Dipartimento di Statistica
"Giuseppe Parenti"

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

WORKING PAPER 2004/04

OOBN For Forensic Identification via a Search in a Database of DNA profiles

David Cavallini, Fabio Corradi,
Giuseppina Guagnano

Università degli Studi
di Firenze

Statistics

OOBN for Forensic Identification Via a Search in a Database of DNA Profiles

David Cavallini
University of Florence,Italy.
Fabio Corradi†
University of Florence,Italy.
Giuseppina Guagnano
University "La Sapienza" of Rome,Italy.

Summary. In this paper we evaluate the evidence for pairs of competitive and exhaustive hypotheses derived from a characteristic observed on a crime sample and on individuals contained in a database. The subject considered here takes into account a debate which has recently appeared in the literature concerning the appropriateness of different sets of hypotheses. First we demonstrate the problem via a computational efficient Bayesian Network obtained by transforming some recognized conditional specific independencies into conditional independencies. Then an Object Oriented Bayesian Network representation is proposed first for a generic characteristic, then considering inheritable DNA traits. In this respect we show how to use the Object Oriented Bayesian Network to evaluate the hypotheses that some individuals genetically related to the database members 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.

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 available evidence.

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

Typically, this happens for time series models where a certain structure, a time slice, is replicated over time, and links between random variables in different time slices are established. This behavior also occurs when we are interested in the relations between members of a set 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 the growth depends on the number of sets involved.

A solution to the problem can be found through the Object Oriented Bayesian Networks (OOBN) approach. It essentially consists in considering classes of objects related to one other at different levels in a well specified hierarchy. The subject is developed in Koller and Pfeffer (1997) and Bangso and Wuillemin (2000) and the goal of this paper is to show how

†E-mail for correspondence: corradi@ds.unifi.it

a dimension dependent problem, almost intractable by making use of a simple BN, can be tackled once it is reformulated as an OOBN.

We consider, specifically, the forensic identification problem arising when a crime sample has been found but there is no clue about its origin. A search in a database (DB) is in order and the scope of the analysis is to evaluate the probability for each member of the database to be the origin of the trace. The problem has found a considerable attention in the literature (see e.g. Donnelly and Friedman, 1999 for a comprehensive and critical review). Here, starting from an intuitive BN representation of the DB search problem for a not inheritable characteristic, we provide a solution in an OOBN form transforming some recognized conditional specific independencies (Geiger and Hackerman, 1996) into conditional independencies, section (3). The OOBN representation proves useful especially when we consider more complex genetic traits and when the relations between individuals of the same lineage is represented via a BN (Dawid et al. 2002). This allows to extend the search to the relatives of the individuals in the database, providing hints also when no-match between the crime sample and one (or more) of the database members is found, section (4). Then we provide the results of a simulation study based on a real database, section (5) and finally we drawn some conclusions.

2. Background

A BN, $\mathfrak{B}_{\mathbf{U}}(\mathcal{D}, \mathbb{P})$ or more succinctly $\mathfrak{B}_{\mathbf{U}}$, is defined as a pair of objects: a Directed Acyclic Graph (DAG), $\mathcal{D}$, whose nodes, $\mathbf{U}$, represent discrete random variables, and a set, $\mathbb{P}$, of Conditional Probability Tables (CPT) which defines the conditional distributions of each vertex given the parents.

Each node is independent of its non-descendants conditional to the parents, so the joint distribution of $\mathbf{U}$ can be factorized as a product of CPTs (Pearl, 1988).

One of the main advantages of codifying a probabilistic model through a BN is the reduction of the computational efforts for calculating the conditional probability of the interesting unobserved nodes (*query* variables) given the observed ones (*evidence*). This task can be achieved by using one of the available propagation algorithms (e.g. Jensen, 2001).

In many real world applications, as in forensic science, the domain is formed by a large number of variables and often the complexity of the related network does not allow a compact representation. 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.

As regards notation, the upper-case letters stand for random variables and corresponding lower-case letters are used to indicate a specified event or state. The vectors of random variables are denoted with bold upper-case letters and a particular realization or configuration is indicated with bold lower-case letters. Last, lower-case Greek letters represent parameters.

3. The Database Search Problem: BN vs OOBN

Let X the 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 $P(X = x) = \theta_x$. Obviously, the following condition holds

$$\sum_{x \in \mathcal{X}} \theta_x = 1. \tag{1}$$

Let N the (finite) 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 $j \in \mathbb{I} = \{1, 2, \ldots, n\}$. Moreover, we define X_c, the characteristic related to the crime scene, and the hypotheses random variable H which has $n + 1$ states. The first n of them represent the originator status of each single individual, that is, $H = j$, with $j \in \mathbb{I}$, means that the origin of the trace is the j-th individual in the DB while the last, $H = \texttt{rest}$, is referred to the hypothesis that the donor of the trace is outside the DB.

The basic assumptions of the model are:

i. the individuals in the DB are not stochastically related, i.e., $\forall j \neq t$, $X_j \perp\!\!\!\perp X_t$;

ii. the characteristic of the individuals in the DB is *pure*, i.e. is independent of hypotheses variable H, $\forall j$, $X_j \perp\!\!\!\perp H$;

iii. for $H = j$ the characteristics of the rest of the individuals, $\mathbf{X}_{-j}$, are independent X_c, i.e., $\forall j$, $X_c \perp\!\!\!\perp \mathbf{X}_{-j} \mid H = j$ where $\mathbf{X}_{-j} = \{X_i : i \in \mathbb{I} \setminus \{j\}\}$;

iv. for $H = \texttt{rest}$ the set of attributes of the individuals, $\mathbf{X} = \{X_j; j \in \mathbb{I}\}$, is independent of X_c, that is, $\mathbf{X} \perp\!\!\!\perp X_c \mid H = \texttt{rest}$ and $P(X_c = x \mid H = \texttt{rest}) = \theta_x$ with $x \in \mathcal{X}$;

v. for $H = j$ the trace observed on the crime scene is left with error and this error is symmetric, $\forall j$,

$$P(X_c = x \mid X_j = \hat{x}, H = j) = \begin{cases} \beta & \text{if } x = \hat{x} \\ \frac{1-\beta}{m-1} & \text{if } x \neq \hat{x} \end{cases} \tag{2}$$

with $x, \hat{x} \in \mathcal{X}$ and $\beta \in (0, 1]$;

vi. no other clue is available in advance, so the prior probability on H is not informative $P(H = j) = 1/N$ and $P(H = \texttt{rest}) = 1 - n/N$.

Note that (*iii*) and (*iv*) is 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 (Dawid, 1979), since, in the latter, the independence assertions between variables do not vary according to the values of the conditioning sets.

The graphical structure, which derives from the assumptions (*i*) and (*ii*), is depicted in Figure (1) and the CPTs attached to the nodes are specified according to the assumptions (*iii*)-(*vi*).

The proposed naive network does not feature any conditional independence, so, for some evidence, the probability updating does not take advantage of the graphical representation.

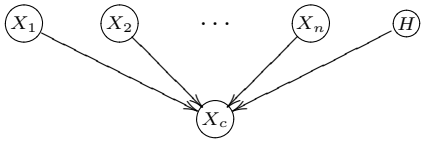

Fig. 1. A DAG for the DB search problem.

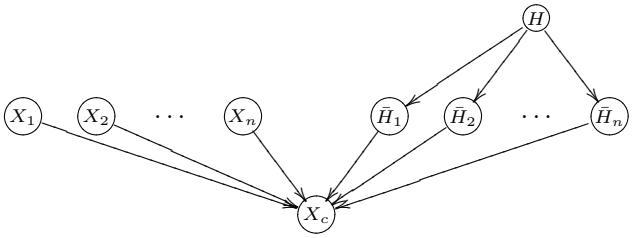

Fig. 2. The augmented DAG obtained from Figure (1).

Moreover, the size of the CPT of X_c increases exponentially with respect to the number of individuals in the DB, so that the propagation becomes rapidly unfeasible. As we will make further, this problem becomes relevant when an inheritance characteristic is considered and some unobserved individuals related to any member of the DB are considered for the forensic identification problem.

Our goal 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 $\bar{\mathbf{H}} = \left\{\bar{H}_j : j \in \mathbb{I}\right\}$ is added and a new network, $\mathfrak{B}_{\hat{\mathbf{U}}}$, is defined on the augmented domain $\hat{\mathbf{U}} = \mathbf{U} \cup \bar{\mathbf{H}}$ 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:

$$\hat{P}(\bar{H}_j = 1 \mid H = i) = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$\hat{P}(X_c \mid \mathbf{X}, \bar{\mathbf{H}} = \bar{\mathbf{h}}) = \begin{cases} P(X_c \mid X_j, H = j) & \text{if } \bar{\mathbf{h}} = \mathbf{1}_j \\ P(X_c \mid H = \texttt{rest}) & \text{if } \bar{\mathbf{h}} = \mathbf{0} \\ m^{-1} & \text{otherwise} \end{cases} \tag{4}$$

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 attached to each node $\bar{H}_j$, specified as in (3), is 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

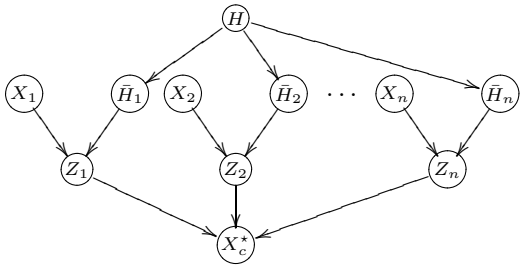

Fig. 3. The augmented DAG of Figure (2) after the divorce.

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:

$$\sum_{\bar{\mathbf{H}}} \hat{P}(X_c, \mathbf{X}, \bar{\mathbf{H}}, H) = \sum_{j=1}^{n} \hat{P}(X_c, \mathbf{X}, \bar{\mathbf{H}} = \mathbf{1}_j, H) + \hat{P}(X_c, \mathbf{X}, \bar{\mathbf{H}} = \mathbf{0}, H) = P(X_c, \mathbf{X}, H). \quad (5)$$

Since the hypotheses are mutually exclusive, all configurations of $\bar{\mathbf{H}}$ not equal to the $\mathbf{1}_j$s and $\mathbf{0}$ have a probability zero to realize. For this reason, in the marginalization (5), we consider only the relevant configurations of $\bar{\mathbf{H}}$.

The main consequence of the above result concerns the probability 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 $\mathfrak{B}_{\mathbf{U}}$ or $\mathfrak{B}_{\hat{\mathbf{U}}}$.

Step 2. Here a *divorcing* technique (Jensen, 2001) is applied. The idea is to introduce a set of mediating variables between the parents and their child of a large converging connection. The role of the mediating variables is 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 the network $\mathfrak{B}_{\hat{\mathbf{U}}}$, is to add n mediating variables $\mathbf{Z} = \{Z_j : j \in \mathbb{I}\}$, 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. We denote it with $\tilde{\mathcal{D}}$ while we use $\tilde{\mathbf{U}}$ for indicating its domain, that is, $\tilde{\mathbf{U}} = \{\mathbf{X}, \bar{\mathbf{H}}, H, \mathbf{Z}, X_c^\star\}$. 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 $\mathcal{X}^\star = \mathcal{X} \cup \{\text{NAN}\}$ where the state labelled NAN is an instrumental event.

The CPTs specification of the nodes $\mathbf{X}$,$\bar{\mathbf{H}}$ and H remains unchanged with respect to $\mathfrak{B}_{\hat{\mathbf{U}}}$. Imposing the CSI conditions

$$\forall j, Z_j \perp\!\!\!\perp X_j \mid \bar{H}_j = 0, \quad (6)$$

the rest of CPTs are specified as follows

$$\tilde{P}(Z_j = x \mid \bar{H}_j = 0) = \theta_x \quad (7)$$

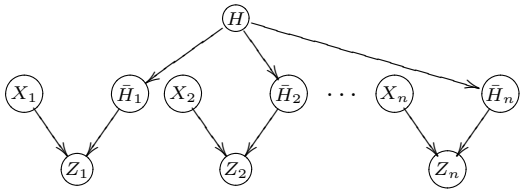

Fig. 4. The network obtained after dropping the $X_c^\star$ node and the related incidental arcs from the DAG in Figure (3)

$$\tilde{P}(Z_j = x \mid X_j = \hat{x}, \bar{H}_j = 1) = \begin{cases} \beta & \text{if } x = \hat{x} \\ \frac{1-\beta}{m-1} & \text{if } x \neq \hat{x} \end{cases} \tag{8}$$

$$\tilde{P}(X_c^\star = \bar{x} \mid \mathbf{Z} = \mathbf{z}) = \begin{cases} 1 & \text{if } \bar{x} = \text{NAN or } \forall j, \bar{x} = z_j \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

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 $\mathfrak{B}_{\tilde{\mathbf{U}}}$ and $\mathfrak{B}_{\tilde{\mathbf{U}}}$.

PROPOSITION 3.1. *For each $x \in \mathcal{X}$ and for a given constant $C(x)$, depending on x, the following relation holds:*

$$\hat{P}(X_c = x, \mathbf{X}, \bar{\mathbf{H}}, H) = C(x) \cdot \sum_{\mathbf{Z}} \tilde{P}(X_c^\star = x, \mathbf{X}, \bar{\mathbf{H}}, H, \mathbf{Z}) \tag{10}$$

Finally, combining (5) with (10), we obtain the main result:

$$P(X_c = x, \mathbf{X}, H) = C(x) \cdot \sum_{\mathbf{Z}, \bar{\mathbf{H}}} \tilde{P}(X_c^\star = x, \mathbf{X}, \bar{\mathbf{H}}, H, \mathbf{Z}) \tag{11}$$

where, as usual, $x \in \mathcal{X}$. The above equation establishes that for calculating the posterior probability of the hypotheses variable H we can use $\mathfrak{B}_{\tilde{\mathbf{U}}}$ instead of $\mathfrak{B}_{\mathbf{U}}$.

Step 3. As explained in the proof of **PROPOSITION** 3.1, 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 from $\tilde{\mathfrak{D}}$ the node $X_c^\star$ as well as its incidental arcs. Moreover, we use the characteristic observed on the crime scene for evidencing each vertex Z_j.

The new graph, depicted in Figure (4), is conspicuous for a repetitive structure with respect to the individuals in the DB. For each of them 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.

A more compact representation can be achieved by transforming the proposed network in an OOBN framework. Considering the approach proposed by Bangso and Wuillemin

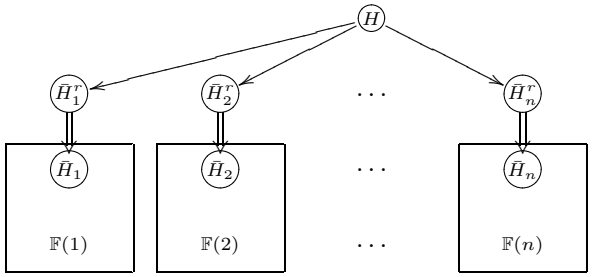

Fig. 5. The OOBN representation for the DB search problem derived from Figure (4).

(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 instantiantion of the class $\mathbb{F}(j)$, with $j \in \mathbb{I}$, 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 are drawn. Finally, the CPTs related to the variables $\bar{H}_j^r$ are specified as in (3).

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

4. OOBN for Inheritable Nuclear DNA Traits

A DNA profile concerns measurements on several well specified locations of the DNA, called *loci*. For each locus we observe two alleles, one inherited from the father and the other from the mother, even if their origin is not recoverable. In this paper we assume independence of the alleles within each locus and between the loci, i.e., we assume Hardy-Weinberg (H-W) and linkage equilibrium. 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 inheritable 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 $t \leq u$.

This inheritance allows us to consider also, as the possible donor of the crime sample, individuals never typed but genetically related to the DB members. In this way the no-match case, the most common, but unfortunately also the less useful outcome of the DB search, could increase the probability for some *compatible* individuals to be the origin of the trace. Compatible individuals are defined as 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, has a compatible child if he/she shares an allele with the crime sample at each considered locus.

Here, we consider a pedigree, $\mathcal{F}$, constituted by a generic individual (i), his parents (0 and 1), his sibling (s), his partner (p) and his brother (b). Note that, the labels 0 and 1 are referred to a generic parent and not specifically to the mother or father because the information concerning inheritance is not available. Since this pedigree is built around a generic individual we call it a *one-generation-around* pedigree.

In this perspective, the variables H and H_j^r, shown in Figure (5), have a new meaning. The j-th state of H, with $j \in \mathbb{I}$, is referred to the hypothesis that the donor of the

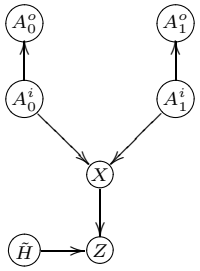

Fig. 6. The individual class

trace belongs to the family of the j-th individual of the DB while $H = \texttt{rest}$ concerns the possibility that trace was left by someone not included in the considered families. Remark that, N, used to specify the prior on H, has to be interpreted as the number of the families composing the reference population.

Furthermore, every variable $\bar{H}_j^r$ takes values in $\bar{\mathcal{F}} = \mathcal{F} \cup \texttt{rest}$. The state `rest` concerns the hypothesis that the trace is left by none of the considered family while the statement $\bar{H}_j^r = q$, with $q \in \mathcal{F}$ means that the donor of the trace is exactly the q-th member of $\mathcal{F}$. Since, we do not have any clue in advance the CPT attached to each referenced node $\bar{H}_j^r$ is specified as follows

$$P(\bar{H}_j^r = \bar{h} \mid H = i) = \begin{cases} 1/6 & \text{if } j = i \text{ and } \bar{h} \neq \texttt{rest} \\ 1 & \text{if } j \neq i \text{ and } \bar{h} = \texttt{rest} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where $i, j \in \mathbb{I}$ and $\bar{h} \in \mathcal{F}$.

In this respect the class $\mathbb{F}$ includes the one-generation-around pedigree and the set of hypotheses variables related to a generic family. Considering the *Allele Network* proposed by Lauritzen and Sheehan (2002), we provide an OOBN representation of $\mathbb{F}$. To do so, we need to define two other classes: the *Individual* ($\mathbb{I}$) and the *Segregation* ($\mathbb{S}$) class.

The individual class' inner structure is represented in Figure (6). If no information about the individual's parents is available, the allele input nodes A_0^i e 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 $\tilde{H}$ representing the originator status of a generic individual. To provide the transmission of the individual genetic characteristics to the siblings, a copy of the alleles is expressed by output nodes (A_0^o e A_1^o), the other vertexes X and Z being interior nodes. The variable X denotes the observable genotype and its CPT is specified as follows

$$P(X = (a_r, a_u) \mid A_0^i = a_h, A_i^1 = a_t) = \begin{cases} 1 & \text{if } (h = r \text{ and } t = u) \text{ or } (h = u \text{ and } t = r) \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

while Z plays the instrumental role explained in section (3) and its CPT is built according to equations (6), (7) and (8).

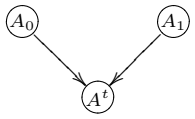

Fig. 7. The segregation class

The segregation class' structure, Figure (7), has two alleles input nodes and provides the selection mechanism to generate the transmitted allele A^t via the following CPT, which reflects the first Mendelian law:

$$P(A^t = a_r \mid A_0 = a_t, A_1 = a_u) = \begin{cases} 1 & \text{if } r = t = u \\ 0.5 & \text{if } (r = t \text{ and } r \neq u) \text{ or } (r = u \text{ and } r \neq t) \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

In the overall, 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 alleles donor while t denotes the member that receives the allele after the segregation. The links among the instantiations of 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 $\tilde{H}^r_q$. All of them are mixed by the only input node $\bar{H}$ and the related CPTs are built as follows

$$P(\tilde{H}^r_q = 1 \mid \bar{H} = u) = \begin{cases} 1 & \text{if } q = u \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

with $u \in \bar{\mathcal{F}}$ and $q \in \mathcal{F}$. In Figure (8) we give a simple example of $\mathbb{F}$ assuming that $\mathcal{F} = \{0, 1, \mathtt{i}\}$.

Under the linkage equilibrium assumpion we build, for each locus $l \in \mathbb{L} = \{1, 2, \ldots, k\}$, a BN, $\mathfrak{B}_l$, as in Figure (5). The genotype of the j-th individual observed on the l-th locus is used for entering evidence on the node $\mathfrak{B}_l.\mathbb{F}(j).\mathbb{I}(\mathtt{i}).X$. Moreover, all the interior vertexes $\mathfrak{B}_l.\mathbb{F}(j).\mathbb{I}(\mathtt{q}).Z$, with $j \in \mathbb{I}$ and $q \in \mathcal{F}$, receive as evidence the (same) genotype observed on the l-th locus of the crime sample.

The propagation provides all the probabilities we need to compute the Weight of Evidence (WE). This measure is used to evaluate the support given by the genetic evidence ($\mathcal{E}$) to an identification hypothesis of interest ($\mathcal{H}$). The WE cannot be read directly by the net, but it can be derived from

$$\text{WE} = \frac{P(\bar{\mathcal{H}})}{P(\mathcal{H})} \cdot \frac{P(\mathcal{H} \mid \mathcal{E})}{P(\bar{\mathcal{H}} \mid \mathcal{E})}. \tag{16}$$

If linkage equilibrium holds the overall WE is just the product of each single WE evaluated for each locus.

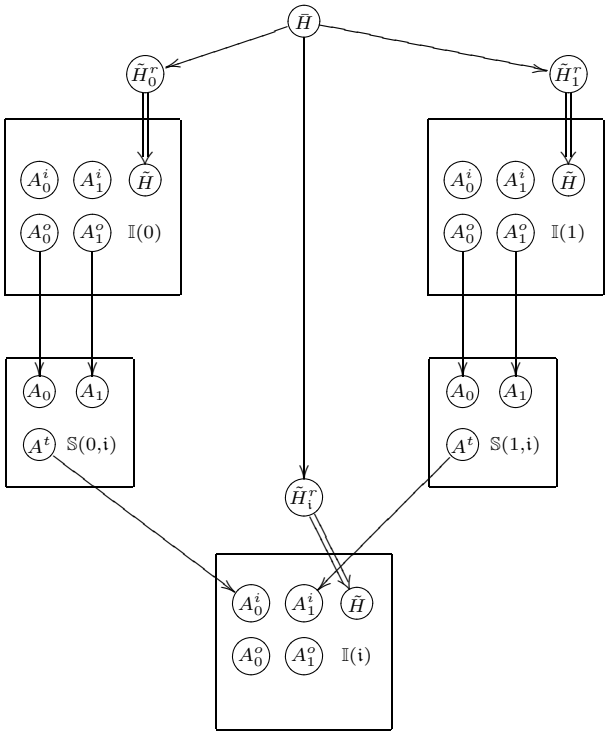

Fig. 8. The family class $\mathbb{F}$ when $\mathcal{F} = \{0, 1, \mathrm{i}\}$.

5. An Application Using a Real DB

Now let us to give account of some simulations on a real (small) DB containing 100 observations on 13 loci. The members of the DB are unrelated and we assume that all the one-generation-around individuals belong to the originator population.

The size of the originator population was set to one million and the prior on H is assumed to be uniform.

For each observed individual, we generated two crime samples obtained respectively from the posterior marginal distribution of the child's and the sibling's genotypes. We call them the child-crime-samples and the sibling-crime-samples.

For each child-crime-sample, we evaluate two hypotheses: one concerns the identification of the child for each member of the DB, the other considers the possibility that the crime sample comes from a generic member of each one-generation-around family. Similar computations are provided if the sibling-crime-samples are used.

Concerning the identification of a child, 98 out of 100 of the WEs supporting the *correct* identification hypothesis have the highest values compared to the WEs relating the simulated child to the other 99 families. In the same simulation, the remaining 2 WEs have got the second highest values. The identification of a *brother* was a slightly less successful, since he is not a direct lineage: 91 out of 100 WEs have got the highest values; 4 of 100 the second highest value, finally the less successful case assumed the seventh highest WE.

As a comment, it must be noted that our simulation is conservative in nature since, for instance, in sampling a sibling-crime-sample we do not know the relatives' genotypes but we sample from their posterior distribution conditional to the genotype of just one of their siblings. In real cases, where the relatives' genotypes are *knows* by nature, brothers' genotypes are often very similar: for each locus, the fact that only one of the parents is homozygote is sufficient for the probability that brothers share one allele to be equal to one and the probability they are identical is equal to 0.5.

6. Conclusions

The use of BN to provide an evaluation of the weight of evidence for forensic identification purposes is a new but already well established approach 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 there is a list of well identified individuals, not apparently related to the crime, 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 possible donors of the crime sample, 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. In this view, even a subject in the original DB thrown out from the hypotheses evaluation, should be kept since the possibility to evaluate genetically related individuals would not be excluded.

Proof of Proposition 3.1

The joint marginal distribution of $\{\mathbf{X}, \bar{\mathbf{H}}, H\}$ is the same in the two BNs $\mathfrak{B}_{\hat{\mathbf{U}}}$ and $\mathfrak{B}_{\tilde{\mathbf{U}}}$, so (10) becomes

$$\hat{P}(X_c = x, \mid \mathbf{X}, \bar{\mathbf{H}}) = C(x) \cdot \sum_{\mathbf{Z}} \tilde{P}(X_c^\star = x, \mid \mathbf{Z}) \cdot \prod_{j=1}^{n} \tilde{P}(Z_j \mid x_j, \bar{H}_j) \tag{17}$$

When the variable $X_c^\star$ receives an evidence $x \in \mathcal{X}$ it is easy to show that after the reduction (9) can be written as product of n potential ϕ_j, that is

$$\hat{P}(X_c^\star = x \mid \mathbf{Z}) = \prod_{j=1}^{n} \phi_j(Z_j) \tag{18}$$

where

$$\phi_j(Z_j = \hat{x}) = \begin{cases} 1 & \text{if } \hat{x} = x \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

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

The equation (19), which defines a *finding* on Z_j, establishes that all mediating variables take value x with probability 1. So, combining equations (18) and (19) with (17) we obtain that

$$\hat{P}(X_c^\star = x, \mid \mathbf{X}, \bar{\mathbf{H}}) = C(x) \cdot \prod_{j=1}^{n} \tilde{P}(Z_j = x \mid X_j, \bar{H}_j). \tag{20}$$

If $\bar{\mathbf{H}} = \mathbf{1}_j$ then from (4) and (6) we have that

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

The third part of the right side of the above equation is a product which involves $n-1$ terms. From (7), each one of them is equal to θ_x so, comparing (2) with (8) 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 (4) and (6), the equation (20) becomes

$$P(X_c = x, \mid H = \texttt{rest}) = C(x) \cdot \prod_{j=1}^{n} \tilde{P}(Z_j = x \mid \bar{H}_j = 0).$$

Finally, from condition (iv) and equation (7) we obtain again that $C(x) = \theta_x^{1-n}$.

REFERENCES

O. Bangso and P-H. Wuillemin (2000). Object Oriented Bayesian Networks A Framework for Topdown Specification of Large Bayesian Networks and Repetitive Structures. Technical Report, *BBS group Department of Computer Sciences, University of Aalborg.*

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 (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society*, B41, 1-31.

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.

A. P. Dawid and J. Mortera and V.L. Pascali and D. Van Boxel (2002). Probabilistic Expert Systems for Forensic Inference from Genetic Markers. *Scandinavian Journal of Statistics*, 29, 577-595.

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.*

S. L. Lauritzen and N. A. Sheehan (2002). Graphical Model for Genetic Analyses. Technical Report R-02-2020, *Department of Mathematical Sciences, University of Aalborg.*

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

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

J. Pearl (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.*Morgan Kaufmann Publishers.*