OOBN FOR FORENSIC IDENTIFICATION THROUGH SEARCHING A DNA PROFILES' DATABASE

David Cavallini
Department of Statistics "G. Parenti"
University of Florence (Italy)
cavallin@ds.unifi.it

Fabio Corradi
Department of Statistics "G. Parenti"
University of Florence (Italy)
corradi@ds.unifi.it

Abstract

In this paper we evaluate forensic identification hypotheses conditionally to the characteristics observed both on a crime sample and on individuals contained in a database. First we solve the problem via a computational efficient Bayesian Network obtained by transforming some recognized conditional specific independencies into conditional independencies. Then we propose an Object Oriented Bayesian Network representation, first considering a generic characteristic, then inheritable DNA traits. In this respect we show how to use the Object Oriented Bayesian Network to evaluate hypotheses concerning the possibility that some unobserved individuals, genetically related to the individuals profiled in the database, are the donors of the crime sample.

1 INTRODUCTION

Bayesian Networks (BN) are a powerful and compact representation of complex statistical models that exploit some recognized conditional independencies among random variables. A BN is defined as a pair of objects: a Directed Acyclic Graph (DAG) whose nodes represent discrete random variables, and a set of Conditional Probability Tables (CPT) which defines the conditional distributions of each vertex given the parents.

One of the reasons to represent a statistical model as a BN is the possibility to use well-established and effective algorithms to solve the inferential issue, i.e. to compute the distribution of some variables of interest conditionally to the evidence by using one of the available propagation algorithms (e.g. Jensen, 2001).

A limit in the representation of a BN arises when the number of random variables in the model increases due to some features of the problem.

Typically, this happens for time series models where a certain structure, a time-slice, is replicated over time, so that links between random variables in different time slices are established. This also occurs when we are interested in the relations between sets of random variables and when some specified relations between the sets must be taken into account. In the former case the model increases its dimensions over time, in the latter its growth depends on the number of sets involved.

In this respect, a new approach, stemmed from the *Object Oriented* language, has been introduced in the last few years. This modelling tool, called *Object Oriented Bayesian Network*, provides a useful technique capable of building a BN by merging pieces of simple BNs. Each item is an instantiation of a well-defined class which can be modified in order to accomplish the maintenance requirements. An update in the structure or in the CPTs of a class is automatically extended to all instantiations of that class. The subject is developed in Koller and Pfeffer (1997) and Bangso and Wuillemin (2000).

Here, we specifically deal with the forensic identification problem arising when a crime sample has been found but there is no clue about its origin. Searching a database (DB) of previously collected items is a common practice and the scope of this analysis is to assess the probability for each member of the database to be the origin of the trace. The problem has found considerable attention in the literature, but only not inheritable characteristics were considered, see e.g. Stockmarr (1999), Donnelly and Friedman (1999), Dawid (2001) and Meester and Sjerps (2003).

The aim of this paper is to show how this dimension-dependent problem, once opportunely formulated as a BN, can be effectively tackled. First, we provide a theoretical contribution transforming some recognized conditional specific independencies (Geiger and Heckerman, 1996) into conditional independencies, Section (2). Then, since the resulting BN shows many different repetitive structures, we propose an OOBN solution. The use of OOBN to model genetic data for identification was previously experienced by Dawid (2003) with special attention to the possibility of mutations.

The DB search problem is first developed for a not inheritable characteristic, but our real aim is to consider more complex genetic traits in order to extend the search to the relatives of the individuals profiled in the database, providing hints also when no match between the crime sample and one (or more) of the database members is found, Section (3.2). In Section (4) we provide the results of a simulation study based on a real database, emphasizing some computational issues. Finally we draw some conclusions.

2 EQUIVALENT BN FOR THE DB SEARCH PROBLEM

Let X the discrete population characteristic (or attribute) considered for the forensic identification problem. With $\mathcal{X}$ we indicate the set of the m states of X. The parameter θ_x, with $x \in \mathcal{X}$, is the probability that X is in state x, that is $P(X = x) = \theta_x$ and $\sum_{x \in \mathcal{X}} \theta_x = 1$. Uncertainty about these probabilities, derived from an inference process, could be introduced but this will not be considered here.

Let N the size of the reference population and n the number of the individuals in the DB. For each of them we define a random variable X_j with $j \in \mathcal{J} = \{1, 2, \ldots, n\}$. Also, we define X_c, the characteristic related to the crime scene, and the hypothesis variable H which has $n+1$ states. The first n of them represent the originator status of each individual, i.e. $H = j$, with $j \in \mathcal{J}$, means that the origin of the trace is the j-th individual in the DB, while the last, $H = \mathtt{r}$, is referred to the hypothesis that the trace's donor is outside the DB.

To specify the DB search model we adopt some common and reasonable assumptions:

i. the individual characteristics in the DB are jointly independent;

ii. the individual characteristics are jointly independent of the hypothesis variable, i.e. $\mathbf{X} \perp\!\!\!\perp H$ where $\mathbf{X} = \{X_j : j \in \mathcal{J}\}$;

iii. if the individual j is the originator of the trace the crime sample is observed without error $X_j \equiv X_c \mid H = j$;

iv. for $H = \mathtt{r}$ the individual attributes are jointly independent of the characteristic involved in the crime scene, i.e., $\mathbf{X} \perp\!\!\!\perp X_c \mid H = \mathtt{r}$ and $P(X_c = x \mid H = \mathtt{r}) = \theta_x$ with $x \in \mathcal{X}$;

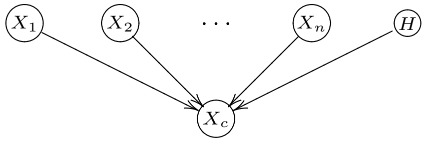

Figure 1: A DAG for the DB search problem.

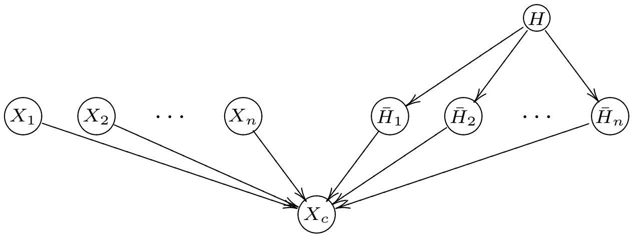

Figure 2: The augmented DAG obtained from Figure (1).

v. no other clue is available in advance, so the prior probability on H is $P(H = j) = 1/N$ and $P(H = \mathtt{r}) = 1 - n/N$.

The graphical structure, depicted in Figure (1), derives only from the assumptions (*i*) and (*ii*) while the CPTs are specified according to the assumptions (*iii*)-(*v*). Note that (*iii*) and (*iv*) imply a whole set of $n+1$ independence statements: for each value of H a different assertion of independence holds. This form of independence is known as *Conditional Specific Independence* (CSI) (Geinger and Heckerman, 1996), which differs from the usual definition of conditional independence since, in the latter, the independence assertions between variables do not vary according to the values of the conditioning sets.

The proposed network does not feature any conditional independence, so, for some evidence, the probability updating does not take advantage of the graphical representation. Moreover, the size of the CPT of X_c increases exponentially according to the number of individuals in the DB, so that the propagation becomes rapidly unfeasible. Our scope is to provide a more efficient solution by introducing a set of instrumental nodes in order to allow local computations. The result is attained in three steps.

Step 1. First, a set of binary random variables $\bar{\mathbf{H}} = \left\{\bar{H}_j : j \in \mathcal{J}\right\}$ is added and a new network is defined on the augmented domain, as in Figure (2).

The marginal distribution of the variables X_j and H does not change with respect to the original network and the remaining CPTs are defined as follows:

$$\hat{P}(\bar{H}_j = 1 \mid H = i) = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$\hat{P}(X_c \mid \mathbf{X}, \bar{\mathbf{H}} = \bar{\mathbf{h}}) = \begin{cases} P(X_c \mid X_j, H = j) & \text{if } \bar{\mathbf{h}} = \mathbf{1}_j \\ P(X_c \mid H = \mathtt{r}) & \text{if } \bar{\mathbf{h}} = \mathbf{0} \end{cases} \tag{2}$$

where $\mathbf{0}$ and $\mathbf{1}_j$ are vectors of size n. Each element of $\mathbf{0}$ is 0 while the i-th element of $\mathbf{1}_j$ is 0 $\forall i \neq j$ and 1 for $i = j$.

The CPTs for each node $\bar{H}_j$, specified as in (1), are the probabilistic translation of the deterministic logical `if-then` relation, i.e., $\forall j$ `if` $H = j$ `then` $\bar{H}_j = 1$ and $\forall i \neq j$, $\bar{H}_i = 0$. Thus, each variable $\bar{H}_j$ represents the originator status for the j-th individual and the deterministic relation is a consequence of the assumption that the characteristic observed on the crime scene was left by only one individual belonging to the reference population.

It is easy to prove that:

$$\sum_{\bar{\mathbf{H}}} \hat{P}(X_c, \mathbf{X}, \bar{\mathbf{H}}, H) = \sum_{j=1}^{n} \hat{P}(X_c, \mathbf{X}, \bar{\mathbf{H}} = \mathbf{1}_j, H) + \hat{P}(X_c, \mathbf{X}, \bar{\mathbf{H}} = \mathbf{0}, H) = P(X_c, \mathbf{X}, H). \tag{3}$$

Since the hypotheses are mutually exclusive, all configurations of $\bar{\mathbf{H}}$ not equal to the $\mathbf{1}_j$s and $\mathbf{0}$ have zero probability to realize. For this reason, in the marginalization (3), we consider only the relevant configurations of $\bar{\mathbf{H}}$.

The main consequence of the above mentioned result concerns the updating of the query variable H. In fact, for any evidence on $\mathbf{X}$ and X_c, the posterior probability of the hypotheses variable can be calculated indifferently by using the BNs of Figure (1) or Figure (2).

Step 2. Here a *divorcing* technique (Jensen, 2001) is applied. The idea is to introduce a set of mediating variables between parents and children in a large converging connection to lead some parents to divorce. The main advantage of this method is the reduction of the computational efforts because the original clique, $\{\mathbf{X}, X_c, \mathbf{H}\}$, is broken into a tree of smaller cliques.

A reasonable way to divorce the parents of node X_c in Figure (2)'s network is to add n mediating variables $\mathbf{Z} = \{Z_j : j \in \mathcal{J}\}$, which take values in $\mathcal{X}$, so that each pair of variables X_j and H_j are married. Figure (3) illustrates the DAG after divorcing. There, the node $X_c^\star$ represents the characteristic related to the crime scene which has been redefined for convenience. In particular $X_c^\star$ takes values in $\mathcal{X}^\star = \mathcal{X} \cup \{\text{NA}\}$ where the state labelled NA is an instrumental event to make the conditional distribution of $X_c^\star$ well defined also for $\mathbf{Z}$ values different from those allowed in this context. The $\mathbf{Z}$ can be considered as *private* copies of the crime sample, reproducing its value for each of the members of the DB.

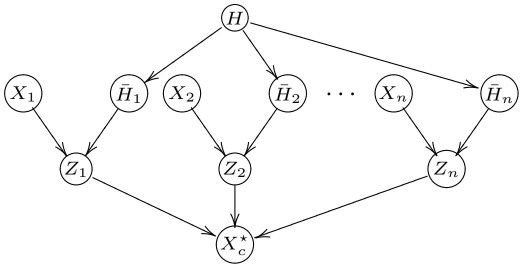

Figure 3: The augmented DAG of Figure (2) after the divorce.

The CPTs specification of the nodes $\mathbf{X}$, $\bar{\mathbf{H}}$ and H remains unchanged with respect to the BN of Figure (2). Imposing the CSI conditions

$$\forall j, Z_j \perp\!\!\!\perp X_j \mid \bar{H}_j = 0, \tag{4}$$

the rest of CPTs are specified as follows

$$\tilde{P}(Z_j = x \mid \bar{H}_j = 0) = \theta_x \tag{5}$$

$$\tilde{P}(Z_j = x \mid X_j = \hat{x}, \bar{H}_j = 1) = \begin{cases} 1 & \text{if } x = \hat{x} \\ 0 & \text{if } x \neq \hat{x} \end{cases} \tag{6}$$

$$\tilde{P}(X_c^\star = \bar{x} \mid \mathbf{Z} = \mathbf{z}) = \begin{cases} 1 & \text{if } \bar{x} = \text{NA or } \forall j, \bar{x} = z_j \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $\bar{x} \in \mathcal{X}^\star$ and $x, \hat{x}, z_j \in \mathcal{X}$.

The following proposition provides the probabilistic relation between the networks in Figure (2) and Figure (3).

PROPOSITION 2.1 *For each $x \in \mathcal{X}$ and for a given quantity $C(x)$, depending on x, the following relation holds:*

$$\hat{P}(X_c = x, \mathbf{X}, \bar{\mathbf{H}}, H) = C(x) \cdot \sum_{\mathbf{Z}} \tilde{P}(X_c^\star = x, \mathbf{X}, \bar{\mathbf{H}}, H, \mathbf{Z}) \tag{8}$$

Proof in the Appendix.

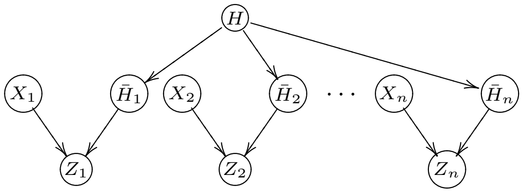

Figure 4: The network obtained by dropping the $X_c^\star$ node and the related incidental arcs from the DAG in Figure (3)

Finally, combining (3) with (8), we obtain the main result:

$$P(X_c = x, \mathbf{X}, H) = C(x) \cdot \sum_{\mathbf{Z}, \bar{\mathbf{H}}} \tilde{P}(X_c^\star = x, \mathbf{X}, \bar{\mathbf{H}}, H, \mathbf{Z}) \quad (9)$$

The above equation establishes that for calculating the posterior probability of the hypotheses variable H we can use the network of Figure (4) instead of that in Figure (3).

Step 3. As explained in the proof of **PROPOSITION** 2.1, during the propagation each valid evidence on $X_c^\star$ is transferred to all mediating variables. So, operationally, we build a new DAG merely by dropping the node $X_c^\star$ as well as its incidental arcs, Figure (4). Moreover, we use the characteristic observed on the crime scene for evidencing each vertex Z_j.

3 OOBN FOR THE DB SEARCH

The graph depicted in Figure (4) is conspicuous for its repetitive structure. For each individual profile in the DB the same BN is built and all the networks are mixed by the hypotheses variable H which is the only parent of every $\bar{H}_j$. Therefore, a set of conditional independence assertions appears, i.e., given H, each triple (Z_j, $\bar{H}_j$, X_j) is independent of the rest of the variables so that, for calculating the posterior distributions of H, local computations are allowed.

3.1 NOT INHERITABLE TRAITS

A more compact representation can be achieved by transforming the proposed network into the OOBN framework. As in Bangso and Wuillemin (2000), we define a class, $\mathbb{F}$, containing a simple BN, $\bar{H} \rightarrow Z \leftarrow X$, where the node $\bar{H}$ is an input node while X and Z are interior nodes. For each instantiation of the class $\mathbb{F}(j)$, with $j \in \mathcal{J}$, we build a binary random variable $\bar{H}_j^r$ which is *referenced* node of $\mathbb{F}(j).\bar{H}$. They are connected through a *reference* link ($\Rightarrow$), that is $\bar{H}_j^r \Rightarrow \mathbb{F}(j).\bar{H}$. Moreover, a set of arcs from the general hypotheses variable H pointing towards each referenced node is drawn. Finally, the CPTs related to the variables $\bar{H}_j^r$ are specified as in (1).

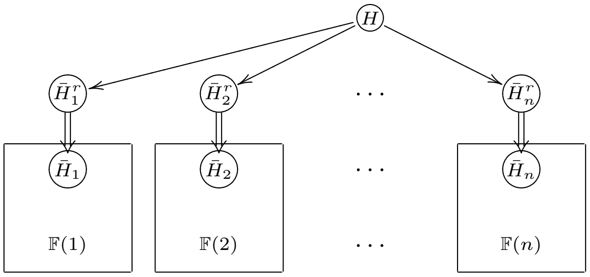

Figure 5: The OOBN representation for the DB search problem derived from Figure (4).

Figure (5) illustrates the OOBN representation for the DB search problem as the basic model to solve the forensic identification issue.

3.2 INHERITABLE DNA TRAITS

A DNA profile involves measurements on several well-specified locations of the DNA, called *loci*. For each locus we observe a genotype i.e. two alleles, one inherited from the father and the other from the mother, even if their origin is not distinguishable. For a generic locus we define two random variables A_0 and A_1 whose states, $a_1, a_2, \ldots, a_m$, are the inherited alleles. In addition, we consider a further random variable X whose states represent the genotypes, i.e., an ordered pair of alleles (a_t, a_u) with $t \leq u$. In this paper we assume Hardy-Weinberg (H-W) conditions and linkage equilibrium. H-W implies that parents are not related so that the inherited alleles in a genotype are independent. Linkage equilibrium refers to the independence among loci in the same individual. This is justified since the loci considered for identification are chosen far enough in the genome to make plausible that they are generated by different meiosis processes.

The genetic inheritance allows us to consider, as the possible donors of the crime sample, also individuals never typed but related to the DB members. In this way the no-match case, the most common in practice, but unfortunately the less useful, could originate *compatible* unobserved individuals, i.e. those having a positive probability for the characteristic observed on the crime sample, conditional to all the available evidence. For instance, a DB member not matching the crime sample but sharing at least one allele for each considered locus has a compatible child.

Here, we consider a pedigree, $\mathcal{F}$, constituted by a generic individual (i), their parents (0 and 1), their child (c), their partner (p) and their brother (b). Note that Labels 0 and 1 refer to a generic parent and not specifically to the mother or father because this information is not available. Since each pedigree is built around a member of the DB we call it a *first-degree-relative* pedigree. This choice is essentially due to the fact that, in the expectation of a significant hint about the trace's donor, we cannot explore too far from each individual in the DB. Refinements of the search could be achieved if familiar connections between the DB members were known. This kind of information is not usually recorded in a DB but, if available, could be exploited to relate two or more familiar classes with suitable links.

In this new perspective, the variables H and H^r_j shown in Figure (5) have a new meaning.

The j-th state of H, with $j \in \mathbb{J}$, refers to the hypothesis that the donor of the trace belongs to the family of the j-th individual of the DB, while $H = \mathtt{r}$ concerns the possibility that the trace was left by someone not included in the considered families. Every variable $\bar{H}^r_j$ takes values in $\bar{\mathcal{F}} = \mathcal{F} \cup \mathtt{r}$. The state $\mathtt{r}$ concerns the hypothesis that the trace is left by none of the considered family's members, while the statement $\bar{H}^r_j = q$, with $q \in \mathcal{F}$ means that the donor of the trace is the q-th member of the jth family.

Since, by definition, we have no clue about the donor of the trace, all the considered individuals are assumed to have the same prior probability to be the searched person. Within each family we assume that six persons are the possible suspects but, obviously, some of them might be ruled out if, e.g., they were in jail or dead. To refine the analysis we define an indicator variable $\mathrm{J}_{h,j} \in \{0,1\}$ for the relevance of the q-th person in the j-th family. Moreover, with $k_j = \sum_{q \in \mathcal{F}} \mathrm{J}_{q,j}$ we indicate the number of the relevant persons in the j-th family. The prior on H is $P(H = j) = k_j/N$, and

$$P(\bar{H}^r_j = q \mid H = i) = \begin{cases} \mathrm{J}_{q,j}/k_j & \text{if } j = i \text{ and } q \neq \mathtt{r} \\ 1 & \text{if } j \neq i \text{ and } q = \mathtt{r} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $i, j \in \mathbb{J}$ and $q \in \mathcal{F}$.

For inheritable DNA traits the class $\mathbb{F}$ includes the first-degree-relative pedigree and the set of hypotheses variables related to a generic family. Considering the *Allele Network* proposed by Lauritzen and Sheehan (2003), we provide an OOBN representation of $\mathbb{F}$ through defining two other classes: the *Individual* ($\mathbb{I}$) and the *Segregation* ($\mathbb{S}$) class.

The individual class $\mathbb{I}$ is represented in Figure (6). If no information about the individual's parents is available, the allele input nodes A^i_0 and A^i_1 depend on the reference population parameters, otherwise they are determined by the transmitted alleles. Another input node is the binary random variable $\tilde{H}$ representing the originator status of a generic individual. To provide the transmission of the individual genetic characteristics to the child, a copy of the alleles is expressed as output nodes (A^o_0 and A^o_1) and the other vertexes X and Z being interior nodes. The variable X denotes the observable genotype and its CPT is specified as follows

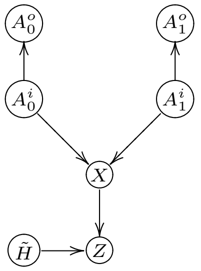

Figure 6: The individual class

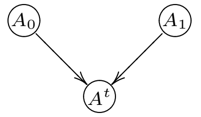

Figure 7: The segregation class

$$P(X = (a_r, a_u) \mid A^0_i = a_h, A^1_i = a_t) = \begin{cases} 1 & \text{if } (h = r \text{ and } t = u) \text{ or } (h = u \text{ and } t = r) \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

The segregation class, Figure (7), has two alleles as input nodes and provides the selection mechanism to generate the transmitted allele A^t via the following CPT, which reflects the first Mendelian law:

$$P(A^t = a_r \mid A_0 = a_t, A_1 = a_u) = \begin{cases} 1 & \text{if } r = t = u \\ 0.5 & \text{if } (r = t \text{ and } r \neq u) \text{ or } (r = u \text{ and } r \neq t) \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

On the whole the family class $\mathbb{F}$ is defined by a set of instantiations of $\mathbb{I}$, $\mathbb{I}(q)$, and $\mathbb{S}$, $\mathbb{S}(q,t)$, with $q, t \in \mathcal{F}$ and $q \neq t$. The index q is referred to the donor while t denotes the member who receives the allele in the segregation. The links among the instantiations of

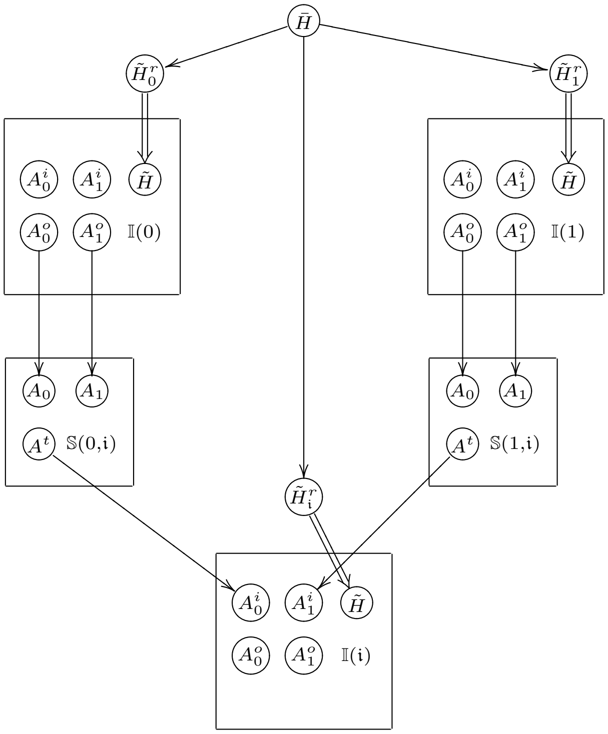

Figure 8: The family class $\mathbb{F}$ when $\mathcal{F} = \{0, 1, \mathrm{i}\}$.

the basic classes, $\mathbb{I}$ and $\mathbb{S}$, are drawn according to the biological relationships and each input node $\mathbb{I}(q).\tilde{H}$ has its own referenced vertex $\tilde{H}_q^r$. All of them are mixed by the input node $\bar{H}$ and the related CPTs are built as follows

$$P(\tilde{H}_q^r = 1 \mid \bar{H} = u) = \begin{cases} 1 & \text{if } q = u \\ 0 & \text{otherwise} \end{cases} \qquad (13)$$

with $u \in \bar{\mathcal{F}}$ and $q \in \mathcal{F}$. In Figure (8) we give a representation of $\mathbb{F}$, assuming, to simplify the picture, that $\mathcal{F} = \{0, 1, \mathrm{i}\}$.

The OOBN specified above deals with a single specific locus and it aims at the evaluation of the marginal posteriors for all the identification hypotheses.

In forensic practice, about 13-15 loci are usually typed for each individual and the support to the hypotheses is required conditionally to all the evidence.

Fortunately, this evaluation can be performed by using the results of the locus-specific nets, since linkage equilibrium still holds conditionally to the crime sample and the identification hypotheses. In fact, given an individual, the genotype distribution in a specific locus assumes the value of the genotype observed on the crime sample with probability one if identification is assumed; otherwise it follows the reference population distribution i.e. it never depends on the genotypes observed on other loci.

To give details, define: $\mathbb{L} = \{1, 2, \ldots, k\}$ the set of the loci; $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ the genotypes observed on the DB members and $\mathbf{x}_c = \{x_{c,1}, \ldots, x_{c,k}\}$ the crime samples observed on the considered loci.

If linkage equilibrium holds, the posterior of the identification hypothesis expressed in odds form, concerning, e.g., the q-th individual of the j-th family is:

$$\frac{P(H_j^r = q \mid \mathbf{x}, \mathbf{x_c})}{P(H_j^r \neq q \mid \mathbf{x}, \mathbf{x_c})} = \prod_{i=1}^{k} \frac{P(\mathbf{x}_i \mid x_{c,i}, H_j^r = q)}{P(\mathbf{x}_i \mid x_{c,i}, H_j^r \neq q)} \cdot \frac{P(H_j^r = q)}{P(H_j^r \neq q)}, \qquad (14)$$

since $\forall i,\ P(x_{c,i} \mid H_j^r = q) = P(x_{c,i} \mid H_j^r \neq q)$. The first term on the RHS of (14) is the likelihood ratio (LR) and can be evaluated making use of the results provided by each locus-specific net after propagation, in fact, $\forall i \in \mathbb{L}$

$$\frac{P(\mathbf{x}_i \mid x_{c,i}, H_j^r = q)}{P(\mathbf{x}_i \mid x_{c,i}, H_j^r \neq q)} = \frac{P(H_j^r = q)}{P(H_j^r \neq q)} \cdot \frac{P(H_j^r \neq q \mid \mathbf{x}_i, x_{c,i})}{P(H_j^r = q \mid \mathbf{x}_i, x_{c,i})}. \qquad (15)$$

Merits and difficulties to provide results as posteriors or LRs are discussed below.

1) The posterior probability of the hypothesis directly provides an answer to the uncertainty about the origin of the crime sample. Since a posterior requires the elicitation of a prior, this forces to deeply understand the meaning of each hypothesis, avoiding misunderstanding. This is a real problem as reveals the controversy between Stockmarr (1999), Dawid (2001) and Meester and Sjerps (2003): there the problem concerned the choice among hypotheses that sound logical. In this work both positions are represented: the Stockmarr's hypothesis is represented by the event $H \neq \mathbf{r}$ and considers the presence of the originator of the trace in the (augmented) DB; The Dawid's individual hypotheses are represented by the set of the H_j^rs. A possible drawback in the use of the posterior is that a large population size often implies very small (marginal) priors for each of the identification hypotheses so that small posteriors are likely to be obtained, wrongly suggesting a failure of the identification trial.

2) The LR is the measure usually provided to evaluate the evidence in a court; it does not imply any choice about priors and can be combined by the judge with others LRs obtained using different sources of evidence. An LR typically emphasizes a *discover*, even

Table 1: The rank distributions of the LR supporting the correct identification hypothesis.

Rank	Child	Brother
1°	54.99%	61.89%
2°	16.24%	10.71%
3°	7.53%	4.26%
4°	4.17%	2.36%
5°	2.90%	2.08%
6°	1.81%	1.27%
7°	1.63%	1.45%
≤ 8°	10.73%	15.98%

if the result might be of difficult interpretation, since the LR is not expressed in a normalized form.

4 APPLICATIONS

Now let us give account of a simulation study on a real DB containing 1102 observations on 10 loci. What is involved is how effective is the DB search in retrieving the origin of the simulated crime samples.

To produce the first simulation, we generated for each observed individual two crime samples obtained respectively sampling from the posterior marginal distribution of the child's and brother's genotypes. We call them the Child Crime Sample (CCS) and Brother Crime Sample (BCS).

Consider first the CCS. For each considered first-degree-relative pedigree we evaluate the hypothesis concerning the identification of the family originating the child. Obviously we expect that the LRs, or the posteriors, evaluated for the family from which the CCS was generated has one of the highest values. Similar computations are provided if the BCSs are used, and the results are in Table (1).

Concerning the identification of a child, in over 85% of the cases, the LR corresponding to the originating family ranks in the top five highest positions; the identification of a *brother* is slightly less successful, since in this case the same figure is just over 80%. In real cases, it seems safe to suggest that the the results' evaluation should include a comparison between the LRs or the posteriors for the families exhibiting the highest values associated to a careful investigative work.

As a comment, it must be noted that our simulation is disadvantaged with respect to real cases. For instance, when we sample a BCS we do not know the relatives' genotypes as the nature knows but our knowledge is restricted to the brother posterior distribution, typically over-dispersed. In real cases, brothers' genotypes are often very similar: for each locus, if one of the parents is homozygote the probability that brothers share one allele is equal to one and the probability they are identical is equal to 0.5.

Table 2: Parameter estimates of the CPU time proposed model

CPU	α	β	θ
Pentium IV	-10.93	1.82	0.01
AMD64	-11.75	1.84	0.01

A further simulation experiment has been achieved, making use of different DB sizes in the range $5000 - 50000$, and loci with a number of alleles varying in the range $5 - 20$. We estimate the dependence of the CPU times (t) required to perform the search with respect to the DB size (n) and the alleles' number (a) according to the model $\log(t) = \alpha + \beta \cdot \log(n) + \theta \cdot \log(a) + e$ where e is the stochastic error with zero mean. Results are in Table (2).

Clearly the estimation of the βs and the θs produced very similar results and the difference in technology is provided by α. Note that there is a very slight dependence on the number of the alleles, due to the adoption of an allele recoding strategy (Lauritzen and Sheehan, 2003). Instead the dependence of t on the DB size is less then quadratic, making the search feasible also when very large DB are involved.

5 CONCLUSIONS

The use of BN to provide an evaluation of the LR for forensic identification purposes is a new but already well-established approach, see Dawid (2003), Mortera and al. (2003) and Corradi et al. (2003).

Here, the BN technology is invoked when there is no clue about the origin of the trace, but a list of well identified individuals, not apparently related to the crime, is available in the DB. This result is all the more effective when an augmented DB is introduced, having assumed that all its members belong to the population of the crime sample's possible donors, even if some of them are not observed. In this new perspective the OOBN approach provides the most striking solution: the *familiar*, the *individual* and the *segregation* classes of hierarchy provide a concise representation of the repetitive part of the problem, saving efforts when *maintenance* operations are required. This could happen, for instance, when we want to introduce the possibility of a mutation in the alleles transmission: in this case a slight modification of the segregation class produces the result. At the same time the proposed solution leaves some room to operate on the single instance of the classes. This is compulsory for our problem since we are required not to consider as possible originator of the crime sample those individuals in the augmented DB who are not included in the donors' population since e.g. dead or in jail. In the OOBN environment this can be realized just by intervening on the hypotheses input nodes concerning each family and detailed for each considered members.

PROOF OF PROPOSITION 2.1

The joint marginal distribution of $\{\mathbf{X}, \bar{\mathbf{H}}, H\}$ is the same in the two BNs of Figure (2) and Figure (3) so (8) becomes

$$\hat{P}(X_c = x, | \mathbf{X}, \bar{\mathbf{H}}) =$$
$$C(x) \cdot \sum_{\mathbf{Z}} \tilde{P}(X_c^\star = x, | \mathbf{Z}) \cdot \prod_{j=1}^{n} \tilde{P}(Z_j \mid X_j, \bar{H}_j). \quad (16)$$

When the variable $X_c^\star$ receives an evidence $x \in \mathcal{X}$ it is easy to show that after the reduction (7) can be written as product of n potential ϕ_j, that is

$$\hat{P}(X_c^\star = x \mid \mathbf{Z}) = \prod_{j=1}^{n} \phi_j(Z_j) \quad (17)$$

where

$$\phi_j(Z_j = \hat{x}) = \begin{cases} 1 & \text{if } \hat{x} = x \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

with $\hat{x} \in \mathcal{X}$.

The equation (18), which defines a *finding* on Z_j, establishes that all mediating variables take value x with probability 1. So, combining equations (17) and (18) with (16) we obtain

$$\hat{P}(X_c = x, | \mathbf{X}, \bar{\mathbf{H}}) = C(x) \cdot \prod_{j=1}^{n} \tilde{P}(Z_j = x \mid X_j, \bar{H}_j). \quad (19)$$

If $\bar{\mathbf{H}} = \mathbf{1}_j$ then from (2) and (4) we have

$$P(X_c = x, | X_j, H = j) = C(x) \cdot$$
$$\prod_{i \neq j} \tilde{P}(Z_i = x \mid \bar{H}_i = 0) \cdot \tilde{P}(Z_j = x \mid X_j, \bar{H}_j = 1). \quad (20)$$

The third part of RHS of (20) involves $n-1$ terms. From (5), each of them is equal to θ_x so, considering (6) and assumption (*iii*) we obtain $C(x) = \theta_x^{1-n}$.

The same result is achieved for $\bar{\mathbf{H}} = \mathbf{0}$ as well. In fact, in that case, considering (2) and (4), the equation (19) becomes

$$P(X_c = x, | H = \mathtt{r}) = C(x) \cdot \prod_{j=1}^{n} \tilde{P}(Z_j = x \mid \bar{H}_j = 0). \quad (21)$$

Finally, from condition (*iv*) and equation (5) we obtain again $C(x) = \theta_x^{1-n}$.

REFERENCES

O. Bangso and P-H. Wuillemin (2000). Top-down Construction and Repetitive Structures Representation in Bayesian Networks. In *Proceedings of the Thirteenth International Florida Artificial Intelligence Society Conference*, 282-286. AAAI Press.

F. Corradi and G. Lago and F. M. Stefanini (2003). The Evaluation of DNA Evidence in Pedigrees Requiring Population Inference. *Journal of the Royal Statistical Society*, A166, 425-440.

A. P. Dawid (2001). Comment on Stockmarr's Likelihood Ratios for Evaluating DNA Evidence When the Suspect is Found Through a Database Search. In *Biometrics*, 57, 976-980.

A. P. Dawid (2003). An Object Oriented Bayesian Network for Estimating Mutation Rates. In *Proceeedings of the Ninth International Workshop on Artificial Intelligence and Statistics, January 3-6 2003, Key West, Florida*, edited by Christopher M. Bishop and Brendan J. Frey.

P. Donnelly and R.D. Friedman (1999). DNA Database Searches and the Legal Consumption of Science Evidence. *Michigan Law Review*, 974, 931-984.

D. Geiger and D. Heckerman (1996). Knowledge Representation and Inference in Similitary Networks and Bayesian Multinets. *Artificial Intelligence*, 82, 45-74.

F.V. Jensen (2001).Bayesian Network and Decision Graphs. *Springer-Verlag*, New York.

D. Koller and A. Pfeffer (1997). Object-Oriented Bayesian Network. *Proceedings of Thirteenth Conference on Uncertainty in Artificial Intelligence*, 302-313.

S. L. Lauritzen and N. A. Sheehan (2003). Graphical models for genetic analyses. *Statistical Science*, 18, 489-514.

R. Meester and M. Sjerps (2003). The Evidential Value in the DNA Database Search Controversy and the Two Stain Problem *Biometrics*, 59, 727-732

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

A. Stockmarr (1999). Likelihood Ratios for Evaluating DNA Evidence When the Suspect is Found Through a Database Search *Biometrics*, 55, 671-677