1

Graphical Models for Forensic Analysis

A. Philip Dawid, University of Cambridge, UK
Julia Mortera, Università Roma Tre, Italy

CONTENTS

List of Figures

List of Tables

Contents

1.1 Introduction

"Forensic" means pertaining to courts of law. Here we are concerned with systems to assist in the evaluation of evidence presented in a criminal or civil court case. Such a case may have a mixed mass of evidence of many kinds, all of it hedged about with uncertainty. We describe how such a case can be helpfully represented by means of a *Bayesian Network* (BN), or *Probabilistic Expert System* (Cowell *et al.* 1999): a directed graphical model describing the various items of evidence and hypotheses, and the probabilistic relationships between them. Such a representation displays clearly the relevance of the evidence to questions of interest, and supports efficient routines to compute the impact of the evidence presented. In many cases the BN can be constructed as an *object-oriented Bayesian network* (OOBN), a top-down hierarchical structure which hides irrelevant detail and simplifies both construction and interpretation.

In § 1.2 we describe by means of a fictitious example the way in which different elements in a case (eye-witness, fibre and blood evidence) can be drawn together into a single coherent story structured as a Bayesian network. We use this example to explain how a BN can be used to discover implicit relationships of relevance and irrelevance in the evidence, which in turn can be used to simplify probabilistic calculations. § 1.3 describes the features of an OOBN, and shows how simple reusable "idioms" can be constructed to represent common features and relationships such eyewitness testimony and identification. § 1.4 briefly describes how a BN can be used to simplify the specification and manipulation of probabilities, in particular the use of "evidence propagation" to compute conditional probabilities taking the evidence into account.

The remainder of this Chapter focuses on DNA evidence (the Appendix gives a very brief glossary of the relevant biological background and terminology). § 1.5 gives examples of the use of OOBNs to handle cases of criminal identification, simple and complex disputed paternity. These examples deal with cases where "clean" single source DNA profiles are available, whereas § 1.6 shows how the methods can be extended to deal with more complex cases, where (for example) a crime trace may contain a mixture of DNA from more than one contributor, in varying proportions. Finally, § 1.7 relaxes some of the simplifying assumptions made so far, to account for such realistic complications as uncertainty about allele frequencies and heterogeneity in the reference population. Network modules that account for these additional features are introduced; these can then be integrated into the variety of identification problems previously described.

1.2 Bayesian Networks for the Analysis of Evidence

In a legal case, we may have various items of evidence, both lay and scientific, with more or less complex relationships. It can often be helpful to represent such relationships in graphical form, as a Bayesian Network (BN). As described in Part I, Chapter 1 of this Handbook, "Conditional independence concept and Markov properties for basic graphs" by Milan Studený, a BN is a directed acyclic graph (DAG), with nodes representing relevant variables in the problem, joined by arrows representing probabilistic dependence, and, for each "child" node in the DAG, a specification of its conditional distribution, given the states of its "parents". This can then be used for further analysis, both qualitative and quantitative. We start by considering purely qualitative properties.

Example 1 (Robbery) We illustrate with a fictional crime story (reproduced with permission from Dawid and Evett (1997).)

Eye witness evidence: An unknown number of offenders entered commercial premises late at night through a hole which they cut in a metal grille. Inside, they were confronted by a security guard who was able to set off an alarm before one of the intruders punched him in the face, causing his nose to bleed. The security guard said that there were four men but the light was too poor for him to describe them and he was confused because of the blow he had received. About 10 minutes later the police found the suspect trying to "hot wire" a car in an alley about a quarter of a mile from the incident. The suspect denied having anything to do with it.

Fibre evidence: A tuft of red acrylic fibres was found on the jagged end of one of the cut edges of the grille. The suspect's jumper was red acrylic. The tuft was indistinguishable from the fibres of the jumper by eye, microspectrofluorimetry and thin layer chromatography.

Blood evidence: A spray pattern of blood was found on the front and right sleeve of the suspect's jumper. The blood on the jumper was of a different type from that of the suspect, but the same as that of the security guard.

The directed acyclic graph of Figure 1.1 contains a number of nodes, corresponding to random events or variables; here a square node corresponds to a variable that has been observed, while a round node indicates an unobserved variable that is required to complete the picture. The arrows leading into any node identify the variables on whose value it is supposed to depend, probabilistically. For example, Y_2, the measurement of the blood type

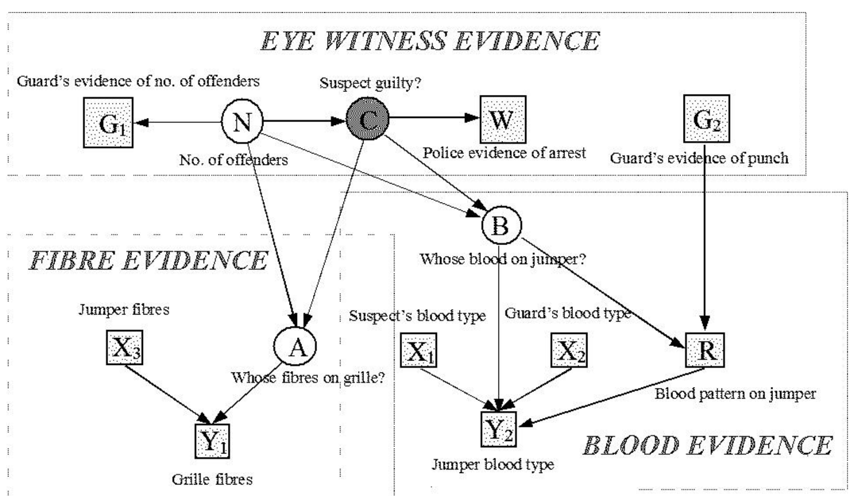
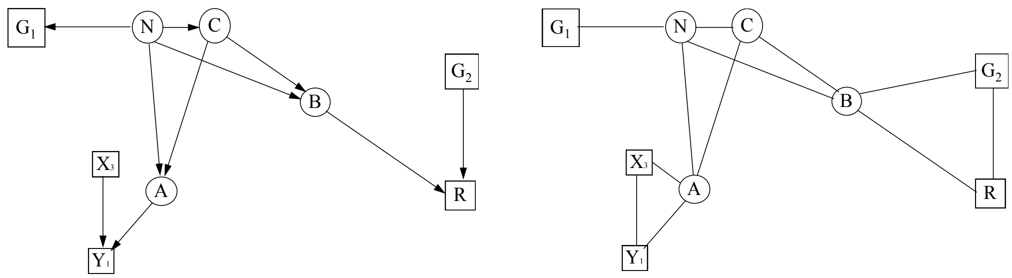


FIGURE 1.1: Directed acyclic graph representing robbery story (adapted with permission from Dawid and Evett (1997).)

of the spray on the jumper is dependent on X_1, the suspect's blood type (because it might be a self stain) and the guard's blood type X_2. But information is also provided by R, describing the shape of the stain, because that sheds light on whether or not it might be a self stain. In turn, the shape of the stain is influenced by the way in which the guard was punched, G_2, and B, the identity of the person who did it. B is in turn influenced by



(a) Ancestral graph (b) Moralised ancestral graph

FIGURE 1.2: $(B, R) \perp\!\!\!\perp (G_1, Y_1) \mid (A, N)$? (adapted with permission from Dawid and Evett (1997).)

variable C, whether or not the suspect was one of the offenders, and also by N, the number of offenders.

This construction of the DAG utilises the concept of conditional independence (Dawid 1979). For example, were we to know N, the number of offenders, and C, whether or not the suspect is one of them, our uncertainty about B, the identity of the person who struck the guard, would (it is supposed) be unaffected by further information about both the eye-witness variables, (G_1, G_2, W), and the fibre variables, (A, X_3, Y_1). In conditional independence notation (Dawid 1979): $B \perp\!\!\!\perp (G_1, G_2, W, A, X_3, Y_1) \mid (N, C)$. This is an example of the general requirement that a variable be independent of its "non-descendants", given its "parents". Similarly, once B is known, then G_1, N, C and W become irrelevant to any variables that are descendants of B in the graph, such as Y_2: $Y_2 \perp\!\!\!\perp (G_1, N, C, W) \mid B$. Note that conditional independence, so interpreted, is a purely qualitative "irrelevance" property, and does not require numerical assessment of any probabilities in the problem. (However, it does impose relationships between these probabilities.)

Further, we have methods for examining a DAG to discover additional, implicit, conditional independence properties. One such method is the "moralisation" criterion of Lauritzen *et al.* (1990), which operates as follows. Let A, B, C be sets of nodes. To query the conditional independence $A \perp\!\!\!\perp B \mid C$:

Ancestral graph Form the subgraph containing just the nodes in A, B and C and their ancestors.

Moralisation "Marry" any unmarried parents of the same child; drop arrows.

Separation Look for a path from A to B avoiding C. If there is none such, deduce $A \perp\!\!\!\perp B \mid C$.

For a description of other, equivalent, graphical criteria, refer again to Part I, Chapter 1 of this Handbook.

As an example, suppose we wish to query the conditional independence property $(B, R) \perp\!\!\!\perp (G_1, Y_1) \mid (A, N)$. Figure 1.2 shows the relevant ancestral graph and its moralisation. We note that, in the latter, every path from B or R to G_1 or Y_1 passes through either A or N, and so deduce that this conditional independence does indeed hold.

Such properties can be helpful in simplifying algebraic manipulations on probabilities. Thus we can express the posterior odds on guilt ($C = c$), given evidence $(G_1, G_2, W, R, X_1, X_2, X_3, Y_1, Y_2) = (g_1, g_2, w, r, x_1, x_2, x_3, y_1, y_2)$, as

$$\frac{\Pr(c \mid g_1, g_2, w, r, x_1, x_2, x_3, y_1, y_2)}{\Pr(\bar{c} \mid g_1, g_2, w, r, x_1, x_2, x_3, y_1, y_2)} = \frac{\Pr(r, x_1, x_2, x_3, y_1, y_2 \mid c, g_1, g_2, w)}{\Pr(r, x_1, x_2, x_3, y_1, y_2 \mid \bar{c}, g_1, g_2, w)} \times \frac{\Pr(c \mid g_1, g_2, w)}{\Pr(\bar{c} \mid g_1, g_2, w)}.$$

The scientific evidence enters only into the first term on the right-hand side, which has the form of a conditional likelihood ratio, given the eyewitness evidence. This term can be simplified by applying the following conditional independence properties, all of which follow from application of the moralisation criterion:

$$\begin{aligned}(X_1, X_2, X_3) &\perp\!\!\!\perp (C, G_1, G_2, W)\\ (R, Y_1, Y_2) &\perp\!\!\!\perp W \mid (C, X_1, X_2, X_3, G_1, G_2)\\ Y_1 &\perp\!\!\!\perp (R, Y_2) \mid (C, X_1, X_2, X_3, N, G_2)\\ Y_1 &\perp\!\!\!\perp (X_1, X_2, G_2) \mid (X_3, C, N)\\ (R, Y_2) &\perp\!\!\!\perp X_3 \mid (X_1, X_2, C, N, G_2).\end{aligned}$$

With a further assumption $G_1 = N$ (the guard's evidence of the number of offenders is accurate), the above properties allow us to simplify the conditional likelihood ratio:[1]

$$\frac{\Pr(r, x_1, x_2, x_3, y_1, y_2 \mid c, g_1, g_2, w)}{\Pr(r, x_1, x_2, x_3, y_1, y_2 \mid \overline{c}, g_1, g_2, w)} = \frac{\Pr(y_1 \mid x_3, c, g_1)}{\Pr(y_1 \mid x_3, \overline{c}, g_1)} \times \frac{\Pr(r, y_2 \mid x_1, x_2, c, g_1, g_2)}{\Pr(r, y_2 \mid x_1, x_2, \overline{c}, g_1, g_2)}.$$

□

1.3 Object-oriented networks

Many problems have a hierarchical or repetitive structure that is not best represented by a "flat" network such as that of Figure 1.1. An "object-oriented Bayesian network" (OOBN) allows such additional structure to be taken into account, to simplify the construction, display and interpretation of the network. In an OOBN, what looks like a single node in a network can in fact be a network in its own right. This generalisation of a BN was first proposed by Laskey and Mahoney (1997).

As an example, Figure 1.3 (created using the commercial software package HUGIN) gives a high-level view of the network of Figure 1.1, showing that, conditional on the unknown identification nodes (N, C), the fibre evidence is independent of the blood and eyewitness evidence—whereas the blood evidence remains dependent on the eyewitness evidence (in fact, through the node G_2).

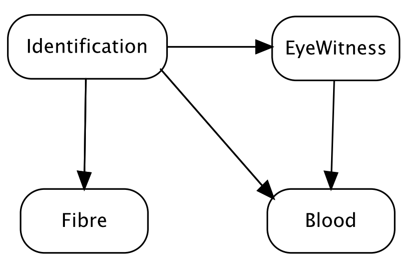


FIGURE 1.3: OOBN for robbery

The internal structure of the individual submodules is shown in Figure 1.4. A thick grey rim denotes an output node, which can be identified with an input node (dashed grey rim) in another module. This is done as shown in Figure 1.5.

[1] Still further simplification is possible, using reasonable properties not all of which are represented in the graph: see Dawid and Evett (1997) for details.

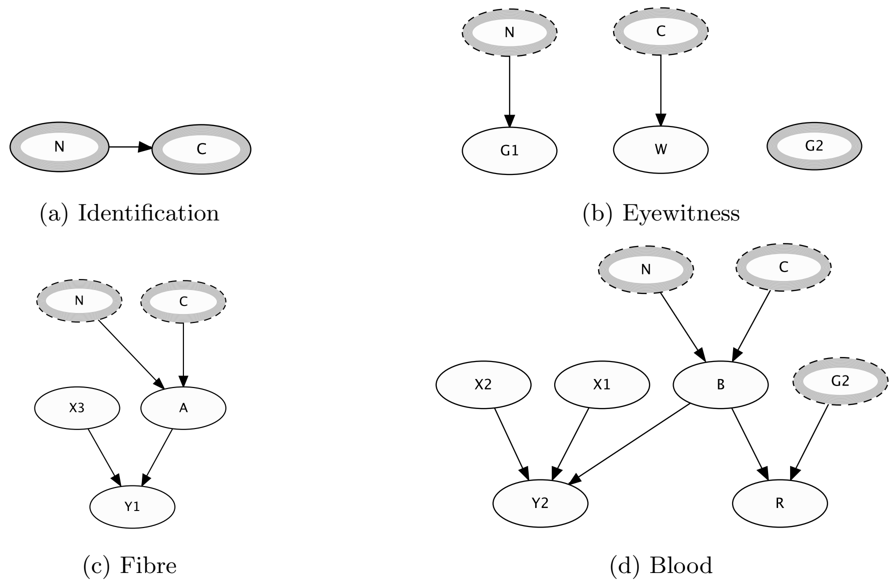


FIGURE 1.4: Submodules for robbery OOBN

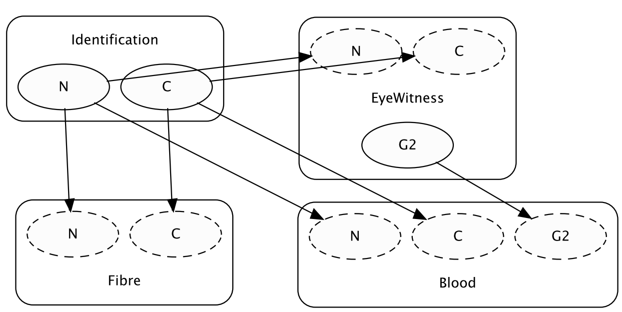


FIGURE 1.5: Expanded view of robbery OOBN

1.3.1 Generic idioms

A particularly valuable application of OOBNs is to develop generic network fragments, or "idioms", that can be reused, both within and across numerous high-level networks (Neil *et al.* 2000; Hepler *et al.* 2007; Fenton *et al.* 2013). We here term such a fragment a *module*, and set it in a **boldface** font. An instance of a module, like any other node in a larger network, will be set in `teletype`, while a value (state) of a node will by indicated by *italic*.

One recurrent idiom describes generic features of eyewitness testimony of an event (Schum and Morris 2007). This is structured into three stages: observational sensitivity ("sensation"), objectivity, and veracity, as represented in the network of Figure 1.6, which builds on the submodules shown in Figure 1.7. "Observational sensitivity" refers to the

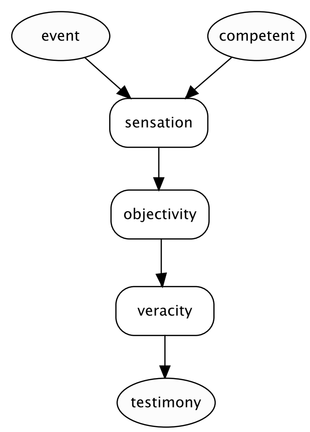


FIGURE 1.6: Testimony idiom (adapted with permission from Dawid *et al.* (2011).)

possibility of mistakes in the witness's perception of the event, due either to his sensory and general physical condition (leading to possible disagreement between the actual and the perceived features of the event), or to the conditions under which the observation is made. The latter aspect is termed "competence". For example, if the witness was hiding under the table, he would not have been competent to observe what was happening. These two processes are incorporated into Figure 1.7a, which models `agreement` as an instance of the generic **filter** module of Figure 1.7b, which allows a random "error" to affect whether or not the output node reproduces the input. The subnetworks `objectivity` and `veracity` in Figure 1.6 are likewise instances of the **filter** module: "Objectivity" means that the witness's belief is a correct interpretation of the evidence of his senses, while "veracity" means that he truthfully reports his belief.

Other recurrent evidential idioms include "identification" (as in Figure 1.9 below), "contradiction", "corroboration", "conflict", "convergence", "explaining away". See Hepler *et al.* (2007) for details.

1.4 Quantitative analysis

Our discussion so far has largely concentrated on the qualitative aspects of a BN representation. Such a representation also allows simplification of the tasks of assigning and manipulating probability distributions for the variables. Rather than specify a very large

(a) Sensation (b) Filter **JM:** Accuracy **:end**

FIGURE 1.7: Testimony modules (adapted with permission from Dawid *et al.* (2011).)

collection of joint probabilities for all the variables in the problem, it is enough (and generally much easier) to specify, for each node, its conditional distribution, given the configuration of states of its "parent" variables. It is then possible to extract, using elegant and efficient computational algorithms, both exact and (for more complex problems) approximate, the marginal distribution for any variable, or ("evidence propagation") its conditional distribution, after taking into account observed values for certain other variables. For details, see Part II, Chapters 1 and 2 of this Handbook. There exist a number of software systems that conduct such computations, including Hugin[2], GeNIe[3], netica[4], AgenaRisk[5],gRain[6], Grappa[7] and (for approximate inference) WinBUGS[8]. All networks shown in this Chapter were created and analysed using Hugin.

1.5 Bayesian networks for forensic genetics

Forensic DNA evidence has special features, principally owing to its pattern of inheritance from parent to child (a very brief introduction to the basic genetics is given in the Appendix.) These make it possible to use it to address queries such as the following:

Criminal case: Did A leave the trace at the scene of the crime?

Disputed paternity: Is individual A the father of individual B?

Immigration: Is A the mother of B? How is A related to B?

Criminal case: mixed trace: Did A and B both contribute to a stain found at the scene of the crime? Who contributed to the stain?

Disputed inheritance: Is A the daughter of deceased B? Is A the son of a contributor to the mixture?

[2] http://www.hugin.com
[3] http://www.bayesfusion.com/genie-modeler
[4] https://www.norsys.com
[5] http://www.agenarisk.com/
[6] https://CRAN.R-project.org/package=gRain/
[7] https://people.maths.bris.ac.uk/~mapjg/Grappa//
[8] http://www.mrc-bsu.cam.ac.uk/software/bugs

Disasters: Was A among the individuals involved in a disaster? Who were those involved?

In a simple criminal identification case we have evidence E that a suspect's DNA profile matches that found at the crime scene. The prosecution hypothesis H_p is that the suspect left the DNA trace, while the alternative defence hypothesis, H_d, might be that another individual randomly drawn from some reference population left the trace. In a simple disputed paternity case, the evidence E will comprise DNA profiles from mother, child and putative father. Hypothesis H_p is that the putative father is the true father, while hypothesis H_d might be that the true father is some other individual randomly drawn from the population. We can also entertain other hypotheses, such as that one of one or more other identified individuals is the father, or that the true father is the putative father's brother.

In a complex criminal case, we might find a stain at the scene of the crime having the form of a *mixed trace*, containing DNA from more than one individual. DNA profiles are also taken from the victim and a suspect. We can entertain various hypotheses as to just who—victim?, suspect?, person or persons unknown?, contributed to the mixed stain.

When we are only comparing two hypotheses H_0 and H_1, the impact of the totality of the DNA evidence E available, from all sources, is crystallised in the *likelihood ratio*, $\text{LR} = \Pr(E \mid H_1)/\Pr(E \mid H_0)$. If we wish to compare more than two hypotheses, we require the full *likelihood function*, a function of the various hypotheses H being entertained (and of course the evidence E):

$$\text{lik}(H) \propto \Pr(E \mid H). \tag{1.1}$$

The proportionality sign in (1.1) indicates that we have omitted a factor that does not depend on H, although it can depend on E. Such a factor is of no consequence and need not be specified, since it disappears on forming ratios of likelihoods for different hypotheses on the same evidence. Only such relative likelihoods are required, not absolute values.

We also now need to specify the prior probabilities, $\Pr(H)$, for the full range of hypotheses H. Then posterior probabilities in the light of the evidence are again obtained from **JM:** Bayes's or Bayes' **:end** theorem, which can now be expressed as:

$$\Pr(H \mid E) \propto \Pr(H) \times \text{lik}(H). \tag{1.2}$$

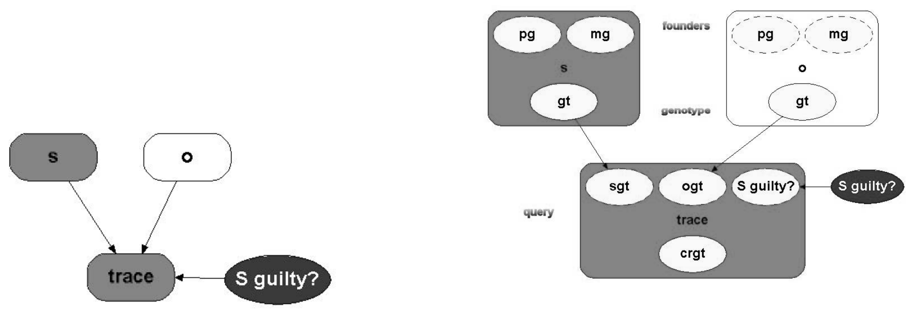
Again the omitted proportionality factor in (1.2) does not depend on H, although it might depend on E. It can be recovered, if desired, as the unique such factor for which the law of total probability, $\sum_H \Pr(H \mid E) = 1$, is satisfied.

1.5.1 Bayesian networks for simple criminal cases

In a simple criminal DNA identification case, the evidence is that the suspect's DNA profile matches a trace found at the scene of the crime. We are interested in testing **JM:** –two mutually exclusive and exhaustive hypotheses– **:end** the prosecution hypotheses H_p : *the crime trace belongs to the suspect* `s` (loosely, 'the suspect is guilty'); versus the defense hypothesis H_d : *the crime trace belongs to another actor*, `o`, *randomly drawn from the population.* Representation of such problems as Bayesian networks was introduced by Dawid *et al.* (2002), and as object-oriented Bayesian networks by Dawid *et al.* (2007).

The OOBN for the case is shown in Figure 1.8, together with its expanded version. Nodes `s` and `o` are each instances of a **founder** network module, with nodes paternal gene `pg`, maternal gene `mg`, and genotype `gt`. Each of the (input) nodes `pg` and `mg` is identified with the single (output) node `gene` of an instance of a simple module **gene**; while the (output) node `gt` is constructed as the unordered combination of `pg` and `mg`. Node `trace` is an instance of the **query** network module shown in Figure 1.9 whereby `trace` is modelled as equal to `sgt` or `ogt`, according as `S guilty?` is *true* or *false*, respectively.

Each genetic marker m is analysed separately. Node `gene` in module **gene** is assigned a distribution corresponding to the allele frequencies for that marker, and `S guilty?` is assigned probability 0.5 for *true*. The observed genotype is entered as evidence at `gt` in `s`, and again at `crgt` in `trace`, and propagated through the network. The resulting odds on *true* at `S guilty?` can then be interpreted as the likelihood ratio in favour of H_p, based on the evidence of a match at marker m. Finally, multiplying these values across all markers delivers the overall likelihood ratio based on the full match.



(a) OOBN, showing actors `s` and `o`, crime trace and hypotheses

(b) Expanded network, revealing variables within each block node

FIGURE 1.8: Network for criminal identification (adapted with permission from Green and Mortera (2009).)

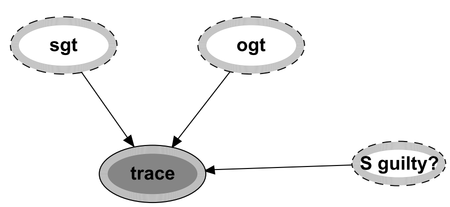


FIGURE 1.9: **query** module: `trace` is identical to `sgt` or `ogt` according as `S guilty?` is *true* or *false*

1.5.2 Bayesian network for simple paternity cases

In a simple case of disputed paternity, a man is alleged to be the father of a child, but disputes this. DNA profiles are obtained from the mother `m`, the child `c`, and the putative father `pf`. On the basis of these data, we wish to assess the likelihood ratio for the hypothesis of *paternity*: H_p: `tf` = `pf`, the true father is the putative father; as against that of *non-paternity*: H_d: `tf` = `af`—where `af` denotes an unspecified alternative father, treated as unrelated to `pf` and randomly drawn from the population.

The disputed pedigree can be represented by the OOBN of Figure 1.10. Nodes `m`, `pf` and `af` are instances of the network module **founder** as in § 1.5.1, while node `tf` is an instance of **query**—its output is a genotype copied from that of `pf` or `af`, according as `tf` = `pf`? is *true* (H_p) or *false* (H_d). Node `c` is an instance of a network module **child**, containing two copies (one for each parent) of the module **mendel**, shown in Figure 1.11, whereby, according to

Mendel's law, the child `c` inherits its parental gene `cg` by a random draw (represented as a fair coin flip `fcoin`) from the maternal and paternal genes, `mg` and `pg`, of the relevant parent.

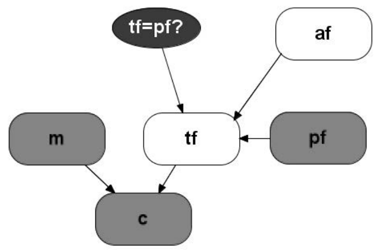


FIGURE 1.10: Pedigree for simple disputed paternity (adapted with permission from Green and Mortera (2009).)

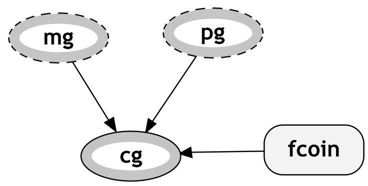


FIGURE 1.11: Module representing Mendelian inheritance

As in § 1.5.1, we analyse the markers one at a time. For each, we assign the relevant allele frequency distribution at each founder gene, enter the observed evidence at `m`, `pf` and `c`, and propagate. This yields a likelihood ratio based on that marker data, and we multiply all these together to obtain the overall likelihood ratio based on the full collection of markers. This can then be combined with the prior odds of paternity, based on external background evidence B, to obtain the posterior odds on paternity.

1.5.3 Bayesian networks for complex cases

A major advantage of OOBN representations is that they make it easy to elaborate the network with additional features (Dawid *et al.* 2007). For example, in the presence of possible mutation, we can modify the network of Figure 1.11 to allow either `mg` or `pg` to mutate, before being possibly selected for transmission to `cg`. Various different mutation models can be constructed and incorporated. Other possible modifications include, for example, allowance for alleles that are not picked up by the instrumentation—a property that can be either inherited (a "silent" allele) or sporadic (a "missed" allele.) Such modifications can typically be confined to low-level networks; the other modules, and the overall high-level structure, are unchanged.

Another advantage is the ability to reuse existing network modules in new combinations, to tell different stories. For example, Figure 1.12 puts together instances of **founder** (at `gf`, `gm`, `m1`, `m2` amd `af`), of **child** (at `pf`, `b1`, `b2`, `c1`, `c2`) and of **query** (at `tf`), to analyse a case where it was impossible to collect DNA from `pf`, the putative father of the child `c1` of mother `m1`, but DNA is available from his two full brothers `b1` and `b2` (all children of grandfather `gf` and grandmother `gm`) and his undisputed child `c2` by a different mother `m2`, as well as from `m1`, `m2` and `c1`.

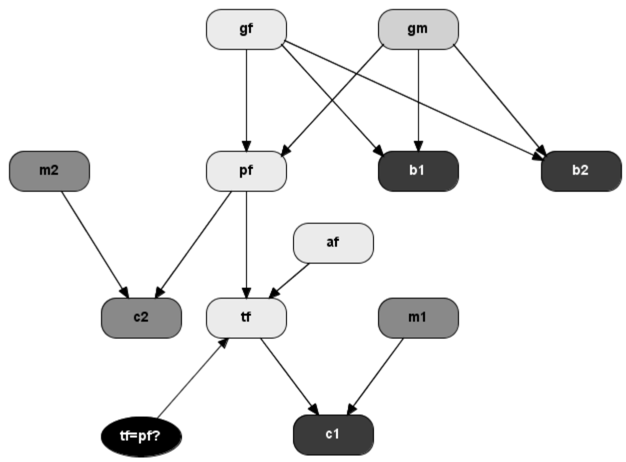


FIGURE 1.12: Disputed paternity with absent putative father

Moreover, the building blocks used in such constructions can themselves be modified, as described above, to incorporate additional features such as mutation.

1.6 Bayesian networks for DNA mixtures

When several actors have contributed to a DNA trace found at a crime scene we will have a *mixed* DNA profile. The presence of 3 or more alleles on any marker indicates that the trace is a mixture from more than one contributor. One might be interested in testing whether the victim and suspect contributed to the mixture, H_p: $v\,\&\,s$, against the hypothesis that the victim and an unknown individual contributed to the mixture, H_d: $v\,\&\,u$. One might alternatively consider an additional unknown individual u_2 instead of the victim, with hypotheses H'_p: $u_2\,\&\,s$ versus H'_d: $u_2\,\&\,u_1$.

1.6.1 Qualitative data

We first describe Bayesian networks for analysing purely *qualitative* data, describing simply which alleles are observed in the trace. Figure 1.13 shows a top-level network which can be used for analysing a mixture with two contributors, $p1$ and $p2$, and a marker in the trace having three alleles A, B and C (the network can be simply modified to account for different numbers of alleles). Nodes `sgt`, `vgt`, `u1gt` and `u2gt` are instances of the network class **founder**, and represent the suspect's, the victim's and two unknown individuals' genotypes. Node `p1gt`, the genotype of $p1$, is an instance of **query**, which selects between the two genotypes `sgt` or `u1gt` according to the *true/false* state of the Boolean node `p1=s?`, representing the hypothesis that contributor $p1$ is the suspect s. A similar relationship holds between nodes `p2gt`, `vgt`, `u1gt` and `p1=v?`. The `target` node is the logical combination of the two Boolean nodes `p1=s?` and `p2=v?` and represents the four different hypotheses described above. Node `Ainmix?` determines whether allele A is in the mixture: this will be so if at least one A allele is present in either `p1gt` or `p2gt`. Similarly for `Binmix?`, `Cinmix?` and `Dinmix?` (where `D` refers to all the alleles that are not observed).

For each marker the gene nodes are populated with the relevant allele frequency distribution, and nodes `p1=s?`, `p1=v?` are modelled as coin-flips. Any available genotype information on the suspect and the victim is entered into nodes `sgt` and `vgt`, *true* is entered at `Ainmix?` and `Binmix?`, and `Cinmix?`, and *false* at `Dinmix?`. This evidence is propagated, after which the probability distribution over the four hypotheses at `target` can be interpreted as a likelihood function, based on the data for that marker. Again, an overall likelihood function is obtained on multiplying these across markers.

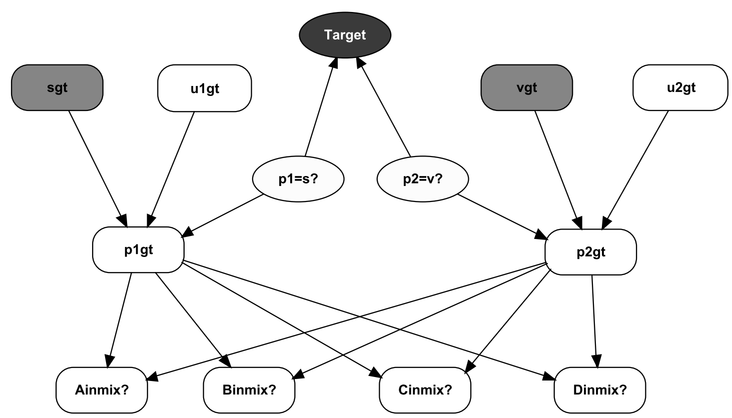


FIGURE 1.13: Bayesian network for DNA mixture from two contributors.

The modular structure of Bayesian networks supports easy extension to mixtures with more contributors by simple modification of the network, so long as the total number of contributors can be assumed known. **JM:** *Possibly rewrite this sentence* Or, if it can be agreed to limit attention to some maximum total number of potential contributors (Lauritzen and Mortera 2002), cases where the number of unknown contributors is itself uncertain can be addressed using a Bayesian network, now including nodes for the number of unknown contributors and the total number of contributors (Mortera *et al.* 2003). **:end** This can be used for computing the posterior distribution of the total number of contributors to the mixture, as well as likelihood ratios for comparing all plausible hypotheses. The modular structure of the Bayesian networks can be used to handle still further complex mixture problems. For example, we can consider together missing individuals, silent alleles and a mixed crime trace simply by piecing together the appropriate modules.

1.6.2 Quantitative data

The networks above only use the qualitative information as to which allele values are present in the mixture and the other profiles. A more sensitive analysis additionally uses measured continuous "peak heights", which give quantitative information on the amounts of DNA involved. This requires much more detailed modelling, but again this can be effected by means of a Bayesian network (Cowell *et al.* 2007b). **JM:** Expand a bit more on the next part **:end** Because the mixture proportion `frac` of DNA contributed by one of the parties is a quantity common across all markers, we must now handle the markers all simultaneously within one "super-network". Figure 1.14 shows the top level network for two contributors, involving six markers (D8, vwa, D21, D18, FGA, Tho1), each an instance of a lower level network **marker** as shown in Figure 1.15. This network is an extended version of the one shown in Figure 1.13, incorporating additional structure to model the quantitative peak height information. In particular, the nodes `Aweight` *etc.* in **marker** are instances of a module that models the quantitative information on the peak height.

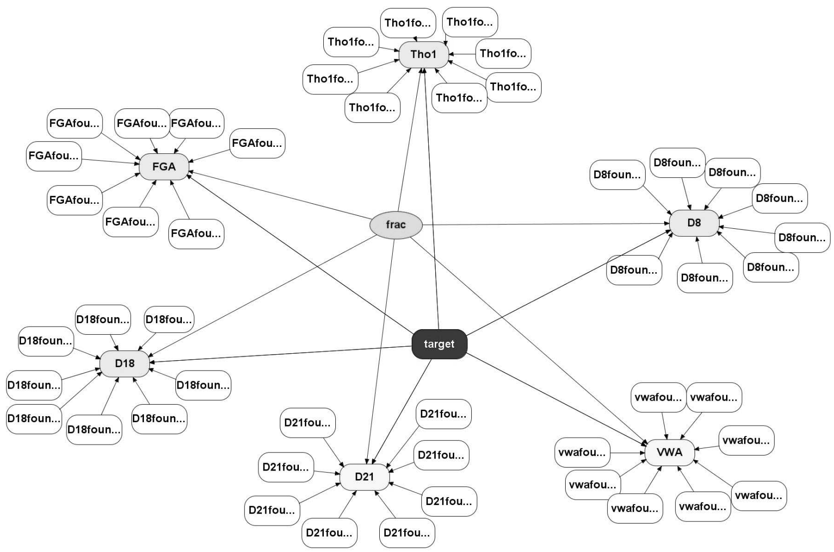


FIGURE 1.14: 6-marker OOBN for mixture using peak areas, 2 contributors (reproduced from Cowell *et al.* (2004)).

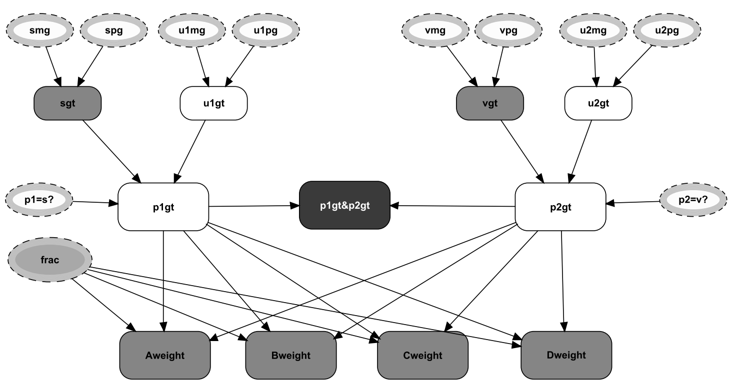


FIGURE 1.15: Network **marker** with four observed allele peaks (reproduced from Cowell *et al.* (2004)).

Cowell *et al.* (2007a); Cowell *et al.* (2007b) analyse the data shown in Table 1.1, taken from Evett *et al.* (1998), involving a 6-marker mixed profile with between 2 and 4 distinct observed alleles and corresponding peak areas per marker, and a suspect whose profile is contained in these. It is assumed that this profile is a mixture either of the suspect and one other unobserved contributor, or of two unknown contributors. Using only the alleles as data, the likelihood ratio for the suspect being a contributor to the mixture is calculated to be around 25,000. On taking account of the peak areas also, this rises 6,800-fold, to about 170,000,000.

Marker	D8			D18			D21			
Alleles	10*	11	14*	13*	16	17	59	65	67*	70*
Peak area	6416	383	5659	38985	1914	1991	1226	1434	8816	8894

Marker	FGA			THO1		VWA			
Alleles	21*	22*	23	8*	9.3*	16*	17	18*	19
Peak area	16099	10538	1014	17441	22368	4669	931	4724	188

TABLE 1.1: Data for mixed trace with two contributors. The starred values are the suspect's alleles.

1.6.3 Further developments on DNA mixtures

Cowell *et al.* (2007a); Cowell *et al.* (2011); Cowell *et al.* (2015) **JM:** further extend the statistical model in § 1.6.2 for the quantitative peak information obtained from an electropherogram of a forensic DNA sample. **:end** A gamma model is used for the peak heights and the model further develops the modelling of various artefacts that can occur in the DNA amplification process. Thus *dropout* of an allele occurs when its associated peak fails to exceed the detection threshold. Another common artefact is *stutter*, whereby an allele at repeat number a that is present in the sample is mis-copied, and appears as a peak at repeat number $a-1$. Yet another artefact is *dropin*, referring to the occurrence of small unexpected peaks in the DNA amplification: this can, for example, be due to sporadic contamination of a sample, either at source or in the forensic laboratory. Current technology allows for the amplification of very small amounts of DNA, even as little as contained within one cell; in such a case many of these artefacts can occur. These artefacts are simply represented in a coherent way in this model.

The model can both find likelihood ratios for evidential calculations, and deconvolve a DNA mixture for the purpose of finding likely profiles of one or more unknown contributors to the mixture. Computation from this model rely on an efficient implementation of Bayesian network techniques. This allows for readily extension to simultaneous analysis of more than one mixture trace. This modelling of peak height information provides for a very efficient mixture analysis.

Recently Mortera *et al.* (2016) applied this model to analyse a complex disputed paternity case, where the DNA of the putative father was extracted from his corpse, which had been inhumed for over 20 years. This DNA was contaminated and appeared to be a mixture of at least two individuals. This case was further analysed in Green and Mortera (2016), which presents general methods for inference about relationships between contributors to a DNA mixture and other individuals of known genotype. The model for relationship inference builds on the approach in Cowell *et al.* (2015), but makes more explicit use of the Bayesian networks in the modelling.

1.7 Analysis of sensitivity to assumptions on founder genes

Many forensic genetics problems, as we have shown, can be handled using structured systems of variables, for which Bayesian networks offer an appealing practical modelling framework, and allow inferences to be computed by probability propagation methods. However, when standard assumptions are violated—for example when allele frequencies are unknown, there is identity by descent or the population is heterogeneous – dependence is generated among founding genes, that makes exact calculation of conditional probabilities by propagation methods less straightforward. The standard assumptions that the allele frequencies are fixed and known, that the individuals actors in the model are independent and that the allele frequency database is homogeneous can all be questioned (Green and Mortera 2009). We now illustrate a couple of these issues.

1.7.1 Uncertainty in allele frequencies

In reality, the allele frequencies assumed when conducting probabilistic forensic inference are not known probabilities, but estimates based on empirical frequencies in a database.

For the criminal case of § 1.5.1, the joint distribution of the founding genes is

$$\prod_m \{p(\texttt{spg}_m)p(\texttt{smg}_m)p(\texttt{opg}_m)p(\texttt{omg}_m)\}\,, \tag{1.3}$$

and all questions about sensitivity can be expressed through modifications to (1.3). Some generate dependence between founding genes.

Following Green and Mortera (2009), assuming the idealisation of a Dirichlet prior and multinomial sampling, the posterior distribution of a set of probabilities is Dirichlet$(M\rho(1), M\rho(2), \ldots, M\rho(k))$, where M is the (posterior) sample size and the ρ's are essentially the database allele frequencies (posterior means). The founding genes (spg, smg, opg, omg) are drawn from this distribution, (conditionally) independently and identically across alleles. This corresponds to the standard set-up for a Dirichlet process model which, by marginalising over the Dirichlet distribution, can be represented in a BN using a Pólya urn scheme. This is represented by the network module shown in Figure 1.16: for further details see Green and Mortera (2009). For efficiency of the probability propagation, in order to create smaller clique tables this network is set up so that all choices are binary, following the "divorcing" procedure (Jensen 1996), whereby auxiliary nodes are introduced in order to reduce the number of incoming edges of a selected node. This module can then be incorporated as a building block in a higher level network that computes inference, for example, about a criminal identification case, a simple or complex paternity testing or a DNA mixture problem. Thus Figure 1.17 shows a network for criminal identification that integrates the network of Figure 1.8a with that of Figure 1.16. Similarly the module in Figure 1.16, representing uncertain allele frequencies, can be integrated into the networks described in § 1.5.2, § 1.5.3, § 1.6. In this way, we can introduce uncertain allele frequencies for the reference population into forensic identification problems.

1.7.2 Heterogeneous reference population

The assumption that the DNA reference population is homogeneous is questionable. The population is typically a mixture of subgroups.

Population heterogeneity raises two kinds of issues in the modelling. First, since unobserved actors are assumed to have genes drawn from a population, results can depend on

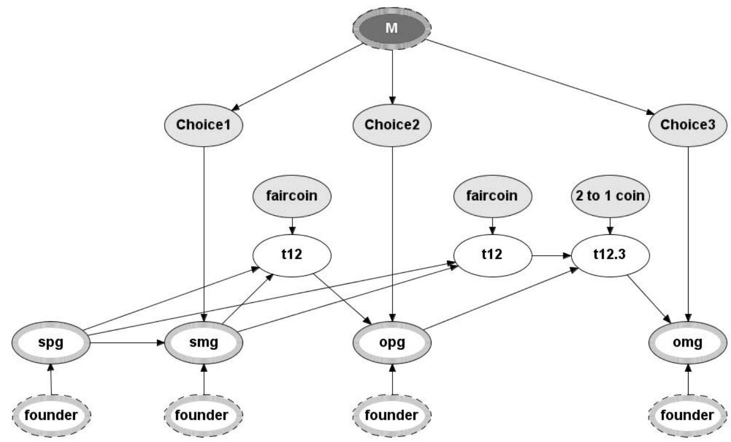


FIGURE 1.16: Sub-network UGF in Figure 1.17 for the Pólya urn scheme (adapted with permission from Green and Mortera (2009).)

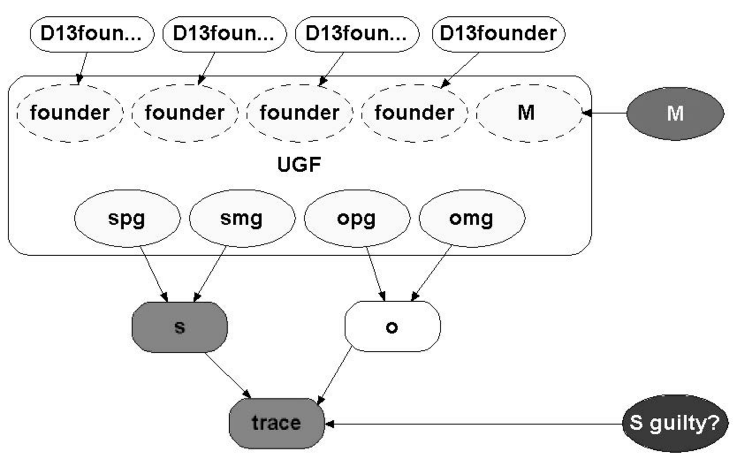


FIGURE 1.17: Network for criminal case with uncertain allele frequencies represented through the Pólya urn scheme

which population (and corresponding allele frequency database) is used. Secondly, when there is uncertainty about which population is relevant, this can induce dependence between actors, observed or not. Additionally, when uncertainty about subpopulation relates to untyped actors, dependence between markers is induced.

The upper level network for sensitivity of inferences to population structure for criminal identification, based on a synthetic population that is a mixture of Afro-Caribbean, Hispanic and Caucasian subpopulations is shown in Figure 1.18.

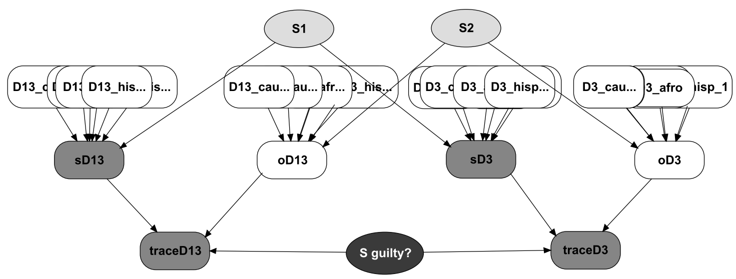


FIGURE 1.18: Network for 2 markers in a criminal identification allowing for subpopulation effect.

Such problems are easily set up as Bayesian networks with the sub-network structure shown in Figure 1.19. The variable `S` identifies the subpopulation, which may be dependent or independent between actors depending on the scenario of interest. Crucially, for each actor, `S` is the same for both genes for all markers, so that mixing across subpopulations is not the same as averaging the allele frequencies and assuming an undivided subpopulation. Note that conditional on subpopulation `S`, every gene at every marker is drawn independently from the appropriate subpopulation gene pool.

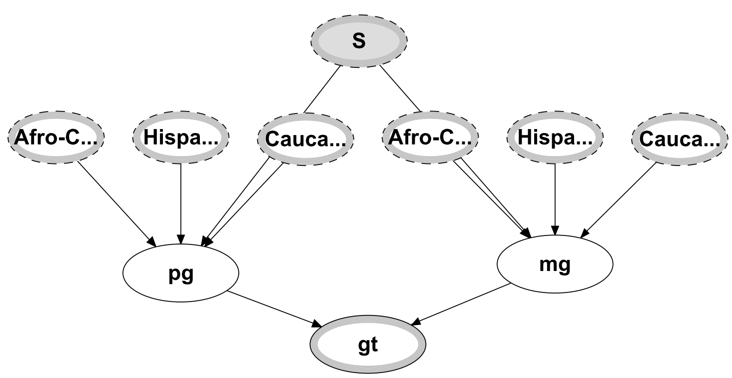


FIGURE 1.19: Network module for a genotype accounting for subpopulation effect (adapted with permission from Green and Mortera (2009).)

1.8 Conclusions

We hope we have stimulated the reader's interest in the application of Bayesian networks for modelling problems in forensic science.

JM: Maybe state that BNs and OOBNs can be usefully used in many branches of forensic analysis beyond those illustrated here. **:end**

We have also aimed to show the usefulness of Bayesian networks for representing and solving a wide variety of complex forensic problems. Both genetic and non-genetic information can be represented in the same network. A particularly valuable feature is the modular structure of Bayesian networks, which allows a complex problem to be broken down into simpler structures that can then be pieced back together in many ways, so allowing us to address a wide range of forensic queries. In particular, using object-oriented Bayesian networks we can construct a flexible computational toolkit, and use it to analyse complex cases of DNA profile evidence, accounting appropriately for such features as missing individuals, mutation, silent alleles and mixed DNA traces, accounting for uncertainty in allele frequencies, heterogeneous populations and also inference about relatedness in DNA mixtures (Green and Mortera 2016).

Appendix. Genetic background

We will introduce some basic facts about DNA profiles; for a more detailed explanation see Butler (2005).

A *gene* is a particular sequence of the four *bases*, represented by the letters A, C, G and T. A specific position on a chromosome is called a *locus* (hence there are two genes at any locus of a chromosome pair). A *DNA profile* consists of measurements on the genotype at a number of *forensic markers*, which are specially selected *loci* on different chromosomes.

Current technology uses around 17–23 *short tandem repeat* (STR) markers. At each marker, each gene has a finite number (up to around 20) of possible values, or *alleles*, generally positive integers. For example, an allele value of 5 indicates that a certain word (e.g. $CAGGTG$) in the four letter alphabet is repeated exactly 5 times in the DNA sequence at that locus. In statistical terms, a gene is represented by a random variable, whose realised state is an *allele*.

In a particular forensic context, we will refer to the various human individuals involved in the case as 'actors'. Each *genotype* consists of an unordered pair of genes, one inherited from the father and one from the mother (though one cannot distinguish which is which). When both alleles are identical the actor is *homozygous* at that marker, and only a single allele value is observed; otherwise the actor is *heterozygous*. An actor's *DNA profile* comprises a collection of genotypes, one for each marker.

Assuming *Mendelian segregation*, at each marker a parent passes a copy of just one of his two genes, randomly chosen, to his or her child, independently of the other parent and independently for each child.

In standard forensic identification problems it is customary to assume the *HardyWeinberg equilibrium*, and that loci are *unlinked*, which corresponds to assuming independence within and across markers. Databases have been gathered from which allele frequency distributions, for various populations, can be estimated for each forensic marker.

Acknowledgements

The authors would like to thank the Isaac Newton Institute for Mathematical Sciences for its hospitality and support during the programme *Probability and Statistics in Forensic Science* which was supported by EPSRC Grant Number EP/K032208/1. We also thank Normal Fenton and Nadine Smit for useful comments on a previous version.

Bibliography

Butler, J. M. (2005). *Forensic DNA typing*. Elsevier, USA.

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., Graversen, T., Lauritzen, S. L., and Mortera, J. (2015). Analysis of DNA mixtures with artefacts (with discussion). *Journal of the Royal Statistical Society: Series C*, **64**, 1–48.

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. (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**, (1), 28–34.

Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2011). Probabilistic expert systems for handling artefacts in complex DNA mixtures. *Forensic Science International: Genetics*, **5**, 202–9.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B (with Discussion)*, **41**, 1–31.

Dawid, A. P. and Evett, I. W. (1997). Using a graphical method to assist the evaluation of complicated patterns of evidence. *Journal of Forensic Sciences*, **42**, 226–31.

Dawid, A. P., Hepler, A. B., and Schum, D. A. (2011). Inference networks: Bayes and Wigmore. In *Evidence, Inference and Enquiry*, Proceedings of the British Academy, Vol. 171, (ed. A. P. Dawid, W. L. Twining, and D. Vasilaki), pp. 119–50. Oxford University Press.

Dawid, A. P., Mortera, J., Pascali, V. L., and van Boxel, D. W. (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics*, **29**, 577–95.

Dawid, A. P., Mortera, J., and Vicard, P. (2007). Object-oriented Bayesian networks for complex forensic DNA profiling problems. *Forensic Science International*, **169**, 195–205.

Evett, I. W., Gill, P. D., and Lambert, J. A. (1998). Taking account of peak areas when interpreting mixed DNA profiles. *Journal of Forensic Sciences*, **43**, 62–9.

Fenton, N., Neil, M., and Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive Science*, **37**, 61–102.

Green, P. J. and Mortera, J. (2009). Sensitivity of inferences in forensic genetics to assumptions about founder genes. *Annals of Applied Statistics*, **3**, 731–63.

Green, P. J. and Mortera, J. (2016). Paternity testing and other inference about relationships from dna mixtures. On ArXiv: `http://arxiv.org/abs/1609.09638`.

Hepler, A. B., Dawid, A. P., and Leucari, V. (2007). Object-oriented graphical representations of complex patterns of evidence. *Law, Probability and Risk*, **6**, 275–93.

Jensen, F. V. (1996). *An Introduction to Bayesian Networks.* UCL Press and Springer Verlag, London.

Laskey, K. B. and Mahoney, S. M. (1997). Network fragments: Representing knowledge for constructing probabilistic models. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, UAI97, pp. 334–41. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, **20**, 491–505.

Lauritzen, S. L. and Mortera, J. (2002). Bounding the number of contributors to mixed DNA stains. *Forensic Science International*, **130**, 125–6.

Mortera, J., Dawid, A. P., and Lauritzen, S. L. (2003). Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology*, **63**, 191–205.

Mortera, J., Vecchiotti, C., Zoppis, S., and Merigioli, S. (2016). Paternity testing that involves a DNA mixture. *Forensic Science International: Genetics*, **23**, 50–4.

Neil, M., Fenton, N., and Nielson, L. (2000). Building large-scale bayesian networks. *Knowl. Eng. Rev.*, **15**, (3), 257–84.

Schum, D. A. and Morris, J. R. (2007). Assessing the competence and credibility of human sources of intelligence evidence: Contributions from law and probability. *Law, Probability and Risk*, **6**, 247–74.