cases and sometimes an investigation based on allele repeat number only can lead to erroneous inference, whereas the inclusion of the peak area information in the analysis gives the correct result. However, the model is extendable to deal with interesting applications to multiple samples. These can occur when the same DNA sample is amplified a number of times providing different results because, for example, the sample is degraded or the DNA proportion of one of the contributors is too low.

Another issue that has been taken into account is the possibility of using a model based on Gamma distributed absolute peak weights. In effect, the authors in Cowell *et al.* (2007b) modelled peak areas using a conditional-Gaussian distribution, but this is an approximation of the quantitative real model for peak areas. Thus, identification and separation problems here have been solved applying a conditional-Gamma distribution for peak areas, taking care of avoiding that Gaussian distributions take negative values. However, the results have been obtained also applying a CG model in order to compare the two models and to show that the CG model is a good approximation.

Unfortunately, in mixtures made up of more than two contributors we met severe computational issues because of an increased complexity due to cliques with huge total size. This problem represented an obstacle for our analysis since we could examine a maximum of three markers only in the mixture (including the Amelogenin) and furthermore we could not predict the genotypes of the three contributors to the mixture. Thus, researching a method to develop a complete analysis of identification and separation of 3-person DNA mixtures overcoming similar computational problems is an issue for future investigations. The methodology of learning in Bayesian networks has many advantages to offer to analysis of DNA mixtures. However, the complexity of models sometimes could require alternative approaches and in this case a solution could be represented by trying to simplify further the network.

Moreover, the analysis is extendable to situations including complications such as drop-out alleles, stutter, etc. The simultaneous analysis of several traces can be useful in presence of such artifacts. In effect, for example, if a drop-out allele is present in a trace, the simultaneous analysis of another mixture that does not contain artifacts, can be useful in recognizing that allelic drop-out. Nevertheless, we hope to pursue this and other aspects in the future.

Acknowledgement

This thesis was supported by *Capitano Gianpietro Lago* and *Tenente Elena Salata* from *Ra.C.I.S.* (Raggruppamento Carabinieri Investigazioni Scientifiche) and by *FSS* (Forensic Science Service) in London for supplying the DNA mixtures. I also thank professor Julia Mortera for her supervision in this thesis.

Bibliography

Butler, J. M. (2005). *Forensic DNA typing*. Elsevier Academic Press, New York.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.

Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2004). Identification and separation of DNA mixtures using peak area information using a probabilistic expert system. Statistical Research Paper 25, Cass Business School, City University.

Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2006). A conditional-gamma bayesian network for dna mixture analyses. *Bayesian Analysis*, **2**.

Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2007a). A gamma model for dna mixture analyses. *Bayesian Analysis*, **2**, (2), 333–48.

Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2007b). Identification and separation of DNA mixtures using peak area information. *Forensic Science International*, **166**, 28–34.

Dawid, A. P., Mortera, J., Pascali, V. L., and Boxel, D. V. (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics*, **29**, (4), 577–95.

Frydenberg, M. (1990). The chain graph markov property. *Scandinavian Journal of Statistics*.

Gill, P., Brenner, C., Buckleton, J., Carracedo, A., Krawxzak, M., Mayr, W., Morling, N., Prinz, M., Schneider, P., and Weir, B.

(2006). Dna commission of the international society of forensic genetics: Recommendations on the interpretation of mixtures. *Forensic Science International*, **160**, 90–101.

Gill, P., Sparkes, R., and Kimpton, C. (1997). Development of guidelines to designate alleles using an str multiplex system. *Forensic Science International*, **89**, 185–97.

Hammersley, J. and Clifford, P. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.

Jensen, F. (2001). *Bayesian Networks and Decision Graphs.* Springer-Verlag, New York.

Koller, D. and Pfeffer, A. (1997). Object-oriented bayesian networks. pp. 302–13. San Francisco.

Laskey, K. B. and Mahoney, S. M. (1997). Network fragments: Representing knowledge for constructing probabilistic models. pp. 334–41. San Francisco.

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

Lauritzen, S. L. (1996). *Graphical Models.* Clarendon Press, Oxford, United Kingdom.

Lauritzen, S. L. and Mortera, J. (2002). Bounding the number of contributors to mixed DNA stains. *Forensic Science International*, **130**, 125–6.

Mortera, J. (2003). Analysis of DNA mixtures using Bayesian networks. In *Highly Structured Stochastic Systems*, (ed. P. J. Green, N. L. Hjort, and S. Richardson), chapter 1B, pp. 39–44. Oxford University Press.

Mortera, J. and Dawid, A. (2007). *Probability and Evidence. In Handbook of Probability Theory with Applications.* T. Rudas. Sage Publications. To appear.

Mortera, J., Dawid, A. P., and Lauritzen, S. L. (2003). Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology*, **63**, 191–205.

Neapolitan, R. E. (2004). *Learning Bayesian Networks*. Pearson Education.

Nilsson, D. and Lauritzen, S. (2000). Evaluating influence diagrams using limids. *Uncertainty in Artificial Intelligence.*

Pearl, J. (1986). A constraint-propagation approach to probabilistic reasoning. *In Uncertainty in Artificial Intelligence.*

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G. (1993). Bayesian analysis in expert systems (with Discussion). **8**, 219–83.

Studený, M. and Bouckaert, R. (1998). On chain graph models for description of conditional independence structures. *Annals of Statistics.*

Tarjan, R. and Yannakakis, M. (1984). Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal on computing.*

Appendices

Appendix A

Details of the object-oriented Bayesian network for 2-person mixtures

In this Appendix we describe the networks used to perform the analyses in chapter 7. All networks have been built using the software Hugin version 6[1].

A.1 OOBN for 2-person mixtures using alleles' repeat number information only

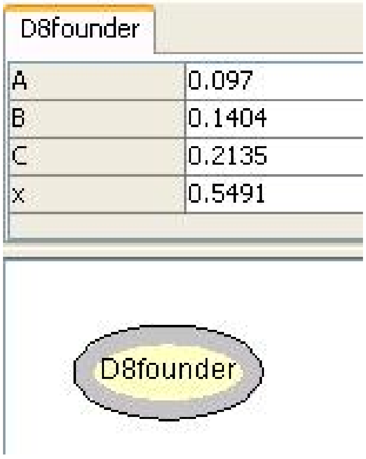
In this section the PES representation used by Mortera *et al.* (2003) to solve forensic identification problems and having the form of a BN has been changed to obtain an OOBN structure. Details of the internal structure are given.

The founder class

The class **founder** contains a single output node. The population gene frequencies in Table 5.5 § 5 are the probabilities associated to the alleles for each marker and that characterize the states of the node. Figure A.1 shows a class **founder** with its probability table; it is referred to the single marker

[1] See www.hugin.com

D8.



D8founder

A	0.097
B	0.1404
C	0.2135
x	0.5491

Figure A.1: Two person mixture. Founder class for marker D8.

The genotype class

The class **genotype** represents an individual's genotype and is shown in Figure A.2. The pair of input nodes `pg` and `mg` are copies of node `founder` of class **founder**. Thus, the (unconditional) distribution of the *founder gene nodes* `pg` and `mg` is specified by the population allele frequencies through the node `founder` in class **founder**. Nodes `pg` and `mg` represent, respectively, the paternal and maternal genes. Now, the genotype of an individual is represented indirectly through a collection of *Boolean nodes*, one for each

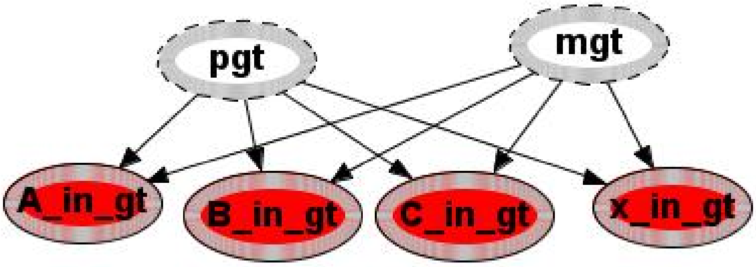


Figure A.2: Two person mixture. Genotype class.

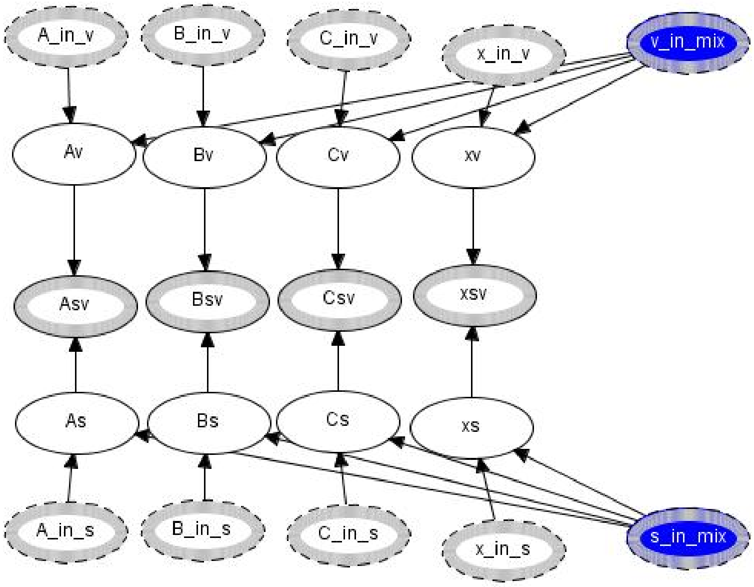


Figure A.3: Two person mixture. Identified genotype class.

allele, termed `A_in_gt`, `B_in_gt`, `C_in_gt` and `x_in_gt`. These nodes are observation nodes and indicates whether or not the individual possesses a specific allele. Their states are given by the logical disjunction[2] of the parents nodes. For example, for the allele A, $A\_in\_gt = \{pg = A\} \cup \{mg = A\}$. This can be translated by the logical expression: ($if(or($`pg`$==$"A", `mg` $==$"A")), *true*, *false*), *i.e.* if either node `pg` or `mg` is A, then node `A_in_gt` is *true*, otherwise is *false*[3].

The identified class

The **identified** class represents the presence in the mixture of a specific allele contributed by at least one of the identified individuals, v and s. This is shown in Figure A.3. The input query node `v_in_mix?` represents the binary query: "is the victim's genotype in the mixture?". Similarly for `s_in_mix?`. The input nodes `A_in_v`, `B_in_v`, `A_in_s`, `B_in_s`, etc. are copies of the similar

[2] Here the term *logical disjunction* indicates the union of two events, $A \cup B$, whilst the term *logical conjunction* indicates their intersection, $A \cap B$.

[3] If a node has function f(A,x,y), then the node takes value x if condition A holds, otherwise takes value y.

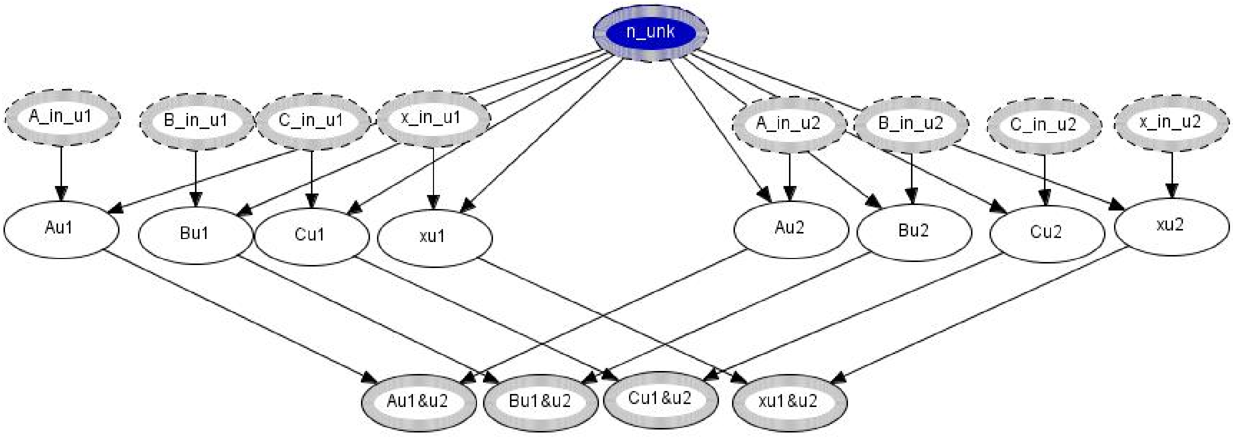


Figure A.4: Two person mixture. Unknown class.

labeled nodes in the class **genotype**. The node `Av` is the logical conjunction `v_in_mix?` $\cap$ `A_in_v`. Thus, its state is *true* if both its parent nodes `v_in_mix?` and `A_in_v` are *true*, *i.e.* if in the mixture there is the allele A contributed by the victim, otherwise its state is *false*. Similarly for the nodes `Bv`, `As`, `Bs`, etc. The node `Avs` indicates the presence of allele A in either the victim or the suspect who contributed to the mixture. Thus, this is *true* if either parent node (`Av` or `As`) is *true* and *false* otherwise.

The unknown class

The **unknown** class represents the presence in the mixture of a specific allele contributed by at least one unknown individual, either *u1* or *u2*. This is shown in Figure A.4. This class is similar to the previous one since both represents the presence in the mixture of a specific allele contributed by at least one individual. Node `n_unknown` specifies the number of unknown individuals in the mixture, therefore has values 0, 1, 2 with same probabilities. The input nodes `A_in_u1`, `B_in_u1`, `A_in_u2`, `B_in_u2`, etc. are copies of the similar labeled nodes in the class **genotype**. Node `Au1` is *true* if `A_in_u1` is *true* and `n_unknown` is either 1 or 2, otherwise is *false*. Similarly for the nodes `Bu1`, `Cu1` and `xu1`. Node `Au2` is *true* if `A_in_u2` is *true* and `n_unknown` is 2, otherwise is *false*. Similarly for the nodes `Bu2`, `Cu2` and `xu2`.

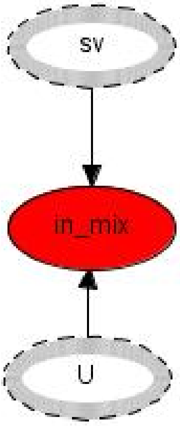


Figure A.5: Two person mixture. Alleleinmix class.

Node `Au1&u2` is the logical disjunction `Au1`∪`Au2`. Thus, its state is *true* if either parent node is *true*, and *false* otherwise.

The alleleinmix class

The class **alleleinmix** represents the composition of the mixture, *i.e.* indicates whether the crime trace contains a certain allelic type. This class is shown in Figure A.5. It contains two *Boolean input nodes* `sv` and `U` which are parents of the observation node `in_mix`. Thus, the node `in_mix` is the logical disjunction of the nodes `sv` and `U`, hence it is *true* if at least one between `sv` and `U` is *true*.

The marker class

The **marker** class represents a specific marker and contains several instances of the classes described so far since it is an upper level. Figure A.6 shows the **marker** class. The query nodes `v_in_mix?` and `s_in_mix?` are *Boolean nodes*. Such nodes indicate whether or not the genotypes of suspect and victim contributed to the mixture, and they have uniform prior probabilities. Node `n_unknown` is the same described in the **unknown** class.

Population allele frequencies, specified in Table 5.5 § 5, define gene

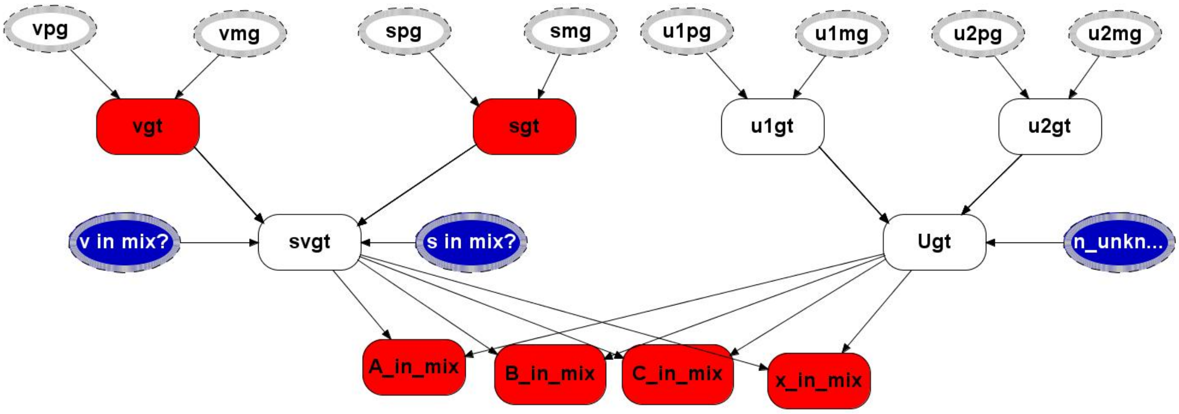


Figure A.6: Two person mixture. Marker class.

nodes `vpg,` `vmg`, `spg`, `smg`, `u1pg`, `u1mg`, `u2pg` and `u2mg`, where, for example `spg` represents the victim's paternal gene, while `smg` is the victim's maternal gene.

Nodes `vgt`, `sgt` `u1gt` and `u2gt` all are instances of the **genotype** class. Evidence on the suspect's and victim's genotype is entered in the network through the nodes `A_in_gt`, `B_in_gt`, `C_in_gt` and `x_in_gt` contained in the instances `vgt` and `sgt`.

The node `svgt` is an instance of the **identified** class, whilst the node `Ugt` is an instance of the class **unknown**.

Nodes `A_in_mix`, `B_in_mix`, `C_in_mix` and `x_in_mix` are all instances of the class **alleleinmix**. Thus, for example, the output node `Asv` in the instance **svgt** is linked to the input node `sv` in the instance **A_in_mix**; whilst, the output node `Au1u2` in the instance **Ugt** is linked to the input node `U` in the instance `A_in_mix`.

The Amelogenin marker class

The **Amelogenin** class is shown in Figure A.7. This class has the same structure of the **marker** class. We show the differences. The nodes `vpg`, `vmg`, `spg`, `smg`, `u1pg`, `u1mg`, `u2pg` and `u2mg` here are not input nodes and they have state space XX for female and XY for male. Thus, no **founder** class is needed. Nodes `vgt`, `sgt`, `u1gt` and `u2gt` are instances of the class **genotype**

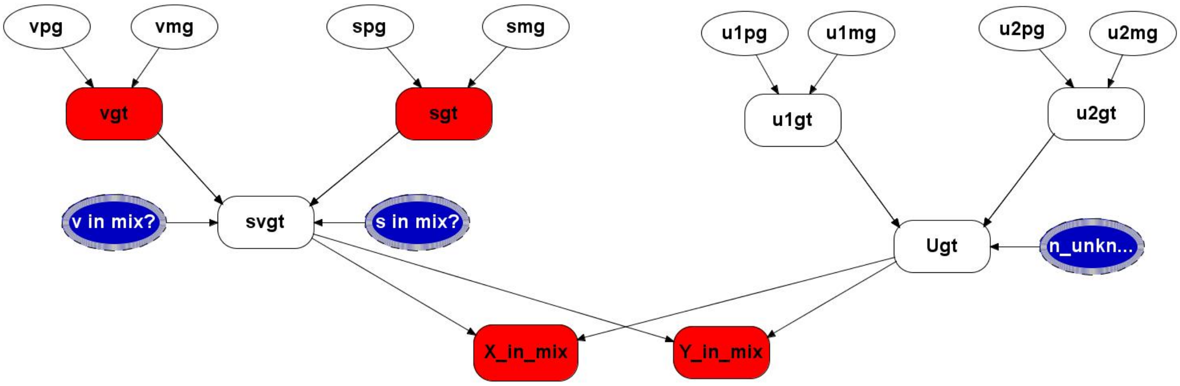


Figure A.7: Two person mixture. Amelogenin marker class.

for the **Amelogenin**. However, the class **genotype** used to build the class **Amelogenin**, has here two *observation nodes* only, termed `X_in_gt` and `Y_in_gt`. Furthermore, the input nodes `pg` and `mg` have state space XX for female and XY for male. The **genotype** class used for the **Amelogenin** is shown in Figure A.8. Furthermore, in the **Amelogenin** nodes `svgt` and `Ugt`

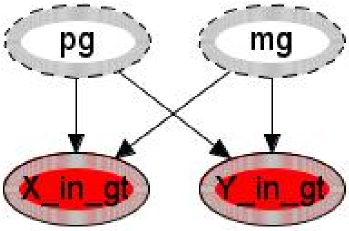


Figure A.8: Two person mixture. Genotype class for Amelogenin marker.

are respectively instances of the classes **identified** and **unknown** used to build the **Amelogenin**. They have the same structure of the corresponding classes described above but with nodes referred to the alleles X and Y rather than the alleles A, B, C, x. These are shown in Figures A.9 and A.10. Finally, the nodes `X_in_mix` and `Y_in_mix` are both instances of the class **alleleinmix**.

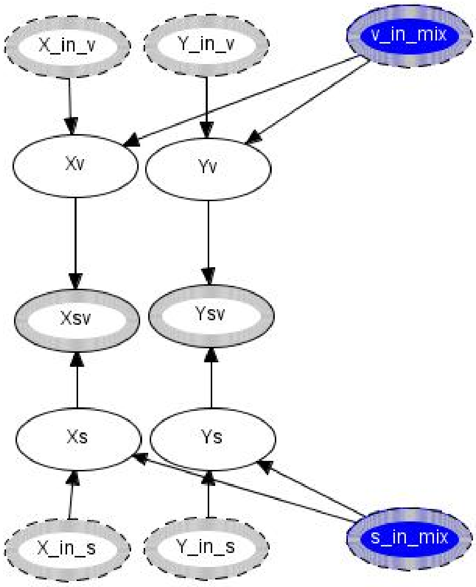


Figure A.9: Two person mixture. Identified class for Amelogenin marker.

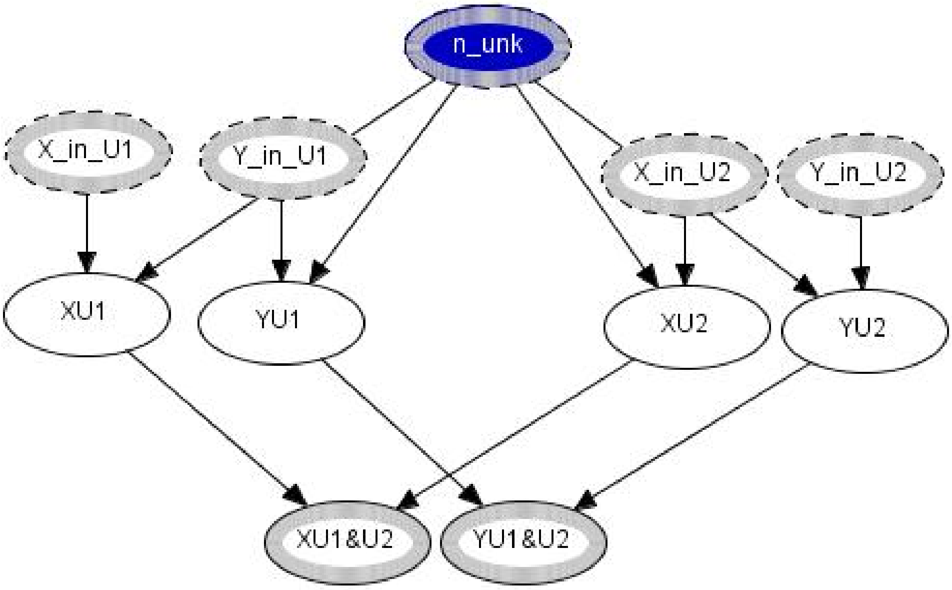


Figure A.10: Two person mixture. Unknown class for Amelogenin marker.

The target class

Figure A.11 shows the **target** class. The **target** class contains the nodes

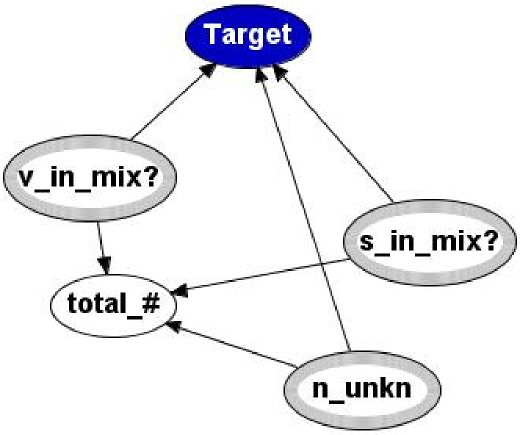


Figure A.11: Two person mixture. Target class.

`v_in_mix?`, `s_in_mix?` and `n_unknown` reported in the **marker** class. These nodes are parents of the node `Target`. The states of the `Target` *query node* represent the several hypotheses under test and are defined by the states of its parent nodes. In other words, its states are made of the aggregation of the states of its parents, being aware that a *false* or a *zero* in the parents' states are not reported in its final state, *e.g.* if the parents' states are `v_in_mix=true`, `s_in_mix=true` and `n_unknown=1`, this node's state will be *v&s&1u*, whilst if they are `v_in_mix=true`, `s_in_mix=false` and `n_unknown=0` its state will be just *v*. Finally, node `total_#` counts all contributors. As a consequence, it has states from 0, if `n_unknown` is 0 and both `s_in_mix` and `v_in_mix` are *false*, to 4, if `n_unknown` is 2 and both `s_in_mix` and `v_in_mix` are *true*.

The master class

The **master** network is shown in Figure A.12. Nodes `D5`, `D8`, `D7`, `D18`, `D16` and `D21` are all instances of **marker** class. `D5`, `D8` and `D16` are markers with three observed alleles A, B and C whilst `D7`, `D18` and `D21` are markers with four observed alleles A, B, C and D. For each marker there are 8 instances of class **founder** linked to the 8 input nodes of the class **marker**. Node

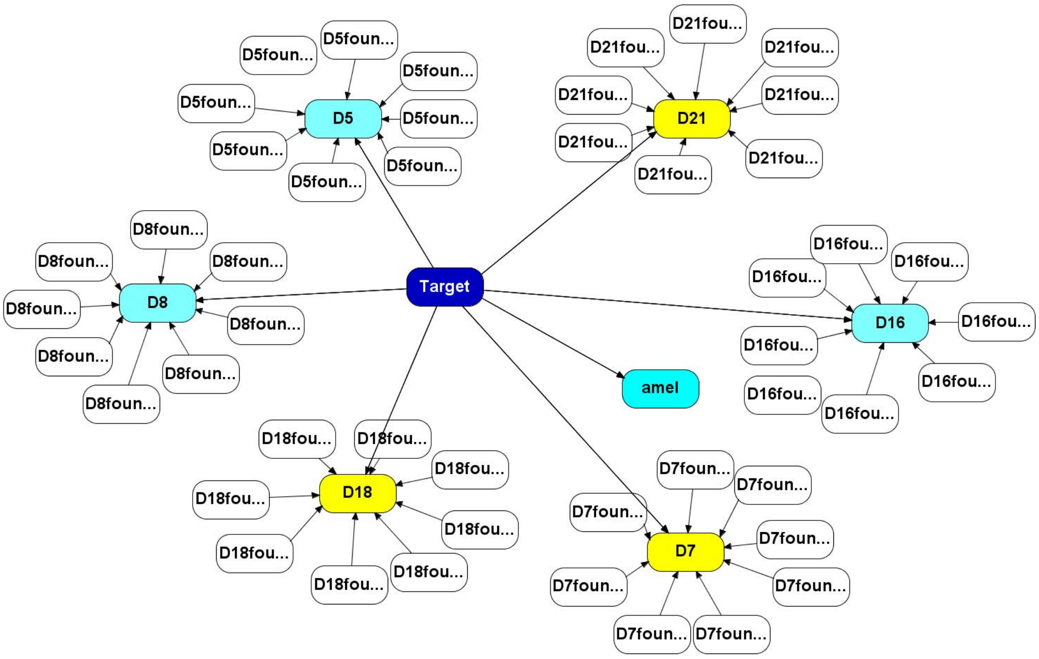


Figure A.12: Two person mixture. Master class

`amel` is an instance of the **Amelogenin** marker class. `Target` is an instance of class **target** and is linked to each marker via its output nodes `v_in_mix?`, `s_in_mix?` and `n_unknown`.

A.2 OOBN for 2-person mixtures including peak area information

In this section the object-oriented Bayesian network which Cowell *et al.* (2007b) used to investigate identification and separation of DNA mixtures using peak area information is described. Particularly, it is shown the single components and their internal structure which have been used in the construction of the master network.

The founder class

The class **founder** is the same described in the previous section.

The genotype class

The **genotype** class represents an individual's genotype **gt**. This class is shown in Figure A.13. It involves two input nodes, the paternal and maternal genes, (`pg` and `mg`), which are chosen independently from the same population with known allele frequencies. The paternal and maternal genes are copies of node `founder` of class **founder** and are parents of the output node `gt` which is their logical combination.

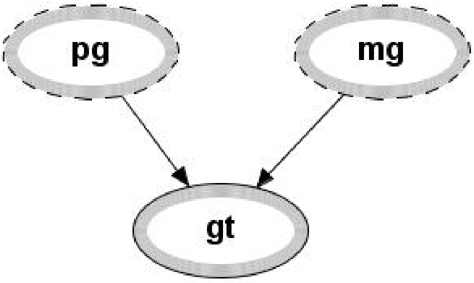


Figure A.13: Two person mixture. Genotype class.

The whichgt class

The class **whichgt** is shown in Figure A.14. This is a **query** class that chooses between two genotypes. It includes three input nodes called `query?`, `ingt` and `othergt` which are linked to an output node `outgt`. The `outgt` probability table is defined by the function: *if*(`query?`==*true*, `ingt`, `othergt`). This expression has meaning: if the Boolean node `query?` is *true*, `outgt` is a copy of the node `ingt`, otherwise it is identical to `othergt`.

The joint class

The combined genotype of the two contributors to the crime trace, `p1` and `p2`, is represented in the class **joint**. Thus, the node `p1gt&p2gt` is the logical combination of the two input genotypes in `p1gt` and `p2gt`. This class is

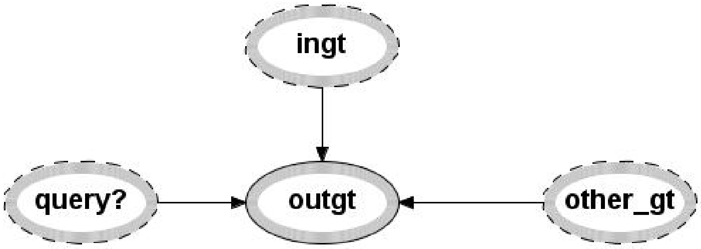


Figure A.14: Two person mixture. Whichgt class.

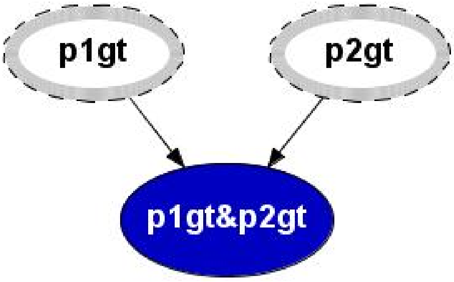


Figure A.15: Two person mixture. Jointgt class.

represented in Figure A.15.

The nalleles class

The class **nalleles** counts the number of alleles in a certain genotype. Figure A.16 shows the class **nalleles**. The output node `nA` counts the number of a particular allelic type in the genotype of the input node `gt`. For example for allele A, `nA` has the expression (if(`gt`$==AA$, 2, if(or(`gt`$==AB$, `gt`$==AC$, `gt`$==Ax$), 1, 0))). This expression has meaning: if the individual's genotype is AA, then `nA` counts 2 alleles, otherwise, if the individual's genotype is either AB, or AC, or Ax, it counts 1 allele, whilst in all the other cases `nA` is zero. In the equation (7.1) in § 7 the variable $n_a^{(i)}$ is modeled in this class.

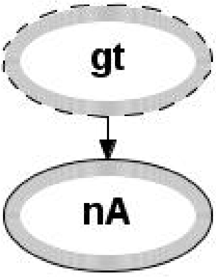


Figure A.16: Two person mixture. Nalleles class.

The alleleinmix class

The class **alleleinmix** represents the composition of the mixture, *i.e.* indicates whether the crime trace contains a certain allelic type. For the sake of brevity in the following lines the class **Aalleleinmix** only is taken into account, but the same structure applies to the other classes of this kind, *i.e.* **Balleleinmix**, **Calleleinmix**, etc. The class **Aalleleinmix** is shown in Figure A.17. Here, the input nodes, representing the genotypes of the two individuals `p1` and `p2`, have identity link to the input node `gt` of the class **nalleles**. The node `Ainmix?` indicates whether a particular allelic type is in the mixture. Thus, it is *true* if at least one of the two unknown contributors has allele A in the genotype. This can be translated by the logical expression: (*if*(*and*(`n1A_nA`==0, `n2A_nA`==0), *false*, *true*)), *i.e.* if both `n1A_nA` and `n2A_nA` counts 0 alleles, then `Ainimix?` is *false*, otherwise it is *true*. Here `n1A_nA` and `n2A_nA` are output nodes of the class **nalleles**. Node `Ainmix?` is an observation node, so that if allele A is measured in the mixture it is set to *true*, and the evidence on the mixture composition concerning allele A propagates from this node to the others.

Additionally, this class computes the mean contribution of a certain allelic type to the peak area. Input node `frac` is the proportion θ of DNA originated by the first contributor `p1`. This parameter is a continuous variable but, for convenience, discrete values are assigned to it in a scale ranging from $[0, 5]$ with step 1 in order to allow evidence propagation in the Bayesian network. `frac` node is linked to node `meanA` through the expres-

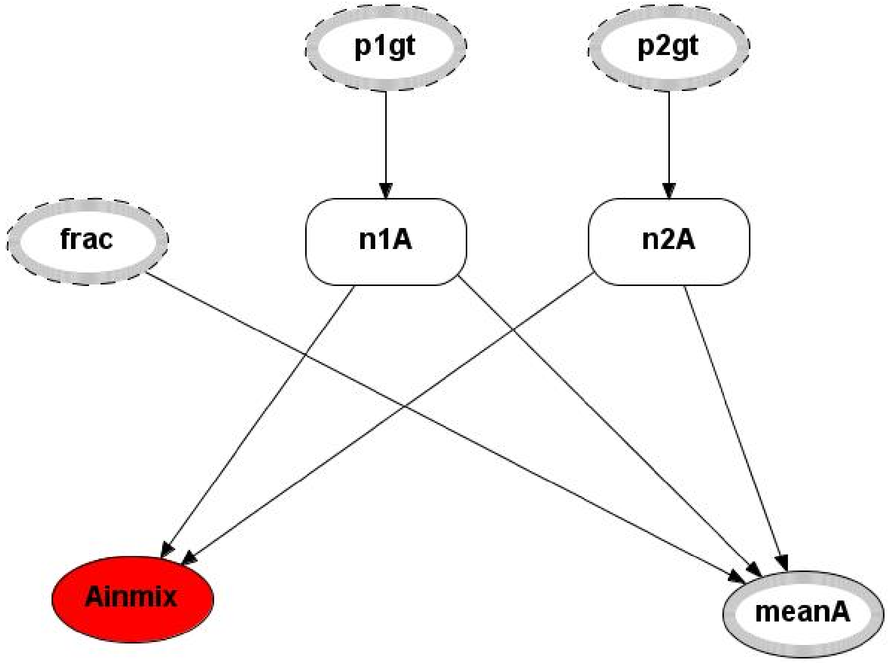


Figure A.17: Two person mixture. Alleleinmix class.

sion `meanA==n1A_nA*frac+n2A_nA*(5-frac)`. This is the same mean of the relative peak weights found in the equation (7.10) in § 7, but it differs by a scale factor of 10. Thus, when we apply a conditional-Gaussian model, before entering evidence on the relative peak weights, these have to be multiplied by 10. It is worth noting that, using a factorization for the conditional-Gamma model, the vector of the likelihoods in equation (7.7) in § 7.2.1 is entered in the node `meanA`.

The peakweight class

The class **peakweight** is shown in Figure A.18. This class models the observable peak weights as described in the conditional-Gaussian approximation model, thus it is not needed when peak areas are modelled with a conditional-Gamma distribution (see § 7.2.1 in § 7). The input node `mean` has identity link to the output node `meanA` in class **alleleinmix**. The unobserved true peak weight is represented by the continuous `area` node. This node has a

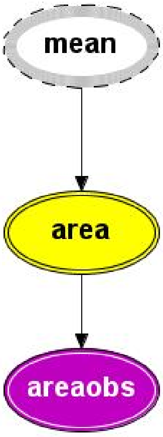


Figure A.18: Two person mixture. Peakweight class.

conditional-Gaussian distribution with mean equal to the value of `meanA` and variance given by 10 × 0.01 × μ, where the factor 10 is due to the fact that, since θ has been scaled of 5, then the mean has been scaled of 10 too. The continuous node `areaobs` is an observational node with mean 0 and variance representing variation in the measurement process. This node receive the evidence on the relative peak weights.

The marker class

The **marker** class represents a specific marker and contains several instances of the classes described so far since it is an upper level. Figure A.19 shows the **marker** class. All the input nodes `smg`, `spg`, `u1mg`, `u1pg`, `vmg`, `vsg`, `u2mg` and `u2pg` have identity links to the node `founder` in the **founder** class. The nodes `sgt`, `u1gt`, `vgt` and `u2gt` are instances of the **genotype** class. They contain respectively information on the suspect's, victim's and the two unknown individual's genotypes. Evidence on suspect and victim is set in the nodes `gt` of `sgt` and `vgt`. Nodes `p1gt` and `p2gt` are instances of the class **whichgt**. The Boolean node `squery` is connected to the input query node `query?` in **p1gt**; the output node `gt` in **sgt** is connected to the input node `ingt` in **p1gt**; the output node `gt` in **u1gt** is connected to the input node

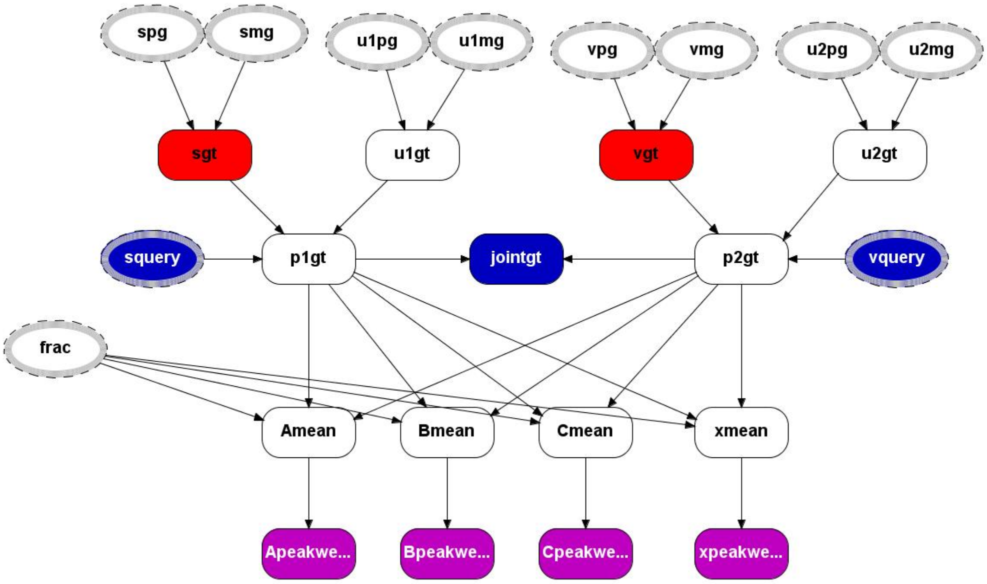


Figure A.19: Two person mixture. Marker class.

`othergt` in **p1gt**. Thus, if the node `squery` is *true* the output node `outgt` in **p1gt** is a copy of the node `gt` in **sgt**, otherwise is a copy of the node `gt` in **u1gt**. Similarly for **p2gt**. The node `jointgt` is an instance of the class **jointgt**. The nodes `Amean`, `Bmean`, `Cmean` and `xmean` are instances of the class **alleleinmix**. Their output node `meanA` is linked to the input node `mean` in the class **peakweight**. The node `frac` copies the corresponding nodes in the class **alleleinmix**. The instances of the class **peakweight** are used when peak areas are modelled with a conditional-Gaussian model.

The Amelogenin marker class

The **Amelogenin** class is shown in Figure A.20. This class has the same structure of the **marker** class. No **founder** class is introduced. Nodes `vgt`, `sgt`, `u1gt` and `u2gt` are instances of the class **genotype** for the **Amelogenin**. However, the class **genotype** used to build the class **Amelogenin**, has here a single output node `gt` with states XX for female and XY for male. The **whichgt** and **joint** classes are unchanged but have their state spaces reduced,

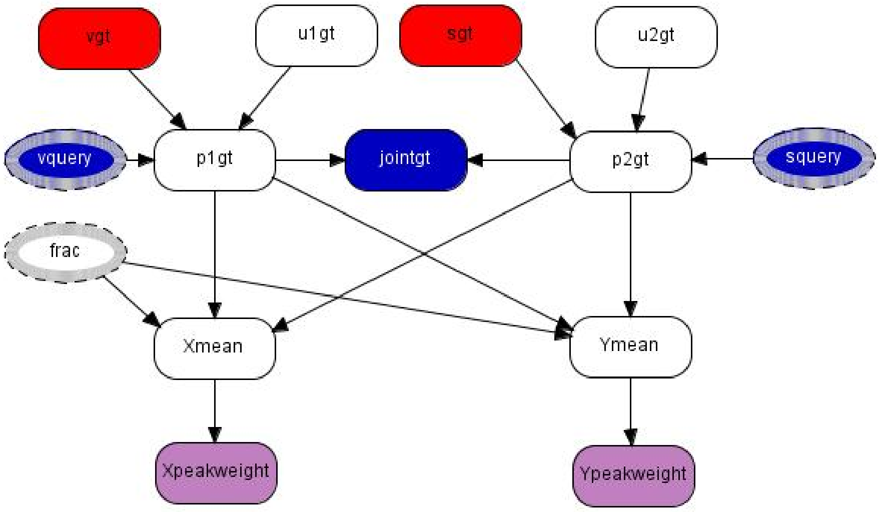


Figure A.20: Two person mixture. Amelogenin marker class.

i.e. they have two states only: XX and XY. In the class **nalleles** the node `nX` (`nY`) counts 1 (1) allele if the parent node `gt` is XY, whilst counts 2 (0) if the parent node `gt` is XX. The class **alleleinmix** is modified in the node `Xinmix` only which is always set to *true*. The **Amelogenin** class has only two instances of the class **alleleinmix** which are termed `Xmean` and `Ymean` and are connected to the nodes `Xpeakweight` and `Ypeakweight` instances of the **peakweight** class.

The target class

Figure A.21 shows the **target** class. The **target** class contains the `target` node where the results are read and the likelihood ratios are computed. This is the logical combination of the two Boolean nodes, `p1=s?` and `p2=v?`. Since `p1=s?` and `p2=v?` have a uniform prior distribution, then the `target` node also has a uniform prior distribution.

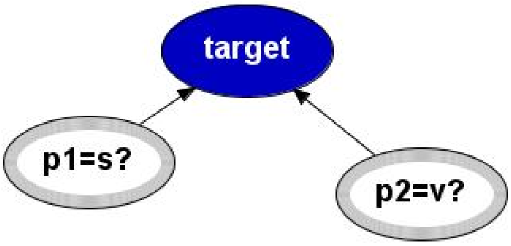


Figure A.21: Two person mixture. Target class.

The master class

The **master** network is given in Figure A.22. This network has been constructed in order to analyse the data in Table 5.4 § 5. Nodes `D5`, `D8`, `D7`, `D18`, `D16` and `D21` are all instances of the **marker** class. Nodes `D5`, `D16` and `D8` are **marker** instances with three observed alleles A, B and C , whilst nodes `D7`, `D18` and `D21` are **marker** instances with four observed alleles A, B, C and D. For each marker there are 8 instances of the class **founder** which are linked to the 8 input nodes of the class **marker**. The node `amel` is an instance of the **Amelogenin** class. The `frac` node is linked to the corresponding `frac` node in the **marker** instances. `Target` is an instance of class **target** and is linked to each marker via its output nodes `p1=s?` and `p2=v?`.

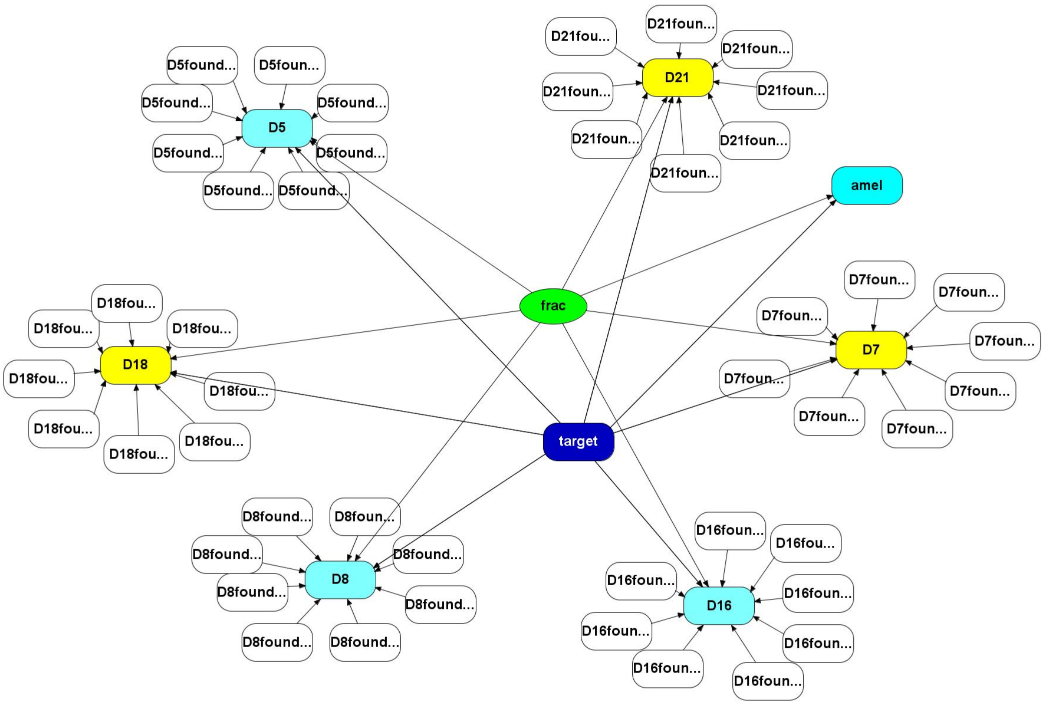


Figure A.22: Two person mixture. Master class.

Appendix B

Details of the object-oriented Bayesian network for two mixed traces

In this Appendix we describe the networks used to perform the analyses in chapter 8.

B.1 OOBN for two DNA mixed traces using alleles' repeat number information only

The modular structure of the object-oriented Bayesian network described in Appendix A.1 is here extended in order to include a second trace in the network.

We describe only the classes that have been changed, *i.e.* the classes **marker**, **Amelogenin**, **target** and **master**; whilst the classes **founder**, **genotype**, **identified**, **unknown** and **alleleinmix** remain unchanged and are described in Appendix A.1.

The marker class

The **marker** class represents a specific marker and contains a number of instances of the classes **founder**, **genotype**, **identified**, **unknown** and **alleleinmix**, since it is an upper level network. Figure B.1 shows the **marker**

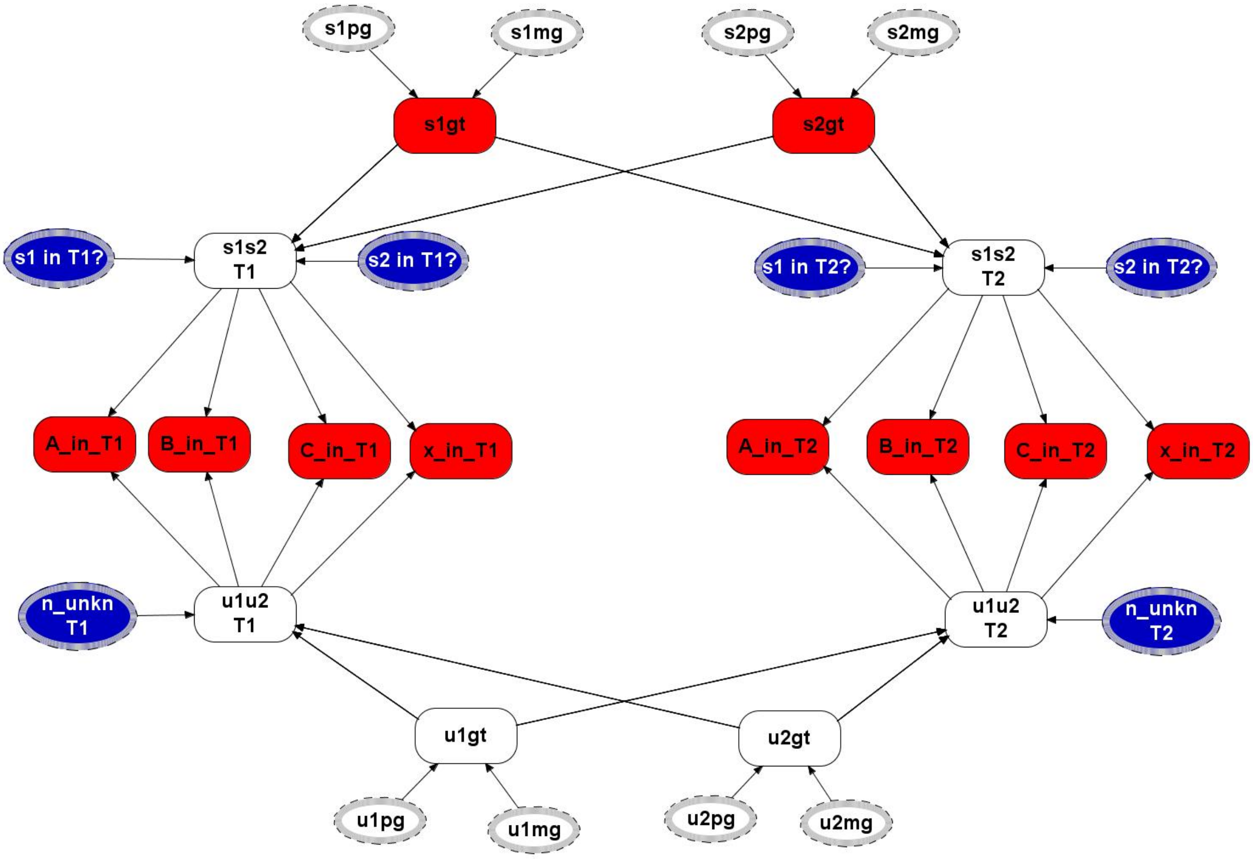


Figure B.1: Two traces. Marker class.

class. Here it is represented for a marker having three observed alleles in the mixture. Population allele frequencies, specified in Table 5.7 in 5, define the probability distribution of the input gene nodes s1pg, s1mg, s2pg, s2mg, u1pg, u1mg, u2pg and u2mg where, for example s1pg represents the first suspect's paternal gene, whilst s1mg is the first suspect's maternal gene.

Nodes s1gt, u1gt and u2gt are all instances of the **genotype** class. Evidence on the suspects' genotypes is entered in the network through the nodes A_in_gt, B_in_gt, C_in_gt and x_in_gt contained in the instances s1gt and s2gt.

Nodes s1s2_T1 and s1s2_T2 are instances of the class **identified**, whilst the nodes u1u2_T1 and u1u2_T2 are instances of the class **unknown**. Note that the letters "T1" and "T2" at the end of the name of each node indicates that we are referring to, respectively, the first or the second trace.

Input nodes `s1_in_T1?`, `s2_in_T1?` and `n_unknown_T1` indicates, respectively, whether *s1* is in the first trace, whether *s2* is in the first trace and the total number of unknown individuals in the first trace. They are identified with the corresponding input nodes contained in the instances `s1s2_T1` and `u1u2_T1`. Similarly for the other query nodes referred to the second trace, *i.e.* `s1_in_T2?`, `s2_in_T2?` and `n_unknown_T2`.

Nodes `A_in_T1`, `B_in_T1`, `A_in_T2`, `B_in_T2`, etc. are all instances of the class **alleleinmix**. Thus, for example, the output node `As1s2` in the instance `s1s2_T1` is linked with the input node `s1s2` in the instance `A_in_T1`; whilst, the output node `Au1u2` in the instance `u1u2_T1` is linked with the input node `U` in the instance `A_in_T1`.

The Amelogenin marker class

The **Amelogenin** class is shown in Figure B.2. This class has the same structure of the **marker** class. We show the differences.

In this class, as in the others, we added all the nodes referred to the second trace. Nodes called with the letters "T1" at the end of the name belong to the first trace, whilst nodes called with the letters "T2" at the end of the name are referred to the second trace. Nodes referred to the first trace are linked to those for the second trace through the nodes that represent the genotypes of the two suspects, `s1gt` and `s2gt`, and to the genotypes of the two unknown individuals, `u1gt` and `u2gt`.

The Target class

Figure B.3 shows the **target** class. The **target** class contains the nodes where the results are read. We describe the class in detail.

The query nodes `s1_in_T1?`, `s2_in_T1?`, `n_unknown_T1`, represent, respectively, the presence of the suspects in the first trace and the number of unknown individuals in the mixture; they have uniform prior probabilities. These nodes are parents of the node `total_#_T1`, which counts all contrib-

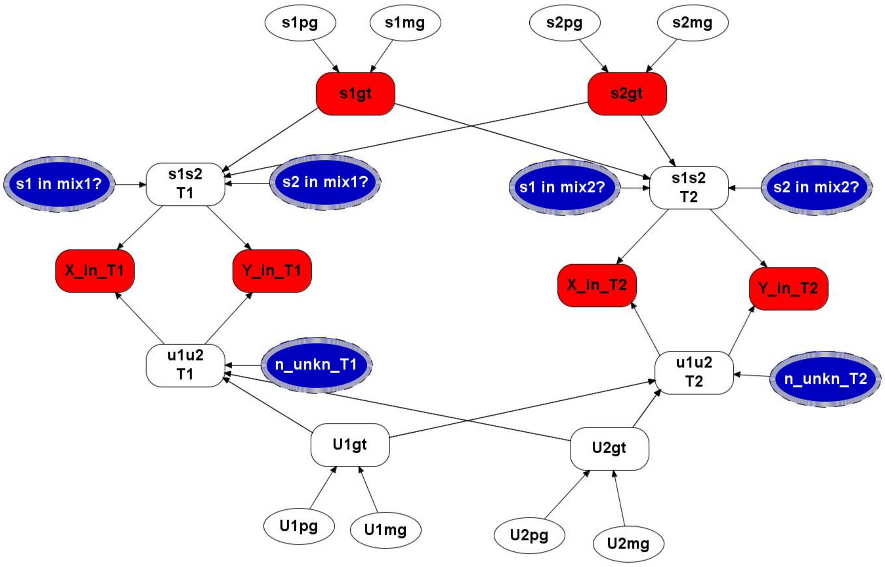


Figure B.2: Two traces. Amelogenin marker class.

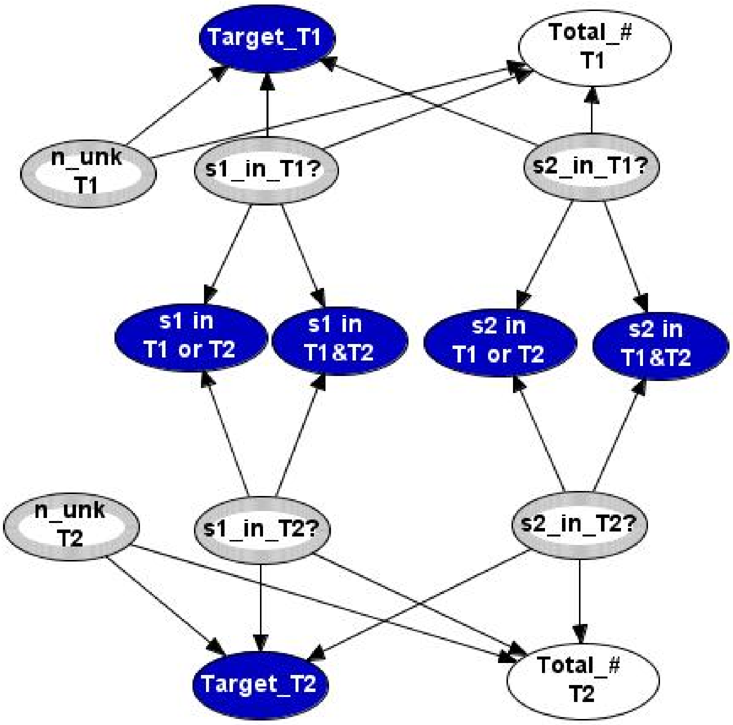


Figure B.3: Two traces. Target class.

utors in the first trace. This node has states from 0, if n_unknown_T1 is 0 and both s1_in_T1? and s2_in_T1? are *false*, to 4, if n_unknown_T1 is 2 and both s1_in_T1? and s2_in_T1? are *true*. Similarly for the nodes s1_in_T2?, s2_in_T2?, n_unknown_T2 and total_#_T2 which are referred to the second trace *T2*.

Furthermore, the nodes s1_in_T1? and s1_in_T2? are connected to the *Boolean nodes* s1_in_T1_or_T2 and s1_in_T1&T2. Node s1_in_T1_or_T2 indicates the presence of the suspect1 in at least one trace. Thus, it is *true* if at least one either s1_in_T1? or s1_in_T2? is *true*. Node s1_in_T1&T2 indicates the presence of the suspect1 in both crime traces. Thus, it is *true* if both s1_in_T1? and s1_in_T2? are *true*. Similarly for the second suspect *s2*.

The states of the Target_T1 and Target_T2 *query nodes* represent the 12 hypotheses under test and are defined by the states of its parent nodes. In other words, its states are made of the aggregation of the states of its parents, being aware that a *false* or a *zero* in the parents' states are not reported in its final state, *e.g.* for the node Target_T1, if the parents' states are s1_in_T1?=true, s2_in_T1?=true and n_unknown_T1=1, this node's state will be *s1&s2&1u*, whilst if they are s1_in_T1?=true, s2_in_T1?=false and n_unknown_T1=0 its state will be just *s1*.

The master class

Figure B.4 shows the **master** class where markers D2, D21 FGA, TH01 and VWA are specified through the instances of **marker** class. They are all markers with three observed alleles in the mixture, except D21 which has four observed alleles *A*, *B*, *C* and *D*. Each marker has 8 instances of class **founder** with their appropriate frequencies and linked with the 8 input nodes of the class **marker**.

The node amel represents the **Amelogenin** class and therefore it does not need of **founder** classes.

Target is an instance of class **target** and it is linked to each marker via its output nodes s1_in_T1?, s2_in_T1?, n_unknown_T1, s1_in_T2?, s2__T2?

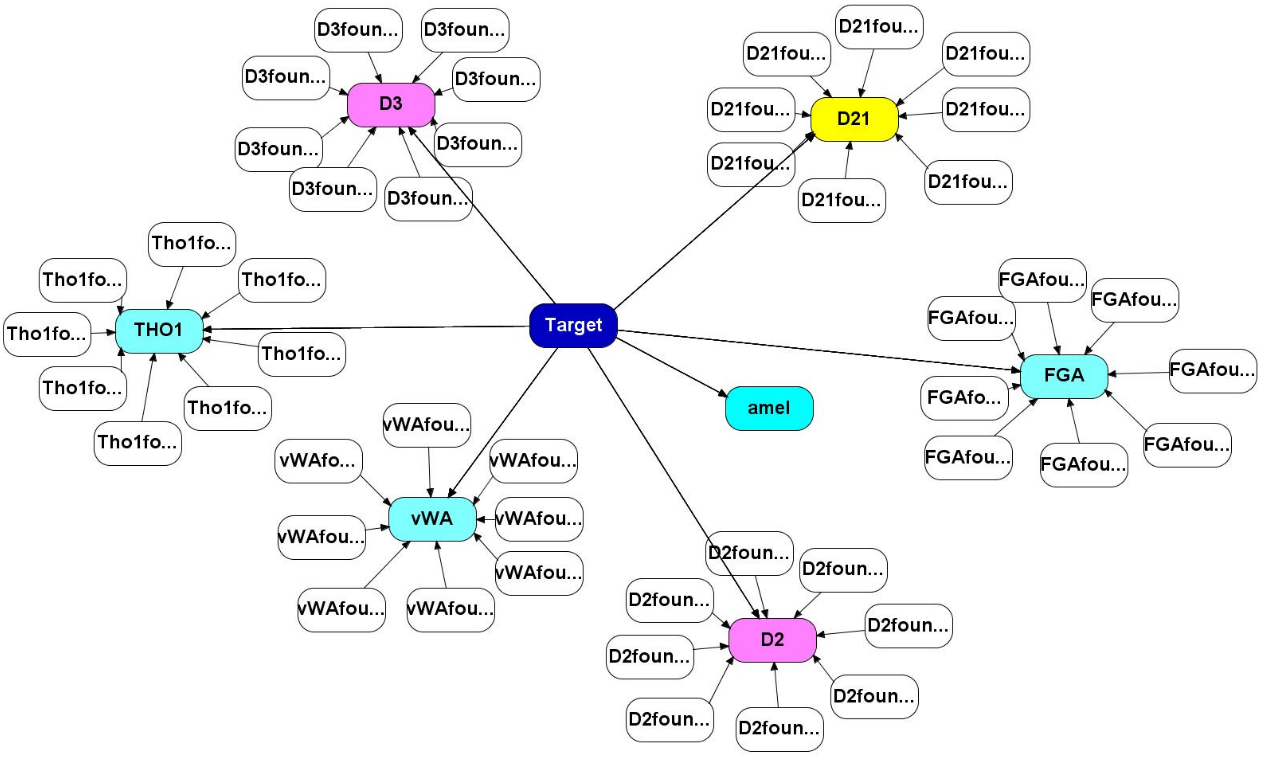


Figure B.4: Two traces. Master class.

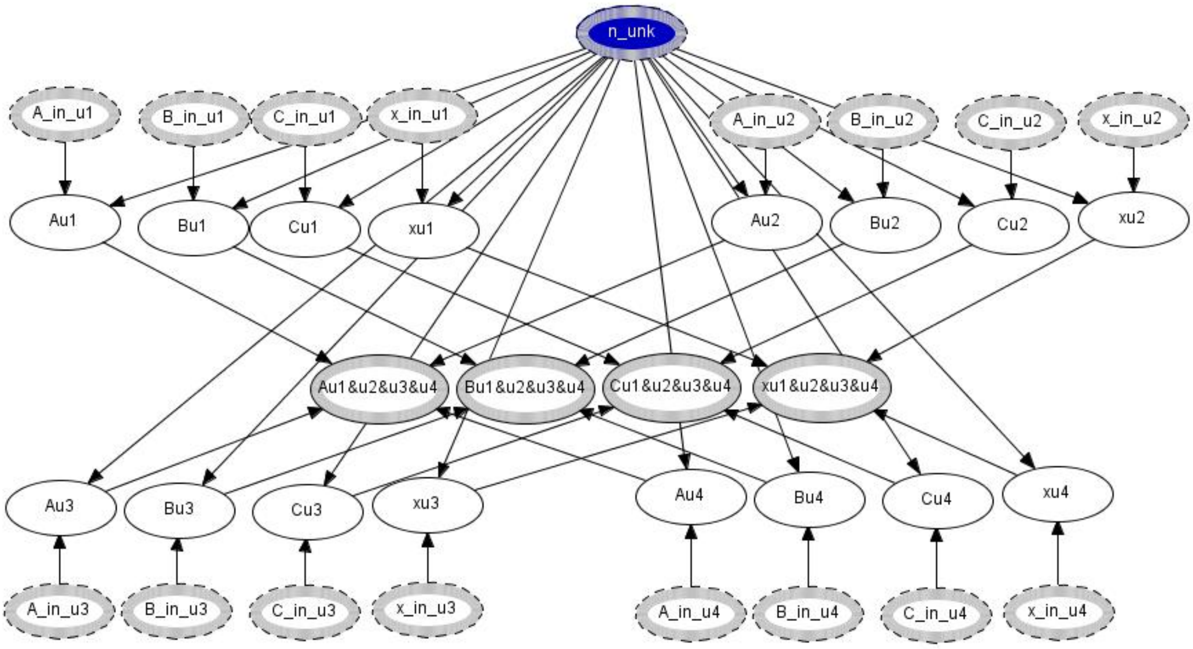


Figure B.5: Two traces. Unknown class.

and n_unknown_T2.

Adding unknown number of contributors

This network can be easily extended to account more unknown contributors. We show only the classes to change in order to add more unknown contributors.

Figure B.5 shows the **unknown** class modified for up to 4 unknown contributors. Here the node n_unknown has the number of its states increased up to 4. Furthermore, we added all the nodes referred to the third and the fourth unknown contributor, *u3* and *u4*. The nodes A_in_u3, B_in_u3, A_in_u4, B_in_u4, etc. are *Boolean* nodes with uniform prior probabilities. Node Au3 is *true* if A_in_u3 is *true* and n_unknown is 3, 4, otherwise is *false*. Similarly for the nodes Bu3, Cu3 and xu3. Node Au4 is *true* if A_in_u4 is *true* and n_unknown is 4, otherwise is *false*. Similarly for the nodes Bu4, Cu4 and xu4. Node Au1&u2&u3&u4 has two more parents (Au3 and Au4) and it is *true* if either Au1, or Au2, or Au3, or Au4 is *true*, otherwise it is *false*. Similarly

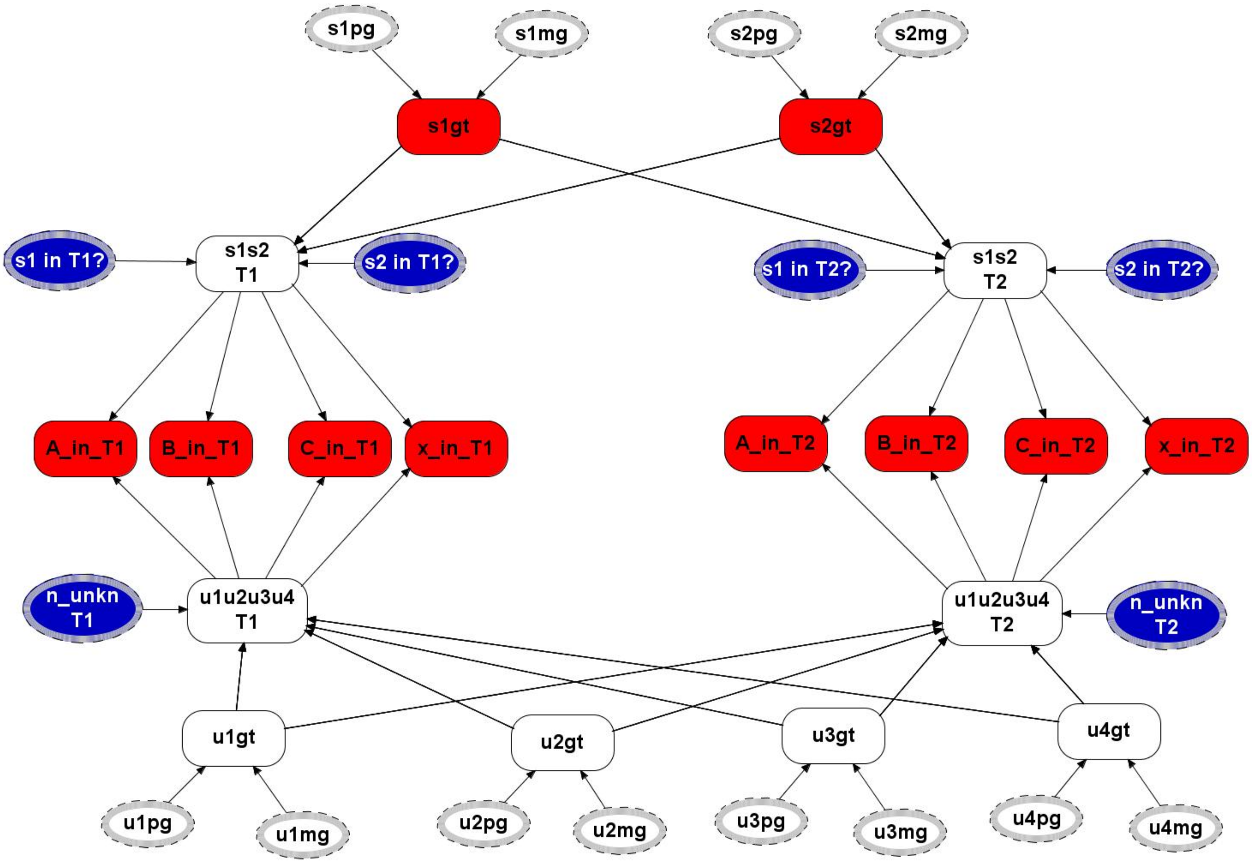


Figure B.6: Two traces with 4 unknown individuals. Master class.

for `Bu1&u2&u3&u4`, `Cu1&u2&u3&u4` and `xu1&u2&u3&u4`.

In the **marker** class shown in Figure B.6 we added the founder nodes `u3pg`, `u3mg`, `u4pg` and `u4mg`. As a consequence, the **master** class contains now 12 instances of the class **founder**, rather than 8, connected with the 12 input nodes in the **marker** class.

Additionally, we introduced the nodes `u3gt` and `u4gt` which are instances of the **genotype** class. The input nodes of these instances are connected to the output nodes of the instances `u1u2u3u4_T1` and `u1u2u3u4_T2`.

All the other classes are unchanged, whilst the **Amelogenin** has similar modifications.

B.2 OOBN for two DNA traces including peak area information

In this section we expand the network used by Cowell *et al.* (2007b) in a way so that we can include two traces in the same network. It models DNA mixtures using both alleles repeat number and peak area information but is not used to make inference on the total number of contributors.

For this network, as for the previous one, we describe only the classes that have been modified, *i.e.* the classes **marker**, **Amelogenin**, **target** and **master**; whilst the classes **founder**, **genotype**, **whichgt**, **joint**, **nalleles**,**alleleinmix** and **peakweight** remain unchanged and are described in Appendix A.2.

The marker class

The **marker** class represents a specific marker and contains instances of the classes **genotype**, **whichgt**, **joint**, **nalleles**,**alleleinmix** and **peakweight**, since it is an upper level. Figure B.7 shows the **marker** class. All the input nodes `s1pg`, `s1mg`, `u1pg`, `u1mg`, `s2pg`, `s2mg`, `u2pg` and `u2mg` have identity links with the node `founder` in the **founder** class. The nodes `s1gt`, `u1gt`, `s2gt` and `u2gt` are instances of the **genotype** class. They contain respectively

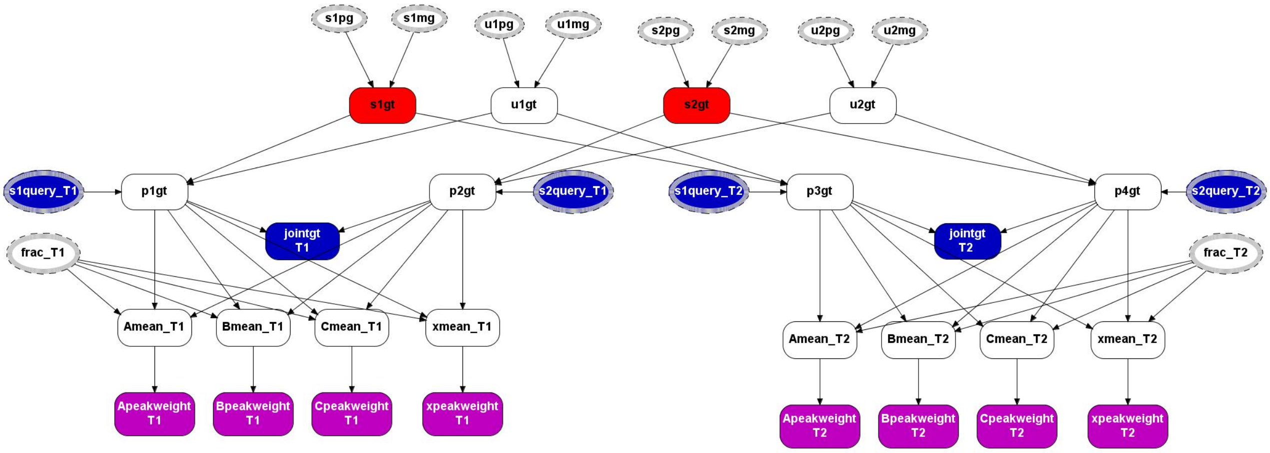


Figure B.7: Two traces. Marker class.

information on the suspect1's, suspect2's and the two unknown individual's genotypes. Evidence on suspects is set in the nodes gt of s1gt and s2gt. Nodes p1gt, p2gt represent the two individuals, p1 and p2, contributors in the first trace *T1*. Nodes p3gt and p4gt represent the two individuals, p3 and p4, contributors in the second trace *T2*. They are all instances of the class **whichgt**. The Boolean node s1query_T1 is connected to the input query node query? in **p1gt**; the output node gt in **s1gt** is connected to the input node ingt in **p1gt**; the output node gt in **u1gt** is connected to the input node othergt in **p1gt**. Thus, if the node s1query_T1 is *true* the output node outgt in **p1gt** is a copy of the node gt in **s1gt**, otherwise is a copy of the node gt in **u1gt**. Similarly for **p2gt**, **p3gt** and **p4gt**. Nodes jointgt_T1 and jointgt_T2 are instances of the class **jointgt**. Here the term *T1* or *T2* at the end of the name of the node indicates that the node is referred to, respectively, the first or the second trace. Nodes Amean_T1, Bmean_T1, Amean_T2, Bmean_T2, etc. are all instances of the class **alleleinmix**. Their output node meanA is linked to the input node mean in the class **peakweight**. When peak areas are modelled with a conditional-Gaussian model, we enter the evidence on the relative peak weights in the classes **peakweight**. Nodes frac_T1 and frac_T2 are linked with the corresponding node frac in the class **alleleinmix**.

The Amelogenin class

The **Amelogenin** class is represented in Figure B.8. This class has the same structure of the **marker** class. No **founder** class is introduced. Nodes s1gt, u1gt, s2gt and u2gt are all instances of the class **genotype**. But, the class **genotype**, used to build the class **Amelogenin**, has here a single output node gt with states XX for female and XY for male. The **whichgt** and **joint** classes are unchanged but have their state spaces reduced, *i.e.* they have two only states XX and XY. In the class **nalleles** the node nX (nY) counts 1 (1) allele if the parent node gt is XY, whilst counts 2 (0) if the parent node gt is XX. The class **alleleinmix** is modified in the node Xinmix

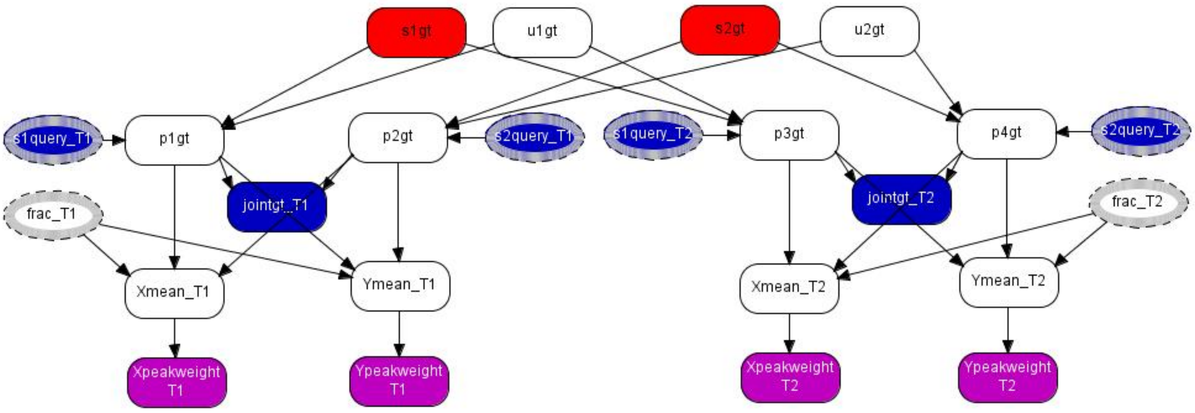


Figure B.8: Two traces. Amelogenin marker class.

only which is set always *true*. The **Amelogenin** class has only two instances for each trace of the class **alleleinmix** which are termed `Xmean_T1`, `Ymean_T1`, `Xmean_T2` and `Ymean_T2`. They are connected to the nodes `Xpeakweight_T1`, `Ypeakweight_T1`, `Xpeakweight_T2` and `Ypeakweight_T2` which are all instances of the **peakweight** class.

The Target class

Figure B.9 shows the **target** class. The **target** class contains the nodes `target_T1` and `target_T2` where the results are read and the likelihood ratios are computed, respectively, for the first and the second trace. Node `Target_T1` is the logical combination of the two *Boolean* nodes, `p1=s1?` and `p2=s2?`; whilst node `Target_T2` is the logical combination of the two *Boolean* node `p3=s1?` and `p4=s2?`.

Since `p1=s1?`, `p2=s2?`, `p3=s1?` and `p4=s2?` have a uniform prior distribution, then the nodes `target_T1` and `target_T2` also have a uniform prior distribution. Additionally, their states are the hypotheses under test for each trace. They are shown in Table B.1.

The *Boolean* node `s1_in_T1_or_T2` indicates the presence of the first suspect in at least one trace. Thus, it is *true* if either `p1=s1?` or `p3=s1?` is *true*. The *Boolean* node `s1_in_T1&T2` indicates the presence of the first suspect in

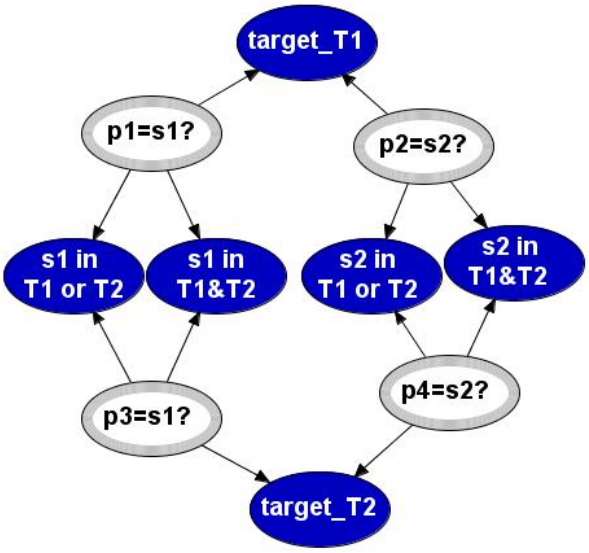


Figure B.9: Two traces. Target class.

Hypotheses under test	
s1&s2	both suspects contributed to the mixture
s1&u	the first suspect and an unknown individual contributed to the mixture
s2&u	the second suspect and an unknown individual contributed to the mixture
2u	two unknown individuals contributed to the mixture

Table B.1: Hypotheses under test.

both the two traces. Thus, it is *true* if both `p1=s1?` and `p3=s1?` are *true*. Similarly for the nodes referred to the second suspect, `s2_in_T1_or_T2` and `s2_in_T1&T2`.

The master class

Figure B.10 shows the **master** class. In the **master** class nodes `D2`, `D21`, `FGA`, `THO1` and `VWA` are all instances of the **marker** class. For each marker, there are 12 instances of the class **founder** linked with the 8 input nodes of the class **marker**. The `frac_T1` and `frac_T2` nodes are linked with the corresponding nodes in the **markers**. The node `amel` represents the **Amelogenin** class and therefore it does not need of **founder** classes. `Target` node is an instance

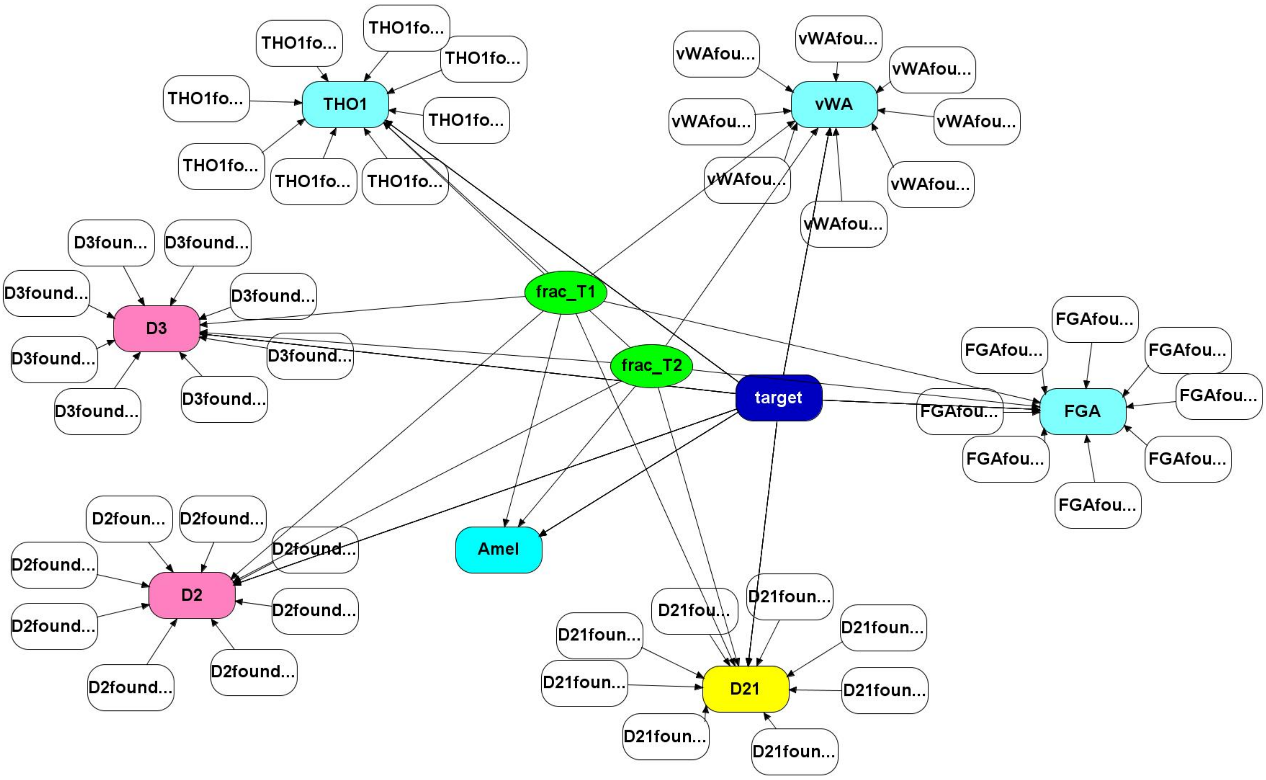


Figure B.10: Two traces. Master class.

of class **target** and is linked to each marker via its output nodes p1=s1?, p2=s2?, p3=s1? and p4=s2?.

Appendix C

Details of the object-oriented Bayesian network for 3-person mixtures

In this Appendix we describe the networks used to perform the analyses in chapter 9.

C.1 OOBN for 3-person DNA mixtures using alleles' repeat number information only

The modular structure of the object-oriented Bayesian network described in Appendix A.1 is extended in order to include a third contributor to the mixed trace.

We describe the classes that have been changed only, *i.e.* the classes **identified**, **unknown**, **marker**, **Amelogenin**, **target** and **master**; whilst the classes **founder**, **genotype**, and **alleleinmix** are unchanged and are described in Appendix A.1.

The identified class

Figure C.1 shows the class **identified**. This class represents the presence in the mixture of a specific allele contributed by at least one of the identified individuals *v*, *s1* and *s2*. We describe the class in detail. The input query

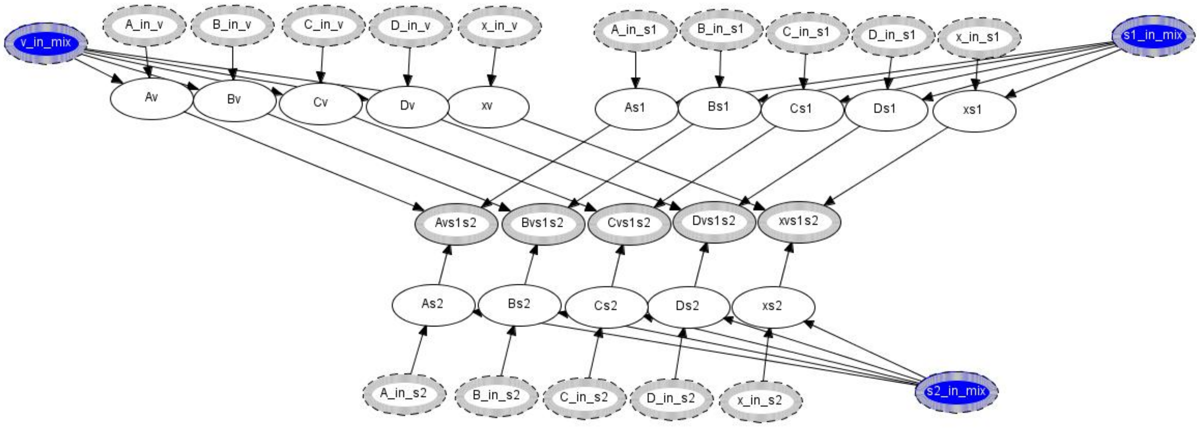


Figure C.1: Three person mixture. Identified genotype class.

node `v_in_mix?` represents the binary query: “is the victim’s genotype in the mixture?”. Similarly for the other input query nodes `s1_in_mix?` and `s2_in_mix?`. The input nodes `A_in_v`, `B_in_v`, `A_in_s1`, `B_in_s1`, `A_in_s2`, `B_in_s2`, etc. are copies of the same labeled nodes of the class **genotype** described in Appendix A.1. The node `Av` is the logical conjunction `v_in_mix?` ∩ `A_in_v`, thus, it is *true* if both its parent nodes `v_in_mix?` and `A_in_v` are *true*. In other words, this is *true* if in the mixture there is the allele A contributed by the victim, otherwise is *false*. Similarly for the nodes `Bv`, `As1`, `Bs1`, `As2`, `Bs2`, etc. The node `Avs1s2` indicates the presence of allele A in either the victim or one suspect who contributed to the mixture. Thus, this is *true* if either one parent node, `Av`, `As1` or `As2`, is *true* and *false* otherwise.

The unknown class

Figure C.2 shows the class **unknown**. This class is similar to the previous one since it represents the presence in the mixture of a specific allele contributed by at least one unknown individual, either `u1` or `u2` or `u3`. We introduce three unknown individuals since this is a network for a 3-person mixture, thus the mixed trace can contain at most the DNA profiles of three unknown individuals. Node `n_unknown` specifies the number of unknown individuals in the mixture, therefore it has values 0, 1, 2, 3 with same probabilities. The input nodes `A_in_u1`, `B_in_u1`, `A_in_u2`, `B_in_u2`, `A_in_u3`, `B_in_u3`, etc. are

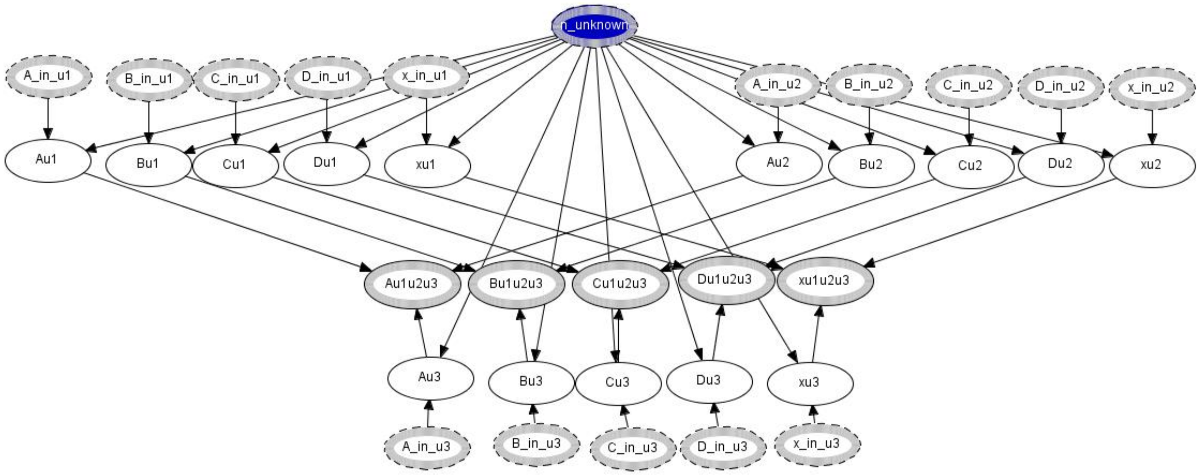


Figure C.2: Three person mixture. Unknown class.

copies of the same labeled nodes of the class **genotype**. Node Au1 is *true* if A_in_u1 is *true* and n_unknown is 1, 2, 3, otherwise is *false*. Similarly for the nodes Bu1, Cu1, Du1 and xu1. Node Au2 is *true* if A_in_u2 is *true* and n_unknown is 2, 3, otherwise is *false*. Similarly for the nodes Bu2, Cu2, Du2 and xu2. Finally, node Au3 is *true* if A_in_u3 is *true* and n_unknown is 3, otherwise it is *false*. Similarly for the nodes Bu3, Cu3, Du3 and xu3. Node Au1u2u3 is the logical disjunction Au1∪Au2∪Au3. Thus, this is *true* if either parent node is *true* and *false* otherwise.

The marker class

The **marker** class represents a specific marker and contains instances of the classes described so far since it is an upper level network. Figure C.3 shows the **marker** class. Here it is represented for a marker containing five alleles in the mixture, A, B, C, D and x. Population allele frequencies, specified in Table 5.9 in § 5, define the probability distribution of the input gene nodes vpg, vmg, s1pg, s1mg, s2pg, s2mg, u1pg, u1mg, u2pg, u2mg, u3pg and u3mg where, for example vpg represents the victim's paternal gene, whilst vmg is the victim's maternal gene.

Nodes vgt, s1gt, s2gt, u1gt, u2gt and u3gt are all instances of the **genotype** class. Evidence on the victim's and the two suspects' genotypes

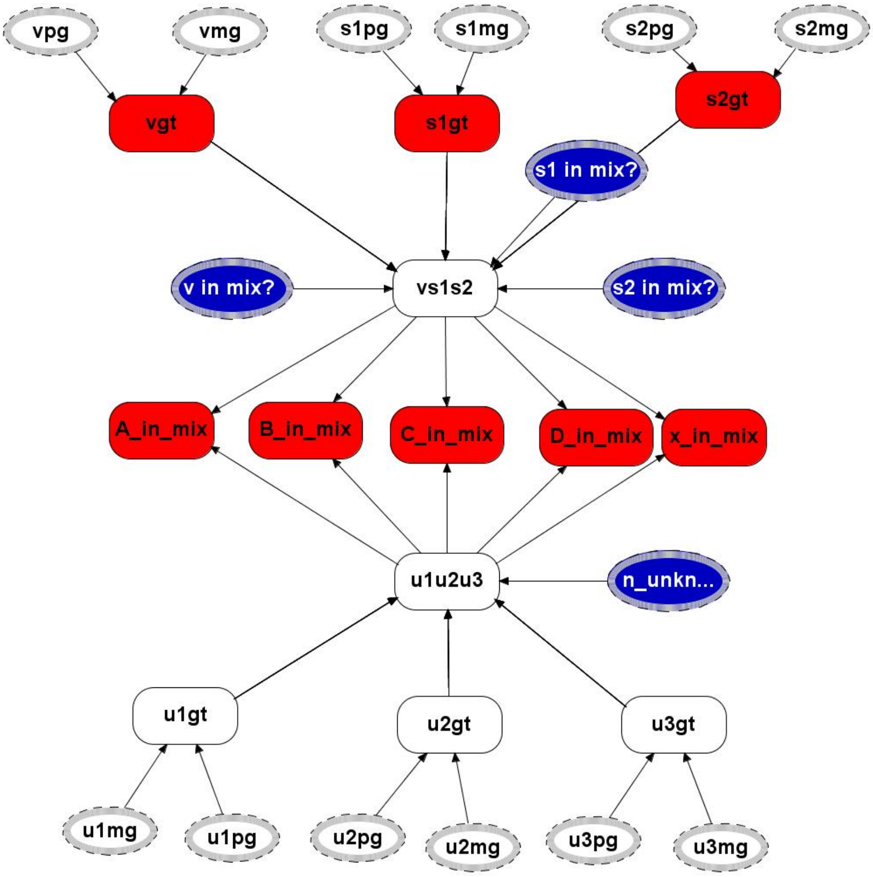


Figure C.3: Three person mixture. Marker class.

is entered in the network through the nodes A_in_gt, B_in_gt, etc. contained in the instances **vgt**, **s1gt** and **s2gt**.

Node vs1s2 is an instance of the class **identified**, whilst the node u1u2u3 is an instance of the class **unknown**.

Input nodes v_in_mix?, s1_in_mix?, s2_in_mix? and n_unknown are identified with the corresponding input nodes contained in the instances **vs1s2** and **u1u2u3**.

Nodes A_in_mix, B_in_mix, C_in_mix, D_in_mix and x_in_mix are all instances of the class **alleleinmix**. For example, the output node Avs1s2 in the class **vs1s2** is linked to the input node vs1s2 in the instance **A_in_mix**; whilst, the output node Au1u2u3 in the class **u1u2u3** is linked to the input node U in the class **A_in_mix**.

The Amelogenin marker class

The **Amelogenin** class is shown in Figure C.4. This class has similar

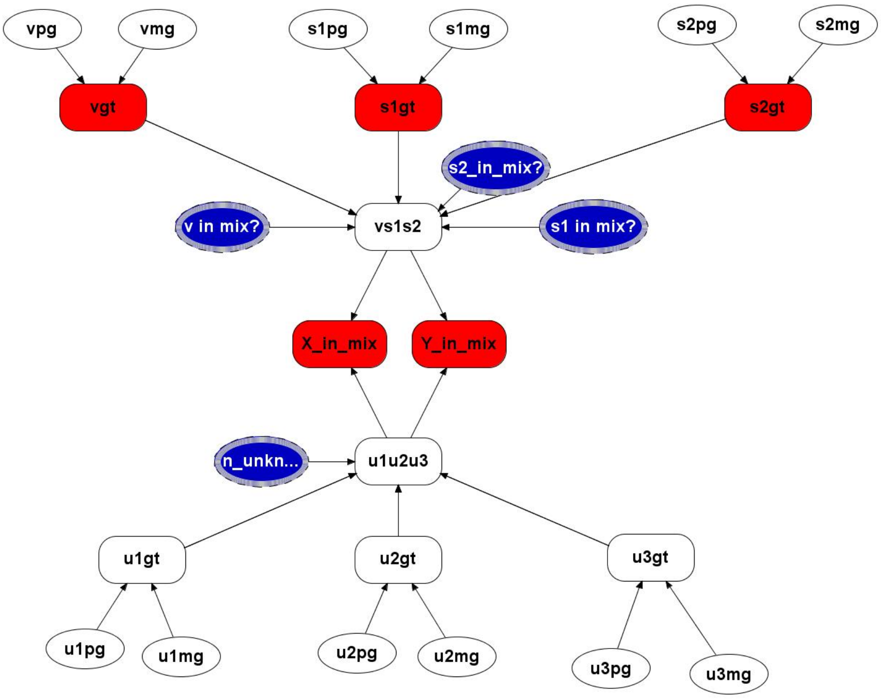


Figure C.4: Three person mixture. Amelogenin marker lass.

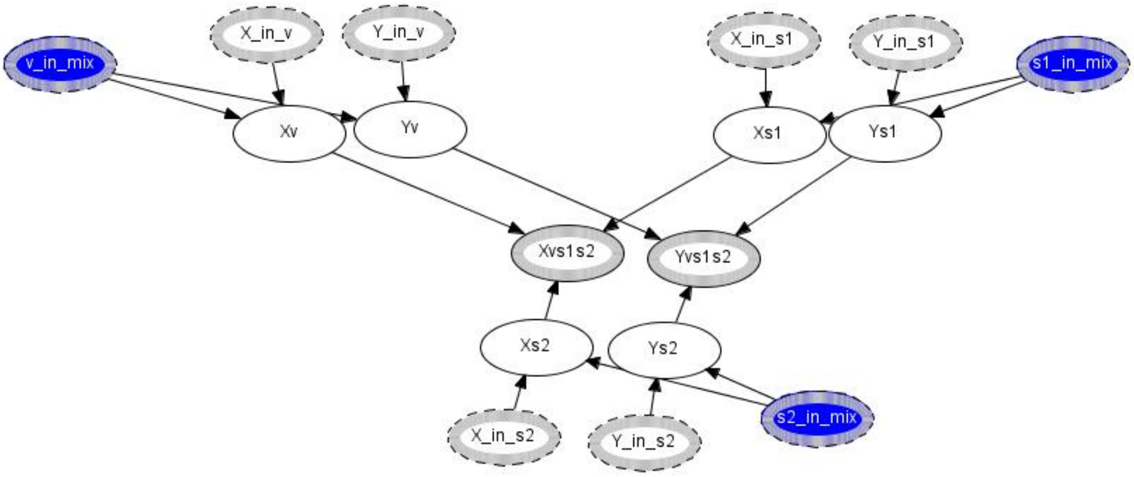


Figure C.5: Three person mixture. Identified genotype class for the Amelogenin marker.

structure of the **marker** class. Here, the nodes `vpg`, `vmg`, `s1pg`, `s1mg`, `s2pg`, `s2mg`, `u1pg`, `u1mg`, `u2pg`, `u2mg`, `u3pg` and `u3mg` are ordinary nodes, thus no **founder** class is needed, and they have state space XX for female and XY for male. Nodes `vgt`, `s1gt`, `s2gt`, `u1gt`, `u2gt` and `u3gt` are instances of the class **genotype** for the **Amelogenin** which is the same class described in Appendix A.1. Furthermore, the node `vs1s2` is an instance of the class **identified**. In the class **identified**, used to build the class **Amelogenin** and shown in Figure C.5, the collection of the allele nodes has two nodes for each group referred to the allele X and Y. Similarly for the class **unknown** shown in Figure C.6. Finally, the nodes `X_in_mix` and `Y_in_mix`, in the **Amelogenin** class, are instances of the class **alleleinmix**.

The Target class

Figure C.7 shows the **target** class. The **target** class contains the nodes where the results are read. We describe the class in detail. Nodes `v_in_mix?`, `s1_in_mix?`, `s2_in_mix?` and `n_unknown` have uniform prior probabilities. These nodes are parents of the node `total_#`, which counts all contributors. Node `total_#` has states from 0, if `n_unknown` is 0 and all `v_in_mix?`, `s1_in_mix?` and `s2_in_mix?` are *false*, to 6, if `n_unknown` is 3 and all

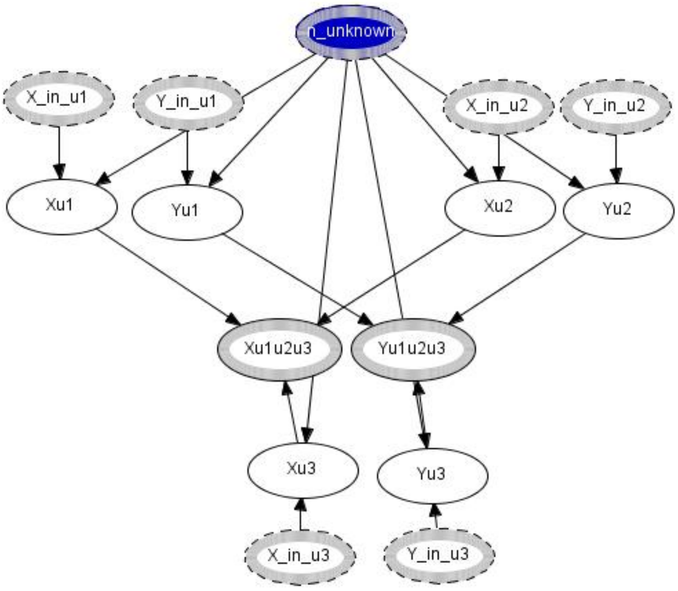


Figure C.6: Three person mixture. Unknown class for Amelogenin marker.

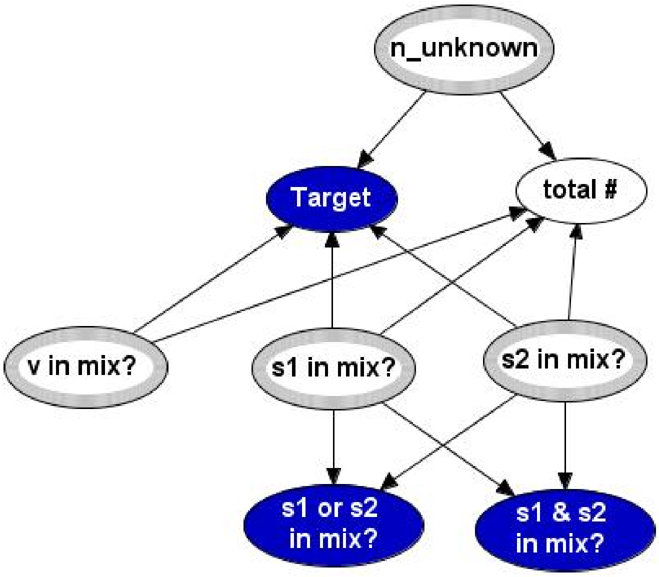


Figure C.7: Three person mixture. Target class.

v_in_mix?, s1_in_mix? and s2_in_mix? are *true*.

Furthermore, the nodes s1_in_mix? and s2_in_mix? are connected to the *Boolean nodes* s1_or_s2_in_mix and s1&s2_in_mix. Node s1_or_s2_in_mix indicates the presence of at least one suspect in the mixture. Thus, this is *true* if at least one either s1_in_mix? or s2_in_mix? is *true*. Node s1&s2_in_mix indicates the presence of both the suspects in the mixture. Thus, this is *true* if both s1_in_mix? and s2_in_mix? are *true*.

The states of the Target *query node* represent the 32[1] hypotheses under test and are defined by the states of its parent nodes. In other words, its states are made of the aggregation of the states of its parents, being aware that a *false* or a *zero* in the parents' states are not reported in its final state, *e.g.* if the parents' states are v_in_mix?=true, s1_in_mix?=true, s2_in_mix?=true and n_unknown=1, this node's state will be *v&s1&s2&1u*, whilst if they are v_in_mix?=true, s1_in_mix?=true, s2_in_mix?=false and n_unknown=0 its state will be *v&s1*.

The master class

Figure C.8 shows the **master** class where markers **D7**, **D8**, and **D21** are specified through the instances of **marker** class. They are all markers with five observed alleles in the mixture, A, B, C, D and x. Each marker has 12 instances of class **founder** with their appropriate frequencies and linked to the 12 input nodes of the class **marker**.

The node amel represents the **Amelogenin** class and therefore it does not need of **founder** classes.

Target is an instance of class **target** and it is linked to each marker via its output nodes v_in_mix?, s1_in_mix?, s2_in_mix? and n_unknown.

Concluding, in this 3-person mixture the third contributor could be represented by a second suspect *s2* or a third unknown individual *u3*. Thus in order to include a third contributor to the mixed trace, we need to add all the nodes referred the second suspect *s2* and the third unknown individual

[1] Obviously, in a specific court case only two competing hypothesis will be needed.

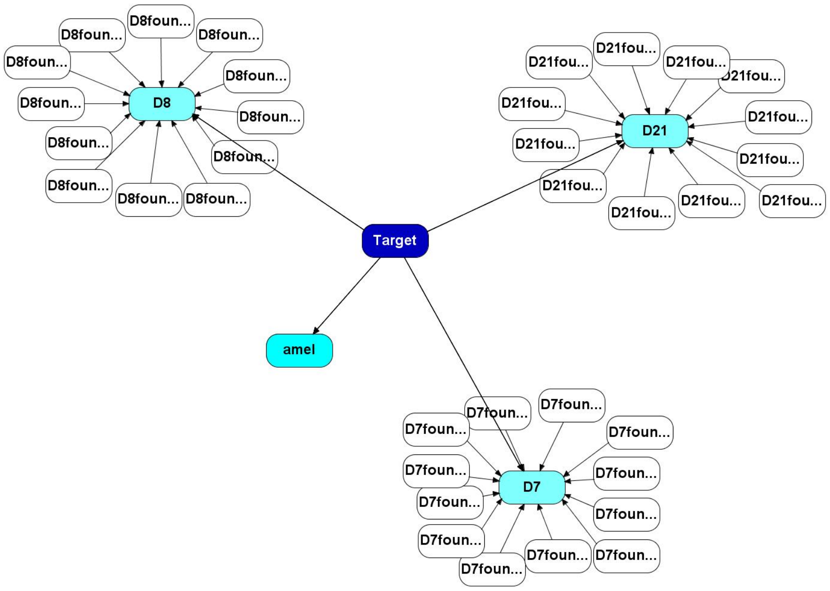


Figure C.8: Three person mixture. Master class.

u3.

Additionally, it is worth noting that this network has been built to be applied in a specific rape case where biological material from the victim and two suspects is considered. However, the network remains unaltered also for different cases that involve, for example, three suspects. In this alternative case only one modification could be introduced: in the **target** class the nodes that indicate the presence of at least one suspect in the crime trace, `s1_or_s2_or_s3_in_mix`, and the presence of all suspects in the mixture, `s1&s2&s3_in_mix`, will have one more parent node `s3_in_mix?`.

C.2 OOBN for 3-PERSON DNA mixtures including peak area information

In this section we expand the network used by Cowell *et al.* (2007b), which includes the nodes referred to peak areas, to mixtures involving three contributors.

For this network, as for the previous ones, we describe only the classes that

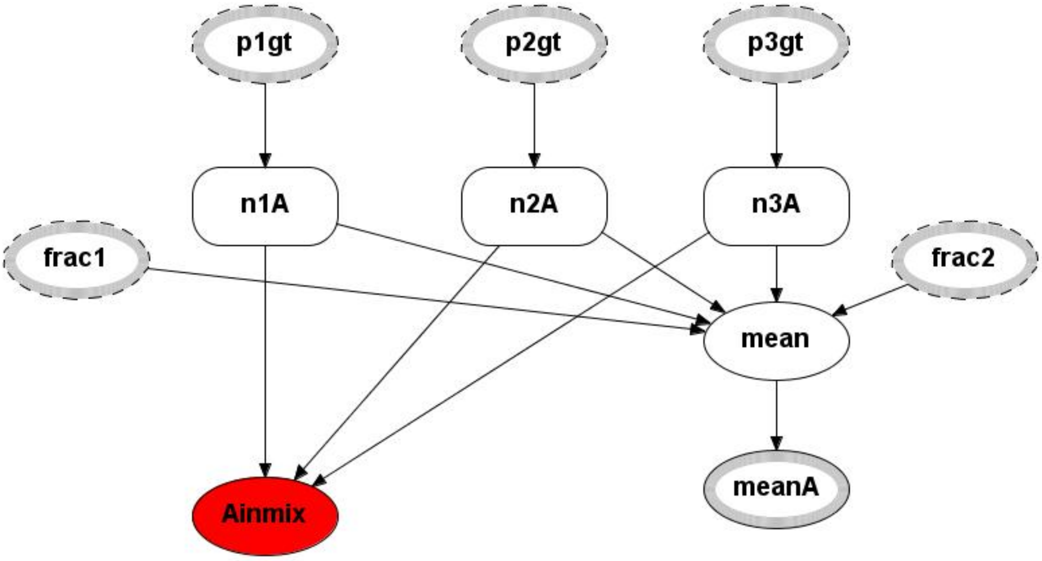


Figure C.9: Three person mixture. Alleleinmix class.

have been changed, *i.e.* the classes **alleleinmix**, **joint**, **marker**, **Amelogenin**, **target** and **master**; whilst the classes **founder**, **genotype**, **whichgt**, **nalleles** and **peakweight** are unchanged and are as described in Appendix A.2.

The alleleinmix class

The class **alleleinmix** represents the composition of the mixture, *i.e.* indicates whether the crime trace contains a certain allelic type. For the sake of brevity in the following lines the class **Aalleleinmix** only is taken into account, but the same structure applies to the other classes of this kind, *i.e.* **Balleleinmix**, **Calleleinmix**, **Dalleleimix xalleleinmix**. The class **Aalleleinmix** is shown in Figure C.9. Here, the input nodes, representing the genotypes of the three individuals `p1`, `p2` and `p3`, have identity link to the input node `gt` of the class **nalleles**. The node `Ainmix?` indicates whether a particular allelic type is in the mixture. Thus, it is *true* if at least one of the three unknown contributors has allele A in the genotype. This can be translated by the logical expression: (*if*(*and*(`n1A_nA`==0, `n2A_nA`==0, `n3A_nA`==0), *false*, *true*)), *i.e.* if all `n1A_nA`, `n2A_nA` and `n3A_nA` count 0 alleles, then Ainimix? is *false*, otherwise is *true*. Here n1A_nA, n2A_nA and n3A_nA are output nodes of the class **nalleles**. This node is an observation node, so that if allele A is measured in the mixture it is set to *true*, and the evidence on the mixture composition concerning allele A propagates from this node to the others.

Additionally, this class computes the mean contribution of a certain allelic type to the peak area. Input node frac1 is the proportion θ_1 of DNA originated from the first contributor p1. Input node frac2 is the proportion θ_2 of DNA originated from the second contributor p2. Both parameters are continuous variables but, for simplicity, discrete values are assigned to it in a scale ranging from $[0,5]$ with step 1 in order to allow evidence propagation in the Bayesian network. Nodes frac1 and frac2 are linked to node mean through the expression (mean==n1A_nA*frac1+n2A_nA*frac2+n3A_nA*(5-frac1-frac2)). This is the same mean of the relative peak weights found in equation (9.4), but it differs by a scale factor of 10. Thus, if we use the conditional-Gaussian model, before entering evidence on the relative peak weights, these have to be multiplied by 10. The state space of such node is discrete but contains some unrealistic values ranging from $[-10, 20]$ with step 1. Node mean is parent of the output node meanA. Since the conditions of sum in (9.3) and inequality in (9.5) in § 9.1 for the DNA proportions must hold (*i.e.* the conditions, respectively, $\theta_1 \geq \theta_2 \geq \theta_3$ and $\theta_1 + \theta_2 + \theta_3 = 5$), the node meanA can assume values from 0 to 10 only. As a consequence, it is defined by the expression: (*if*(*and*(mean≥0, mean≤10), mean, 99)), *i.e.* if its parent node assumes a value in the range $[0, 10]$, the node meanA copies the parent node mean, otherwise assumes the state value 99 representing all those "impossible" or unrealistic states. It is worth noting that, using a factorization for the conditional-Gamma model, the vector of likelihoods in (7.7) in § 7.2.1 is entered in this node meanA.

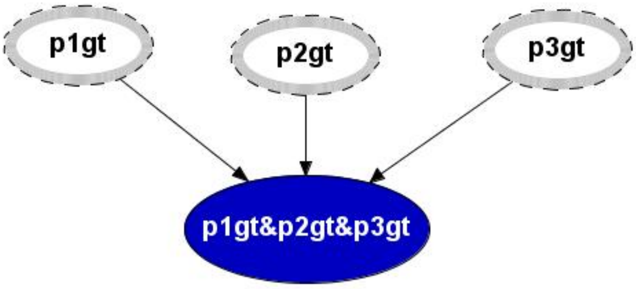


Figure C.10: Three person mixture. Jointgt class.

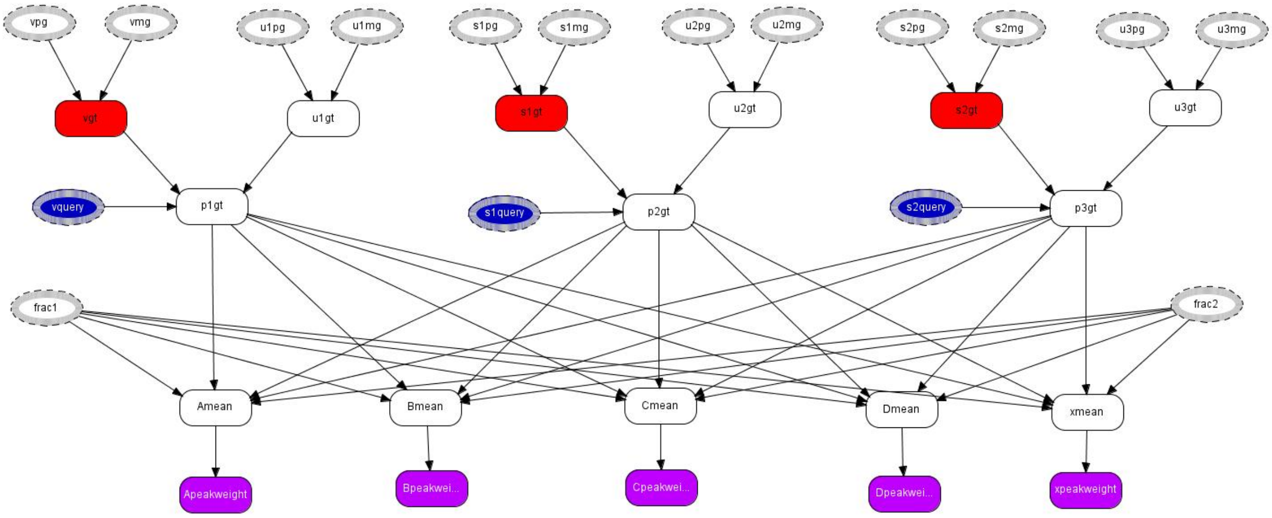


Figure C.11: Three person mixture. Marker class.

The joint class

The combined genotype of the three contributors to the crime trace, `p1`, `p2` and `p3`, is represented in the class **joint**. Thus, the node `p1gt&p2gt&p3gt` is the logical combination of the three input genotypes in `p1gt`, `p2gt` and `p3gt`. It is represented in Figure C.10.

The marker class

The **marker** class represents a specific marker and contains instances of the classes described so far since it is an upper level. This is shown in Figure C.11. All the input nodes `vpg`, `vmg`, `s1pg`, `s1mg`, `s2pg`, `s2mg`, `u1pg`, `u1mg`, `u2pg`,

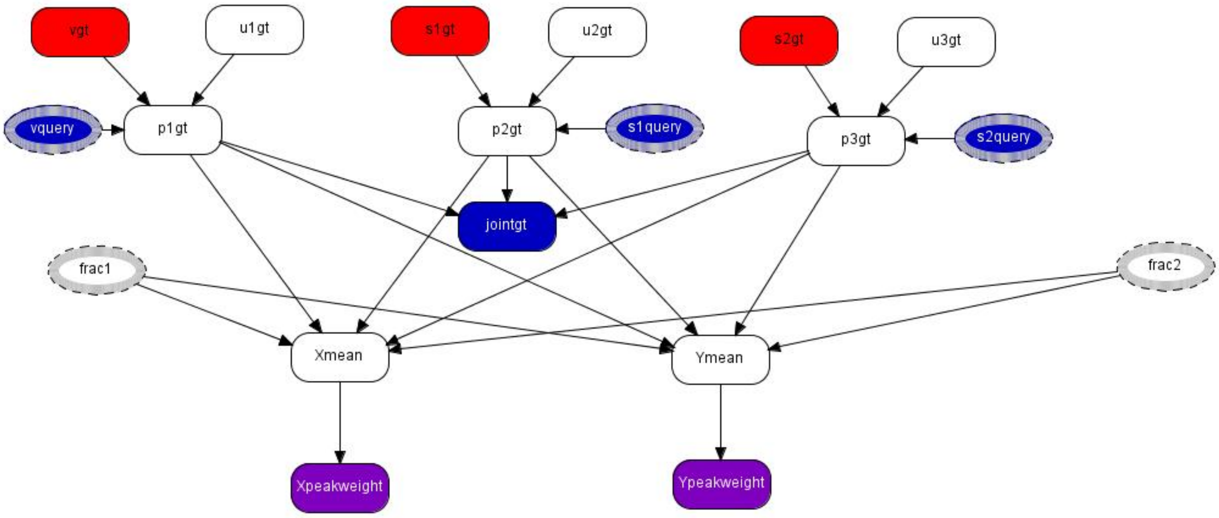


Figure C.12: Three person mixture. Amelogenin marker class.

`u2mg`, `u3pg` and `u3mg` have identity links to the node `founder` in the **founder** class. The nodes `vgt`, `s1gt`, `s2gt`, `u1gt`, `u2gt` and `u3gt` are instances of the **genotype** class. They contain respectively information on the victim's, the two suspects' and the three unknown individuals' genotypes. Evidence on victim and suspect is set in the nodes `gt` of **vgt**, **s1gt** and **s2gt**. Nodes `p1gt`, `p2gt` and `p3gt` are instances of the class **whichgt**. The Boolean node `s1query` is connected to the input query node `query?` in **p1gt**; the output node `gt` in **sgt** is connected to the input node `ingt` in **p1gt**; the output node `gt` in **u1gt** is connected to the input node `othergt` in **p1gt**. Thus, if the node `s1query` is *true* the output node `outgt` in **p1gt** is a copy of the node `gt` in **s1gt**, otherwise is a copy of the node `gt` in **s2gt**. Similarly for **p2gt** and **p3**. The nodes `Amean`, `Bmean`, `Cmean`, `Dmean` and `xmean` are all instances of the class **alleleinmix**. Their output node `meanA`, `Bmean`, `Cmean`, `Dmean` and `xmean` is linked to the input node `mean` in the class **peakweight**. The nodes `frac1` and `frac2` copy the corresponding nodes in the class **alleleinmix**.

The Amelogenin class

The **Amelogenin** class is shown in Figure C.12. This class has the same

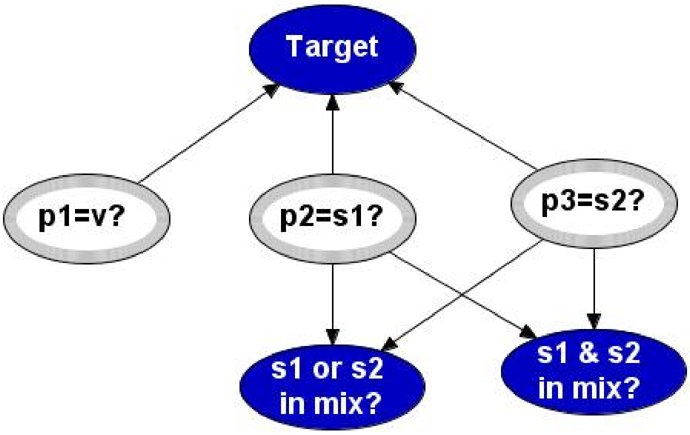


Figure C.13: Three person mixture. Target class.

structure of the **marker** class. No **founder** class is introduced. Nodes `vgt`, `s1gt`, `s2gt`, `u1gt`, `u2gt` and `u3gt` are instances of the class **genotype** for the **Amelogenin**. However, the class **genotype** used to build the class **Amelogenin**, has here a single output node `gt` with states XX for female and XY for male. The **whichgt** and **joint** classes are unchanged but have their state spaces reduced, *i.e.* they have two states only: XX and XY. In the class **nalleles** the node `nX` (`nY`) counts 1 (1) allele if the parent node `gt` is XY, whilst counts 2 (0) if the parent node `gt` is XX. The class **alleleinmix** is modified in the node `Xinmix` only which is always set to *true*. The **marker** class has only two instances of the class **alleleinmix** which are termed `Xmean` and `Ymean` and are connected to the nodes `Xpeakweight` and `Ypeakweight` instances of the **peakweight** class.

The Target class

Figure C.13 shows the **target** class. The **target** class contains the `Target` node where the results are read and the likelihood ratios are computed. This is the logical combination of the three Boolean nodes, `p1=v?`, `p2=s1?` and `p3=s2?`.

Since `p1=v?`, `p2=s1?` and `p3=s2?` have a uniform prior distribution, then the `target` node also has a uniform prior distribution. Thus, the states of `Target` are the hypotheses under test displayed in Table C.1. The node

Hypotheses under test	
s1&s2&v	both suspects and victim contributed to the mixture
s1&s2&u	both suspects and an unknown individual contributed to the mixture
s1&v&u	the first suspect, the victim and an unknown individual contributed to the mixture
s2&v&u	the second suspect, the victim and an unknown individual contributed to the mixture
s1&2u	the first suspect and two unknown individuals contributed to the mixture
s2&2u	the second suspect and two unknown individuals contributed to the mixture
v&2u	the victim and two unknown individuals contributed to the mixture
3u	three unknown individuals contributed to the mixture

Table C.1: Hypotheses under test.

`s1_or_s2_in_mix` indicates the presence of at least one of the two suspects in the mixture. This is *true* if either `p2=s1?` or `p3=s2?` is *true*. The node `s1&s2_in_mix` indicates the presence of both suspects in the mixture. This is *true* if both `p2=s1?` and `p3=s2?` are *true*. It is worth noting that in this context we assume that the identified individuals are two suspects and one victim. However, this class can be easily handled also for crimes where instead the profiles of three possible suspects are available. In this scenario the node `s1_or_s2_in_mix` would be called `s1_or_s2_or_s3_in_mix` and would depend on the three parent nodes `p1=s1?`, `p2=s2?` and `p3=s3?`. Therefore, it would be *true* if at least one among the nodes `p1=s1?` `p2=s2?` and `p3=s3?` is *true*. Similarly for the node `s1&s2_in_mix`.

The master class

Figure C.14 shows the **master** class. In the **master** class nodes `D7`, `D8` and `D21` are all instances of **marker** class. For each marker, there are 12 instances of the class **founder** linked to the 12 input nodes of the class **marker**. Nodes `frac1` and `frac2` represent, respectively, the DNA proportion generating from the contributor `p1` and from `p2`. They are connected to the corre-

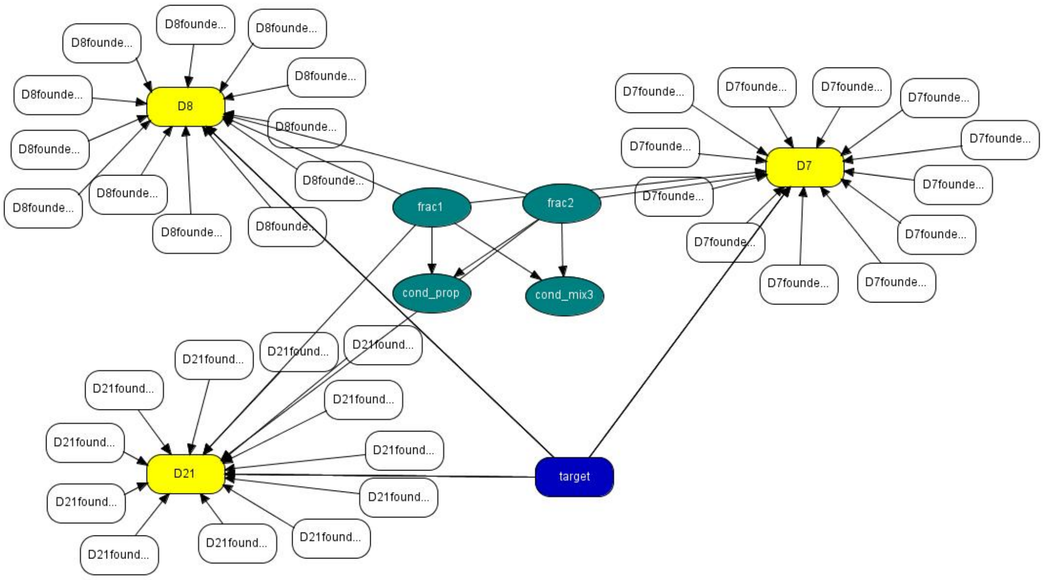


Figure C.14: Three person mixture. Master class.

sponding nodes in the **markers**. Node `amel` represents the **Amelogenin** class and therefore it does not need **founder** classes. Nodes `cond_3mix` and `cond_prop` represent the conditions (9.3) and (9.5) in § 9.1 that must hold for the proportions of DNA θ_1 and θ_2. Thus, the node `cond_prop` is defined by the expression: (*if* (*and* (`frac1`≥`frac2`, `frac2`≥(5-`frac1`-`frac2`), *true*, *false*)), *i.e.* if $\theta_1 \geq \theta_2 \geq (5 - \theta_1 - \theta_2)$, the node `cond_prop` is set *true*, otherwise is set *false*. If the node `cond_prop` are set *true*, then $\theta 1$ represents the DNA proportion originated from the first major contributor, whilst $\theta 2$ represents the DNA proportion originated from the second major contributor. On the contrary, the node `cond_3mix` is defined by the expression: (*if* (`frac1`+`frac2`<5, *true*, *false*), *i.e.* if $\theta_1 + \theta_2 < 5$, the node `cond_3mix` is *true*, otherwise is *false*, where the factor 5 is due to the fact that θ has been discretized assigning to it values in a scale ranging from $[0, 5]$ with step 1.

Finally, `Target` node is an instance of class **target** and is linked to each marker via its output nodes `p1=v?`, `p2=s1?` and `p3=s2?`.