

## **Forensic identification of relatives of individuals included in a database of DNA profiles**

BY DAVID CAVALLINI AND FABIO CORRADI

*Department of Statistics, University of Florence, 50134 Florence, Italy*

cavallin@ds.unifi.it corradi@ds.unifi.it

### SUMMARY

In this paper we evaluate the characteristics observed both on a crime sample and on individuals included in a database to assess the probability of alternative hypotheses concerning identification. The problem is first addressed by considering a generic characteristic and we demonstrate the problem via a computationally efficient Bayesian network. Then we turn our attention to a heritable DNA trait to show how to evaluate the hypotheses that some individuals, genetically related to the members of the database, are the donors of the crime sample. Then the network is extended to cope with many loci. Applications of the method are provided as well as details of computational requirements.

*Some key words:* Bayesian network; Database of DNA profiles; Forensic identification; Search on a database.

### 1. INTRODUCTION

In this paper we consider the forensic identification problem arising when a crime sample is found but there is no clue about its origin and a search in a database is a real possibility.

In the case where only non-heritable characteristics are concerned, or if heritability is not exploited, Dawid (1994) and Dawid & Mortera (1996) found computationally simple solutions to cases arising in a database search involving one or more matches, imperfect evidence and different observation schemes. Simplicity was achieved mainly because all the individuals considered in the evaluation of hypotheses were observed, so that, for those not matching the crime sample, the posterior probability of identification was zero or expressed by formulae depending on laboratory error parameters. Even within this simplified setting the database search problem has led to a considerable debate in the literature, originated by the National Research Council reports (National Research Committee on DNA Forensic Science, 1992, 1996), and followed by Balding & Donnelly (1996) and Stockmarr (1999). A comprehensive and critical review is given in Donnelly & Friedman (1999) and recent contributions are by Dawid (2001) and Meester & Sjerps (2003).

Here, instead, we explicitly consider DNA traits, exploiting heritability to extend the search to some specified but unobserved relatives of the database members.

In principle, to obtain the posteriors for each identification hypothesis, a straightforward application of Bayes' theorem suffices. However, all the unobserved variables accounting for heritability and the individual identification hypotheses must be marginalised out. If

assertions of conditional independence are not exploited, the marginalisation procedure involves as many summations as the product of the dimensions of all unobserved variables, which exponentially increases with the database size. This circumstance makes the search very computationally demanding.

A much more efficient but still general solution can be achieved if the database problem is formulated in terms of a Bayesian network. A Bayesian network is a graphical representation of a statistical model defined on a set of random variables which consists of two objects: a directed acyclic graph,  $G = (V, E)$ , and a set of conditional probability tables. Each node in  $V$  corresponds to a variable in the domain; the absence of directed links between nodes in  $E$  encodes the conditional independencies between relevant sets of variables; and the conditional probability tables represent the conditional distributions of each node of the graph given its parents (Pearl, 1988, pp. 116–22).

Generally speaking, the computation of the conditional probability of unobserved nodes in a Bayesian network is realised efficiently since the required marginalisation is performed in subsets of  $V$  organised in a special graph, called a junction tree, exploiting conditional independencies (Cowell et al., 1999, Ch. 4).

To be more specific, in problems concerning the transmission of genetic evidence, Lauritzen & Sheehan (2003) proved that propagation algorithms designed for Bayesian networks are a generalisation of the well-established peeling algorithm (Cannings et al., 1978) proposed for coping with the problem of evaluating efficiently the probability of unobserved genetic traits. They also proved that Bayesian networks overwhelm in terms of efficiency every marginalisation procedure that does not take into account conditional independence relationships, and they detailed the computational efforts required by different Bayesian-network representations of a pedigree.

In a forensic setting, Dawid et al. (2002), Mortera et al. (2003) and Cowell (2003) proved the usefulness of the Bayesian network approach to cope with a well-defined unobservable identification hypothesis when the possibility of mutations, the presence of multiple contributors to the crime sample and other intricate conditions are relevant.

None of these contributions provides a solution to the database search problem when heritable traits are involved. In this case, we must consider simultaneously the transmission of the genetic information from the database members to their unobserved relatives and the identification hypotheses concerning each individual.

## 2. THE DATABASE SEARCH PROBLEM

Let  $X_j$ , with  $j \in \mathcal{J} = \{1, 2, \dots, n\}$ , be the characteristic or attribute related to the  $j$ th individual of the database. By  $\mathcal{X}$  we indicate the possible values of each  $X_j$  while the parameter  $\theta_x$  denotes the probability that a generic random variable  $X_j$  is in state  $x \in \mathcal{X}$ ; that is, for all  $j \in \mathcal{J}$ ,  $P(X_j = x) = \theta_x$ . Obviously,  $\sum_{x \in \mathcal{X}} \theta_x = 1$ .

The parameters  $\theta = \{\theta_x : x \in \mathcal{X}\}$  are the relative frequencies of the characteristic in the reference population, i.e. in the set of individuals from which the members of the database and the crime sample,  $X_c$ , come. The reference population is defined by auxiliary information strictly related to the case, such as special location, ethnicity and so on: the simplest situation occurs when the reference population is homogeneous and its size,  $N$ , is known, as happens in the so-called island problem (Dawid, 1994). In this paper, these latter circumstances are assumed, but uncertainty about  $N$  and  $\theta$  and the possibility of a structured population could be included in the model.

Moreover, we introduce a hypothesis random variable  $H$  with  $n + 1$  states. The first  $n$  states represent the originator status of each individual; that is,  $H = j$ , with  $j \in \mathcal{J}$ , means that the origin of the trace is the  $j$ th individual in the database. The last state,  $H = r$ , refers to the possibility that the donor of the trace is not in the database.

To specify the database search model we adopt some common and reasonable assumptions.

*Assumption 1.* The individuals' characteristics in the database are mutually independent given  $\theta$ .

*Assumption 2.* The hypothesis variable does not affect the individuals characteristics. To be more specific, for the individuals in the database,  $X \perp\!\!\!\perp H$ , where  $X = \{X_j : j \in \mathcal{J}\}$ .

*Assumption 3.* If individual  $j$  is the donor of the trace the crime sample is observed without error, that is  $X_j = X_c | H = j$ .

*Assumption 4.* For  $H = r$  the set of individual attributes is independent of the characteristic involved in the crime scene, that is  $H \perp\!\!\!\perp H_c | H = r$  and  $P(X_c = x | H = r) = \theta_x$  with  $x \in \mathcal{X}$ .

*Assumption 5.* No other clue is available in advance, so that the prior probability on  $H$  is not informative:  $P(H = j) = 1/N$  and  $P(H = r) = 1 - n/N$ .

The graphical structure, depicted in Fig. 1(a), derives from Assumptions 1 and 2, while the conditional probability tables are specified according to Assumptions 3–5 and the population parameters.

The network in Fig. 1(a) does not feature any conditional independence, so that, for some evidence, the probability updating and the marginalisation activity do not take advantage of the graphical representation.

Our main contribution is to propose an alternative but equivalent Bayesian network, in which assertions of conditional independence appear and local computations are allowed.

The result is attained in three steps.

*Step 1.* First, we introduce a set of binary random variables  $\tilde{H} = \{\tilde{H}_j : j \in \mathcal{J}\}$  representing the originator states of the individuals. The new directed acyclic graph is depicted in Fig. 1(b).

Since the hypotheses are mutually exclusive, the conditional probability tables attached to each node  $\tilde{H}_j$  are specified according to the following deterministic relationships:

- (a) for all  $j \in \mathcal{J}$ ,  $H = j$  if and only if  $\tilde{H}_j = 1$  and, for all  $i \neq j$ ,  $\tilde{H}_i = 0$ ;
- (b) for all  $j \in \mathcal{J}$ ,  $H = r$  if and only if  $\tilde{H}_j = 0$ .

Then the conditional probability table related to  $X_c$  is defined as follows:

$$\hat{P}(X_c = x | X, \tilde{H} = h) = \begin{cases} 1, & \text{if } h = 1_j \text{ and } X_j = x, \\ \theta_x, & \text{if } h = 0, \end{cases} \quad (1)$$

where here  $0$  is a vector of  $n$  zeros and  $1_j$  is an  $n$ -vector with  $j$ th element equal to 1 and the rest equal to zero. Since all configurations of  $\tilde{H}$  not equal to  $1_j$  or  $0$  occur with probability zero, how the rest of (1) is specified is not relevant.

Since  $\tilde{H}$  is fully determined by  $H$ , marginalising with respect to  $\tilde{H}$  the joint probability distribution induced by the directed acyclic graph of Fig. 1(b),  $\hat{P}(X_c, X, \tilde{H}, H)$ , exactly produces  $P(X_c, X, H)$ , that is the joint probability distribution over  $X_c$ ,  $X$  and  $H$  as specified in the original model.

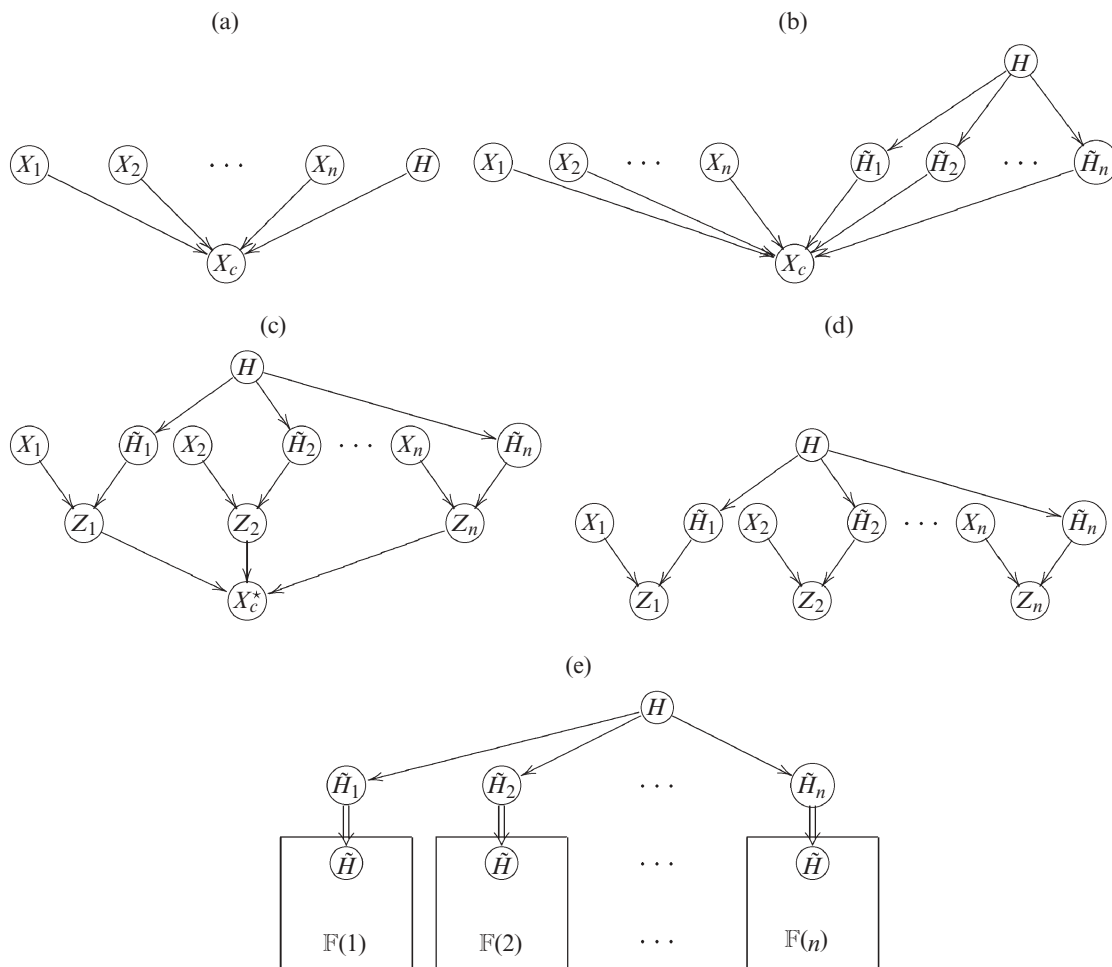


Fig. 1. (a) A directed acyclic graph for the database search problem. (b) The augmented directed acyclic graph. (c) The augmented directed acyclic graph after the divorce. (d) The network then obtained by dropping the  $X_c^*$  node and its incidental arcs. (e) The consequent object-oriented Bayesian network representation for the database search problem.

*Step 2.* Here, we apply a ‘divorcing’ technique (Jensen, 2001, pp. 61–2) by introducing a set of mediating variables to lead some parents to divorce.

A way of divorcing the parents of node  $X_c$  in the network of Fig. 1(b) is to add  $n$  mediating variables  $Z = \{Z_j : j \in \mathcal{J}\}$ , which take values in  $\mathcal{X}$ , so that each pair of variables  $X_j$  and  $\hat{H}_j$  are married. Figure 1(c) illustrates the directed acyclic graph after divorcing. The node  $X_c^*$  represents the characteristic related to the crime scene which has been redefined for convenience. In particular,  $X_c^*$  takes values in  $\mathcal{X}^* = \mathcal{X} \cup \{\text{NA}\}$  where the residual state labelled NA is a never-observed event added to define formally the conditional probability table of  $X_c^*$ , which is specified by imposing the following deterministic relationships.

*Relationships 1.* For all  $x \in \mathcal{X}$ ,  $X_c^* = x$  if and only if, for all  $j \in \mathcal{J}$ ,  $Z_j = x$ .

*Relationships 2.* Here  $X_c^* = \text{NA}$  if and only if  $Z = z$  exists so that at least two elements of  $z$ ,  $z_j$  and  $z_i$ , with  $j \neq i$ , differ, that is  $z_j \neq z_i$ .

To complete the probability distribution construction of the directed acyclic graph of Fig. 1(c), the conditional probability tables related to  $Z$  nodes are built according to the following rules.

*Rule 1.* For all  $j \in \mathcal{J}$ ,  $Z_j \perp\!\!\!\perp X_j | \tilde{H}_j = 0$  and, for all  $x \in \mathcal{X}$ ,  $\tilde{P}(Z_j = x | \tilde{H}_j = 0) = \theta_x$ .

*Rule 2.* For all  $j \in \mathcal{J}$ ,  $Z_j = X_j | \tilde{H}_j = 1$ .

The following proposition, proved in the Appendix, provides the probabilistic relationship between the networks in Figs 1(b) and (c).

**PROPOSITION 1.** *For each  $x \in \mathcal{X}$  the following relationship holds:*

$$\hat{P}(X_c = x, X, \tilde{H}, H) = C(x) \sum_Z \tilde{P}(X_c^* = x, X, \tilde{H}, H, Z), \quad (2)$$

where  $\tilde{P}(\cdot)$  is the probability distribution induced by the directed acyclic graph of Fig. 1(c) and  $C(x) = \theta_x^{1-n}$ .

Finally, from Step 1 and equation (2), we derive the main result: the posterior probability of  $H$  given the evidence on database and crime scene can be calculated by using the Bayesian network either in Fig. 1(a) or in Fig. 1(c).

*Step 3.* As explained in the proof of Proposition 1, during the propagation, any valid evidence about  $X_c^*$  is transferred to all mediating variables. Operationally, therefore, we build a new directed acyclic graph merely by dropping the node  $X_c^*$  and its incidental arcs. Moreover, we use the characteristic observed on the crime scene to provide evidence for each vertex  $Z_j$ .

The new graph, depicted in Fig. 1(d), is conspicuous for its repetitive structure. For each individual in the database the same Bayesian network is built and all the networks are mixed by the hypothesis variable  $H$  which is the only parent of every  $\tilde{H}_j$ . Therefore, a set of conditional independence assertions appears; that is, given  $H$ , each triple  $(Z_j, \tilde{H}_j, X_j)$  is independent of the rest of the variables so that, for calculating the posterior distributions of  $H$ , local computations are allowed.

A more compact representation can be achieved by transforming the proposed network in an object-oriented Bayesian network. Considering the approach proposed by Bangso & Willemin (2000), we define a class,  $\mathbb{F}$ , containing a simple Bayesian network,  $\tilde{H} \rightarrow Z \leftarrow X$ , where the node  $\tilde{H}$  is an input node while  $X$  and  $Z$  are interior nodes. For each realisation of the class  $\mathbb{F}$ ,  $\mathbb{F}(j)$  with  $j \in \mathcal{J}$ , the node  $\tilde{H}_j$  is the ‘referenced’ node of the vertex  $\tilde{H}$  defined within  $\mathbb{F}(j)$ . They are connected through a ‘reference’ link ( $\Rightarrow$ ) which codifies a deterministic identity relationship. Figure 1(e) illustrates the object-oriented Bayesian network representation for the database search problem.

### 3. HERITABLE NUCLEAR DNA TRAITS

A DNA profile concerns 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 recoverable. For a generic locus we define two random variables  $A_0$  and  $A_1$  whose states,  $a_1, a_2, \dots, a_m$ , are the heritable alleles, and a 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 independence among loci and the presence of Hardy–Weinberg equilibrium. The generalisation of the approach to reference populations far from equilibrium requires specific information and more complex models accounting for the presence of population substructures, as illustrated for instance in Pritchard et al. (2000).

The genetic inheritance allows us, while searching for the possible donor of the crime sample, to consider also individuals never typed but related to the database members. In this way the no-match case, in practice the most common and unfortunately the least useful outcome of the database search, could create ‘compatible’ unobserved individuals, which are those having a positive probability for the characteristic observed on the crime sample, conditional on all the available evidence. For instance, a database member sharing at least one allele with the crime sample at each considered locus but not matching the crime sample has a compatible child.

Here, for each individual,  $i$ , in the database we consider a search in the pedigree,  $\mathcal{F}$ , including their parents, labelled as 0 and 1, their child,  $c$ , the other parent,  $p$ , of  $c$  and their sibling,  $s$ . Note that the 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 database we call it a ‘first-degree-relative’ pedigree. The search could be easily extended if more than one member of a family was in the database.

In this new perspective, the variables  $H$  and  $\tilde{H}_j$ , shown in Fig. 1(e), have a new meaning. The  $j$ th state of  $H$ , with  $j \in \mathcal{J}$ , refers to the hypothesis that the donor of the trace belongs to the family of the  $j$ th individual of the database, while  $H = r$  corresponds to the possibility that the trace was left by someone not included in the considered families.

Furthermore, every variable  $\tilde{H}_j$  takes values in  $\tilde{\mathcal{F}} = \mathcal{F} \cup r$ . The state  $r$  corresponds to the hypothesis that the trace was not left by any of the considered family’s members, while the statement  $\tilde{H}_j = q$ , with  $q \in \mathcal{F}$ , means that the donor of the trace is exactly the  $q$ th member of the  $j$ th family.

Since, by Assumption 5, we have no clue about the donor’s identity, all the considered individuals are assumed to have the same prior probability of being the searched person. Obviously, in each family, some of the considered suspects could be ruled out, if for instance they were in jail or dead.

For heritable DNA traits the class  $\mathbb{F}$  includes the first-degree-relative pedigree and the set of hypothesis variables. Considering the allele network proposed by Lauritzen & Sheehan (2003), we provide an object-oriented Bayesian network representation of  $\mathbb{F}$ . We then need to define two other classes, namely the individual class,  $\mathbb{I}$ , and the segregation class,  $\mathbb{S}$ .

The individual class’s inner structure is represented in Fig. 2(a).

If no information about the individual’s parents is available, the probabilities of the allele input nodes  $A_0^i$  and  $A_1^i$  depend on the reference population parameters; otherwise they are determined by the transmitted alleles. Another input node is the binary random variable  $\hat{H}$  representing the originator status of a generic individual. To prove the transmission of the individual’s genetic characteristics to the siblings, a copy of the alleles is expressed as output nodes,  $A_0^o$  and  $A_1^o$ , the other vertices  $X$  and  $Z$  being interior nodes. The variable  $X$  denotes the observable genotype and its conditional probability table is specified as follows:

$$P\{X = (a_r, a_u) | A_0^i = a_q, A_1^i = a_t\} = \begin{cases} 1, & \text{if } (q = r, t = u) \text{ or } (q = u, t = r), \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

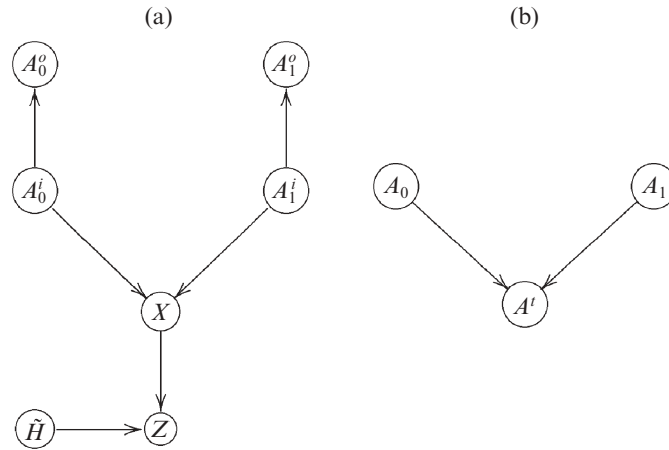


Fig. 2. (a) The individual class  $\mathbb{I}$  and (b) the segregation class  $\mathbb{S}$ .

while  $Z$  plays the instrumental role explained in § 2 and its conditional probability table is built according to Rules 1 and 2 in § 2.

The segregation class’s structure, Fig. 2(b), has two allele input nodes and provides the selection mechanism for generating the transmitted allele  $A^t$  via the following conditional probability table, which reflects the first Mendelian law:

$$P(A^t = a_r | A_0 = a_t, A_1 = a_u) = \begin{cases} 1, & \text{if } r = t = u, \\ 0.5, & \text{if } (r = t, r \neq u) \text{ or } (r = u, r \neq t), \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Overall, the family class  $\mathbb{F}$  is defined by a set of realisations  $\mathbb{I}(q)$  of  $\mathbb{I}$  and  $\mathbb{S}(q, t)$  of  $\mathbb{S}$ , with  $q, t \in \mathcal{F}$  and  $q \neq t$ . The index  $q$  refers to the allele donor while  $t$  denotes the individual that receives the alleles after the segregation. The links among the realisations of the basic classes,  $\mathbb{I}$  and  $\mathbb{S}$ , are drawn according to the biological relationships, and each input node  $\hat{H}$  defined within each  $\mathbb{I}(q)$  has its own referenced vertex,  $\hat{H}_q$ . All of them are mixed by the only input node  $\tilde{H}$ , and the related conditional probability tables are built as follows:

$$P(\hat{H}_q = 1 | \tilde{H} = u) = \begin{cases} 1, & \text{if } q = u, \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

with  $u \in \bar{\mathcal{F}}$  and  $q \in \mathcal{F}$ . Figure 3 gives a simple example of  $\mathbb{F}$  under the assumption that  $\mathcal{F} = \{0, 1, i\}$ .

The object-oriented Bayesian network specified above considers a single specific locus and it aims at the evaluation of the marginal posteriors for all the identification hypotheses.

In forensic practice, 13 to 15 loci are usually typed for each individual and the posteriors of the hypotheses of interest are required conditionally on all the available evidence.

To extend the analysis, let  $\mathbb{F}_l$  be the family class for the  $l$ th locus. For the  $j$ th family the relationship between the realisation of  $\mathbb{F}_l$  classes is expressed by reference links connecting  $\tilde{H}_j$  with every referenced node  $\tilde{H}$  defined within each  $\mathbb{F}_l(j)$ . As an example, Fig. 4 depicts the object-oriented Bayesian network for two loci.

The posterior probability of a hypothesis for an individual evaluates his/her probability to be the origin of the crime sample, but often, instead, the likelihood ratio is provided.

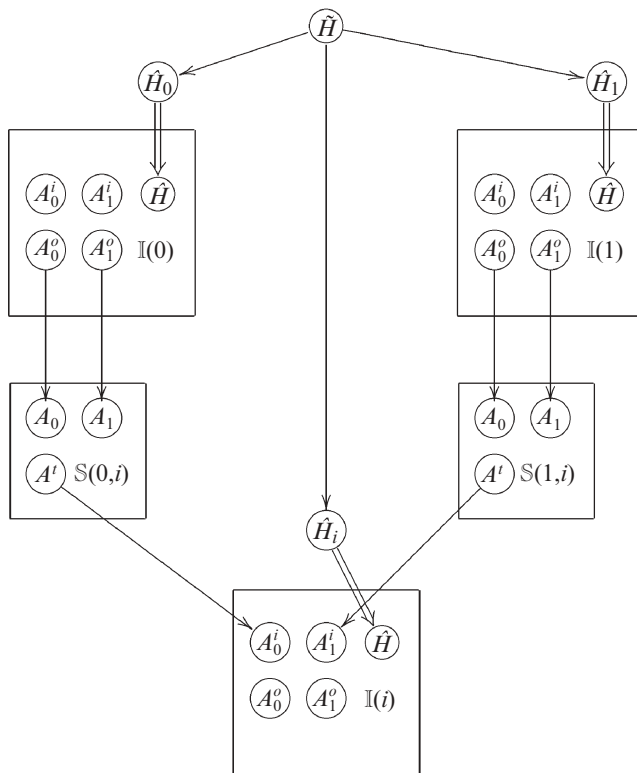


Fig. 3. The family class  $F$  when  $\mathcal{F} = \{0, 1, i\}$ .

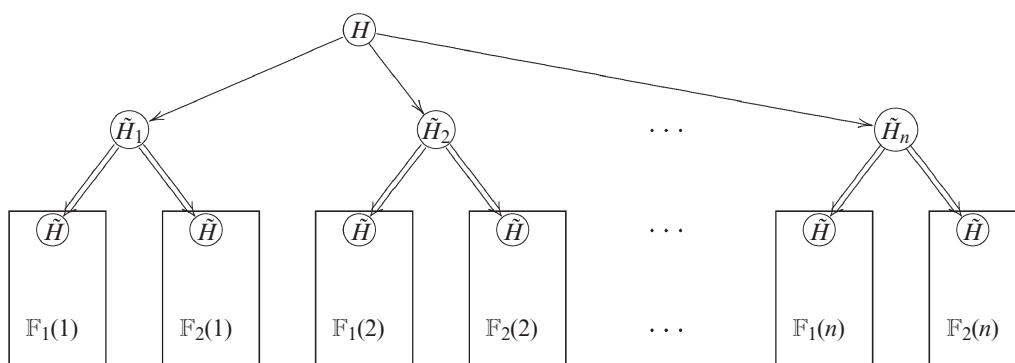


Fig. 4. The object-oriented Bayesian network for a two loci database search.

This latter can be easily derived from the posterior and it is the most classical measure provided by the expert to the court, along with other likelihood ratios obtained using different sources of evidence.

#### 4. APPLICATIONS

We now describe some applications realised by a computer program written in the Java language. First, we simulated a database containing 1000 unrelated individuals on 15 loci. Then, for each individual, we sampled a child and a sibling.

The experiment considers one of the elements of the children set as a trace of unknown origin. Then a search among the families, consisting of 1000 first-degree-relative pedigrees built around the individuals in the database, is performed. If the search is successful, the likelihood ratio evaluated for the family from which the considered trace was generated is expected to have one of the highest values. For this reason this rank, as well as the ranks obtained by replicating the search for the remaining 999 elements of the children set, are representative of the result of the analysis. A similar experiment is performed using the set of siblings as traces to be identified. The results of the two experiments are in Table 1.

Concerning the identification of a child, in over 98% of the cases, the likelihood ratio corresponding to the originating family has the highest value; the identification of a brother is slightly less successful, since in this case the figure is 80%. In real cases, it seems safe to suggest that the results' evaluation should include a comparison between the likelihood ratios for the families exhibiting the highest values, associated with careful investigative work.

A further simulation experiment has been performed, making use of different database sizes in the range 5000–50 000 and a varying number of loci in the range 5–15. We estimate the dependence of the required CPU time,  $t$ , expressed in minutes with respect to the database size,  $n$ , and the number of loci,  $l$ , according to the model  $\log(t) = \alpha + \beta \log(n) + \theta \log(l) + e$ , where  $e$  is the zero-mean stochastic error. Results are in Table 2.

Clearly the estimation of the  $\beta$ 's and the  $\theta$ 's produced very similar results and the difference in technology is accounted for by  $\alpha$ . Since the estimate of  $l$  is close to 1,  $t$  is approximately  $O(n)$  and therefore the search is feasible also when huge databases are involved.

Finally we performed the search on a real database of 2274 records accumulated over the last six years. The database is merely a collection of profiles coming from different

Table 1. *The rank distributions of the likelihood ratio supporting the correct identification hypothesis*

Rank	Sibling	Child
1	79.9%	98.5%
2	9.2%	1.5%
3	3.6%	0%
4	1.4%	0%
5	0.8%	0%
6	0.6%	0%
below 6	4.5%	0%

Table 2. *Parameter estimates of the model for CPU time in minutes, as a function of database size and number of loci required for a database search*

CPU	$\alpha$	$\beta$	$\theta$
Pentium IV	-7.710	0.981	1.172
XEROX	-9.102	1.010	1.118

sources such as crime scenes and relatives of people involved in the cases, along with all the personnel working in the forensic laboratory.

The search consisted of evaluating the likelihood ratio for each profile to be a member of the remaining 2273 families. To reach this result, 2274 searches were performed.

In Table 3 we consider the highest likelihood ratio obtained in each search, distinguishing between individuals with or without a match in the database and with the matching individuals classified according to the number of shared loci.

Matching individuals are to some extent not very interesting for us since they could have been found by the police officers with no need for a probabilistic search. More interesting is to note those 39 individuals who have a very high likelihood ratio even if they have, on average, 5.6 different loci. Increasingly often empirical searches on forensic databases are performed but, up to now, this activity has been carried on in a deterministic manner, by relating only those individuals who differ on 1 or 2 loci or who share at least one allele for each locus.

Traditionally, the search procedure is oriented towards comparing a new DNA profile with those in database. Our results make possible the reconsideration of previously unsolved cases. In this perspective, given the massive amount of computations required, the efficiency achieved is of paramount importance.

Table 3. *Distribution according to a match/no-match classification of the highest likelihood ratio obtained in a search on a real database*

Likelihood ratio	Match on a number of loci		No match
	> 7 loci	≤ 7 loci	
0–10000	0	55	1727
10000–1000000	18	24	185
> 1000000	192	34	39

## 5. DISCUSSION

In this paper the Bayesian network technology is invoked to infer about the originator of a trace when there is no clue about its origin except for a list of well-identified individuals, not apparently related to the crime. This approach is all the more effective when no match is found and an augmented database is introduced, assuming that all its members belong to the population of the crime sample's possible donors, even if many of them are not observed.

In this perspective the object-oriented Bayesian network approach provides the most striking solution, establishing a hierarchy among well-identified objects, that is with concise Bayesian networks representing repetitive parts of the problem.

This solution makes the net more readable and saves effort when maintenance operations are required. For instance, if a mutation in the allele transmission or a silent allele is allowed, a slight modification of the segregation class produces the result.

At the same time the proposed solution leaves some room for operating on some realisations of the classes. This is necessary not only to exclude from the search some individuals clearly not involved in the case, but also to tailor the most suitable pedigree for each considered family, as required by other applications of the methodology.

This point arose in the identification of the victims of a disaster. Many different familial groups were asked to identify among the victims one, or more than one, of their missing relatives. Obviously, it was required to consider more than one family class and to intervene in the network accordingly.

Indeed, the identification of each victim as one of the missing individuals claimed by the family groups is one of the two possible ways of setting the problem up. The other is to evaluate for each missing person the chance of being one of the victims. Obviously the most comprehensive solution is to consider both approaches simultaneously, conditioning the identification hypotheses on all the available evidence. The solution of this new problem will be presented in a forthcoming paper.

#### ACKNOWLEDGEMENT

We wish to thank G. Lago for having posed the problem, the anonymous referees for their hints, A. P. Dawid for an important point he raised on an early version of the paper and J. Mortera for many useful comments and suggestions. The work was partially supported by RaCIS Arma dei Carabinieri, Italy.

#### APPENDIX

##### *Proof of proposition 1*

When the variable  $X_c^*$  receives a piece of evidence  $x \in \mathcal{X}$  it is easy to show that the conditional probability table attached to that node can be written as product of  $n$  ‘findings’ (Cowell et al., 1999, pp. 93–4) which establishes that all mediating variables  $Z_j$  take the value  $x$  with probability 1. Therefore, equation (2) becomes

$$\hat{P}(X_c = x, |X, \tilde{H}) = C(x) \prod_{j=1}^n \tilde{P}(Z_j = x | X_j, \tilde{H}_j). \quad (\text{A1})$$

If  $\tilde{H} = 1_j$ , then from (1) and Rule 1 in § 2 we have that

$$P(X_c = x, |X_j, H = j) = C(x)\theta_x^{n-1} \tilde{P}(Z_j = x | X_j, \tilde{H}_j = 1).$$

Therefore, considering Rule 2 and Assumption 3 in § 2 we obtain that the above equation holds when  $C(x) = \theta_x^{1-n}$ .

The same result is achieved also for  $\tilde{H} = 0$ . In fact, in that case, if we consider (1) and Rule 1 in § 2, equation (A1) can be written as follows:

$$P(X_c = x, |H = r) = C(x)\theta_x^n.$$

Finally, from Assumption 4 we obtain again  $C(x) = \theta_x^{1-n}$ .

#### REFERENCES

- BALDING, D. J. & DONNELLY, P. (1996). Evaluating DNA profile evidence when the suspect is identified through a database search. *Forensic Sci.* **41**, 603–7.
- BANGSO, O. & WUILLEMIN, P. H. (2000). Top-down construction and repetitive structures representation in Bayesian networks. In *Proc. Thirteenth Int. Florida Artif. Intel. Res. Symp. Conf.*, Ed. J. Etheredge and B. Manaris, pp. 282–6. Menlo Park, CA: AAAI Press.
- CANNINGS, C., THOMPSON, E. A. & SKOLNIK, M. (1978). Probability functions on complex pedigrees. *Adv. Appl. Prob.* **103**, 26–61.
- COWELL, R. G. (2003). FINEX: a probabilistic expert system for forensic identification. *Forensic Sci. Int.* **134**, 196–206.
- COWELL, R. G., DAWID, A. P., LAURITZEN, S. L. & SPIEGELHALTER, D. J. (1999). *Probabilistic Networks and Expert Systems*. New York: Springer-Verlag.

- DAWID, A. P. (1994). The island problem: coherent use of identification evidence. In *Aspects of Uncertainty: A Tribute to D. V. Lindley*, Ed. P. R. Freeman and A. F. M. Smith, pp. 159–70. Chichester, U.K.: Wiley.
- DAWID, A. P. (2001). Comment on a paper by A. Stockmarr. *Biometrics* **57**, 976–80.
- DAWID, A. P. & MORTERA, J. (1996). Coherent analysis of forensic identification evidence. *J. R. Statist. Soc. B* **58**, 425–43.
- DAWID, A. P., MORTERA, J., PASCALI, V. L. & BOXEL, D. V. (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scand. J. Statist.* **29**, 577–95.
- DONNELLY, P. & FRIEDMAN, R. D. (1999). DNA database searchers and the legal consumption of science evidence. *Michigan Law Rev.* **974**, 931–84.
- JENSEN, F. V. (2001). *Bayesian Networks and Decision Graphs*. New York: Springer-Verlag.
- LAURITZEN, S. L. & SHEEHAN, N. A. (2003). Graphical models for genetic analyses. *Statist. Sci.* **18**, 489–514.
- MEESTER, R. & SJERPS, M. (2003). The evidential value in the DNA database search controversy and the two stain problem. *Biometrics* **59**, 727–32.
- MORTERA, J., DAWID, A. P. & LAURITZEN, S. L. (2003). Probabilistic expert system for DNA mixture profiling. *Theor. Pop. Biol.* **63**, 191–205.
- NATIONAL RESEARCH COUNCIL COMMITTEE ON DNA FORENSIC SCIENCE (1992). *DNA Technology in Forensic Science*. Washington, D.C.: National Academy Press.
- NATIONAL RESEARCH COUNCIL COMMITTEE ON DNA FORENSIC SCIENCE (1996). *The Evaluation DNA of Forensic DNA Evidence*. Washington, D.C.: National Academy Press.
- PEARL, J. (1988). *Probabilistic Reasoning on Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.
- PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. (2000). Inference in population structures using multilocus genotype data. *Genetics* **155**, 945–59.
- STOCKMARR, A. (1999). Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search. *Biometrics* **55**, 671–7.

[Received December 2004. Revised January 2006]