

Some issues on the identification of relatives of individuals included in a database of DNA profiles

Fabio Corradi

Dept. of Statistics University of Florence, e-mail: corradi@ds.unifi.it

Abstract: 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. First the problem is considered with respect to non inheritable characteristics then we turn our attention to an 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.

Keywords: Bayesian networks; Forensic identification; Database of DNA profiles

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 heritability is not exploited, Dawid and Mortera (1996) and Donnelly and Friedman (1999) found solutions to cases arising in a database search.

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 get the posteriors for each identification hypothesis is straightforward. However if assertions of conditional independence are not exploited, the required marginalization procedure makes the search computationally unfeasible even for a small data base. A solution can be achieved if the database problem is formulated in terms of a Bayesian network (BN), a graphical representation based on a directed acyclic graph and a set of conditional probability tables (CPT) where the computation of the conditional probability of unobserved nodes in a BN is realized locally, exploiting conditional independencies, (Cowell *et al.*, 1999). Lauritzen and Sheehan (2003) and Dawid *et al.* (2002).

2. The database search problem

Let $X_j, j \in \mathcal{J} = \{1, 2, \dots, n\}$, 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 θ_x denotes the probability that a generic random variable X_j is in state $x \in \mathcal{X}$, i.e. for all $j \in \mathcal{J}$, $P(X_j = x) = \theta_x$, where the parameters $\theta = \{\theta_x : x \in \mathcal{X}\}$ are the relative frequencies of the characteristic in the reference population, whose size, N , is assumed known. We also introduce a hypothesis random variable H : the first n states represent the originator status of each individual in the database and $H = \tau$, means that the donor of the trace is not in the database and we also adopt some reasonable assumptions.

Assumption 1. For all $i, j \in \mathcal{J}$, $X_i \perp\!\!\!\perp X_j | \theta$.

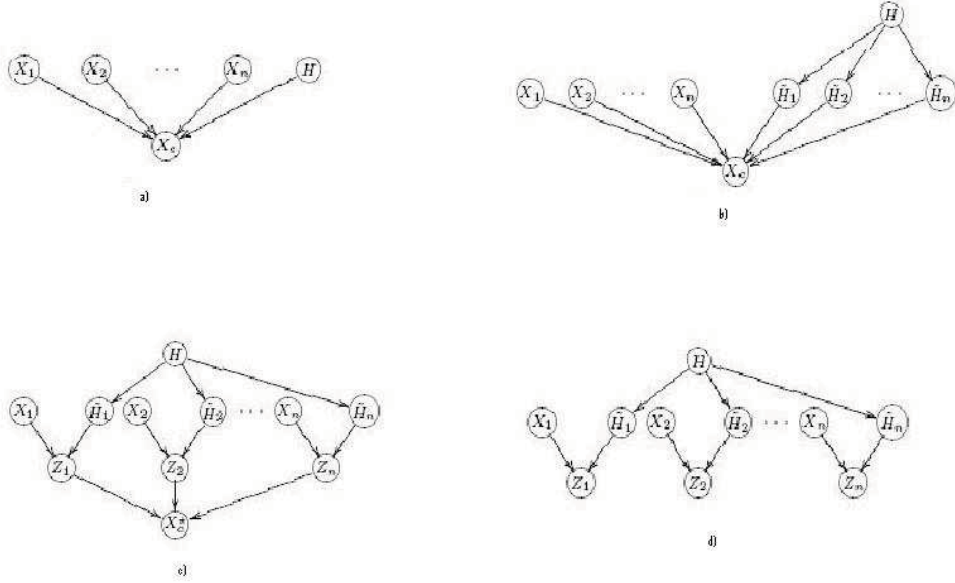
Assumption 2. For all $i, j \in \mathcal{J}$, $X_i \perp\!\!\!\perp H$

Assumption 3. If individual j is the donor of the trace $X_j = X_c | H = j$.

Assumption 4. For $H = r$ and for all $i, j \in \mathcal{J}$, $X_i \perp\!\!\!\perp X_j | H = r$

Assumption 5. No other clue is available, so $P(H = j) = \frac{1}{N}$ and $P(H = r) = \frac{N-n}{N}$.

Figure 1: Some possible directed acyclic graphs for the database search problem



The network in Fig. 1(a), derived from Assumptions 1 and 2 does not feature any conditional independence. An equivalent BN, is obtained first introducing a set of binary random variables $\tilde{H} = \{\tilde{H}_j : j \in \mathcal{J}\}$ representing the originator states of the individuals as depicted in Fig. 1(b). Since the $\tilde{H}_j$ are mutually exclusive, their CPTs can be specified considering:

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

Then, the CPT related to X_c is:

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

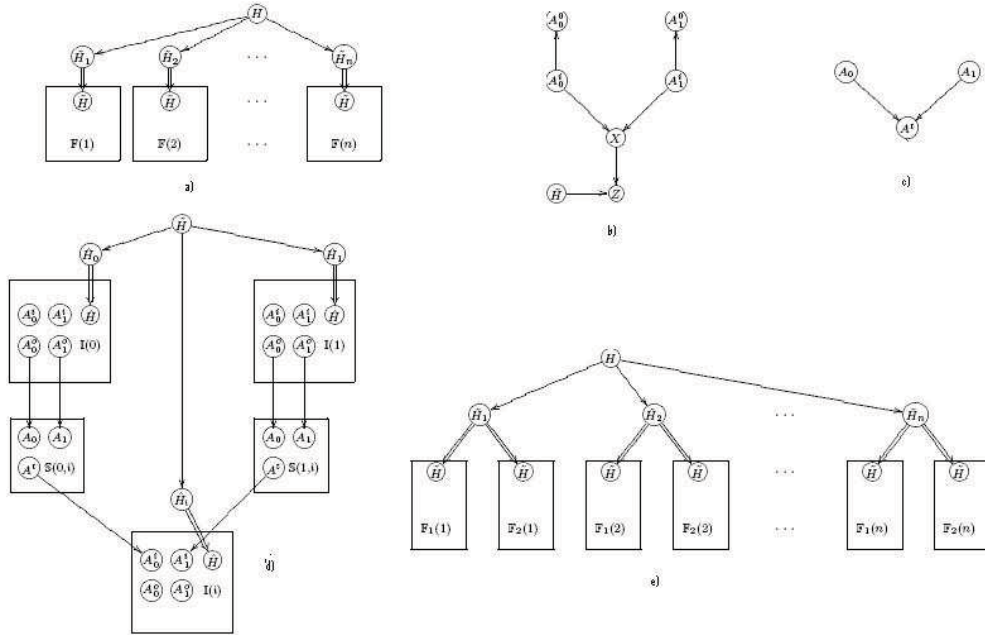
where $\vec{0}$ is a zero vector of size n while $\vec{1}_j$ has the i th element 0 for all $i \neq j$ and 1 for $i = j$. Since $\tilde{H}$ is fully determined by H , marginalizing with respect to $\tilde{H}$ the joint probability distribution induced by the directed acyclic graph of Fig. 1(b), $\hat{P}(X_c, X, \tilde{H}, H)$, exactly produces $P(X_c, X, H)$. Then we apply a 'divorcing' technique (Jensen, 2001) to divorce the parents of node X_c in the network of Fig. 1(b) adding 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 $\tilde{H}_j$

are married, Fig. 1(c). Full details on the equivalence of the models depicted in Fig. 1(b) and Fig. 1(c) are provided in Appendix 1 of Cavallini and Corradi (2006). This obviously implies equivalence of the Fig. 1(a) or in Fig. 1(c) represented models. 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 obtaining the graph depicted in Fig. 1(d): for each individual in the database the same BN is built and mixed by the hypothesis variable H which is the only parent of every $\tilde{H}_j$ and since, given H , each triple $(Z_j, \tilde{H}_j, X_j)$ is independent of the other variables local computations are allowed.

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 X whose states are an ordered pair of alleles (a_t, a_u) with $t \leq u$. Loci are chosen on different chromosomes to assure independence. Hardy-Weinberg equilibrium in the population is assumed.

Figure 2: Object oriented BNs and some related inner classes



Here, for each individual (i) in the database we consider a search in the pedigree, $\mathcal{F}$, including never typed but related to the database members, i.e. their parents, labelled as 0 and 1, their child, c , the other parent, p , of c and their sibling, s . Call it a 'first-degree-relative' pedigree.

In this new perspective, the variables H and $\tilde{H}_j$, shown in Fig. 2(a), 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 $H = \tau$ corresponds to the possibility that the trace was left by someone not included in the considered families.

Considering the allele network proposed by Lauritzen and Sheehan (2003), we provide an object oriented BN representation of $\mathbb{F}$ obtained by defining two other classes: the individual class, $\mathbb{I}$, Fig. 2(b) and the segregation class, $\mathbb{S}$, Fig. 2(c). The individual class links the individual genotype, X to the trace (Z) to determine the posterior of $\hat{H}$, the originator status of the individual. It also provides the transmission of the individual's genetic characteristics to the siblings. The segregation class has two allele input nodes and provides the selection mechanism for generating the transmitted allele A^t via the first Mendelian law. Overall, the family class $\mathbb{F}$ is defined by a set of realizations 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 allele donor while t denotes the individual that receives the alleles after the segregation. The links among the realizations of the basic classes, $\mathbb{I}$ and $\mathbb{S}$, are drawn according to the biological relationships and each input node $\hat{H}$ defined inside to each $\mathbb{I}(q)$ has its own referenced vertex, $\hat{H}_q$. All of them are mixed by the only input node $\tilde{H}$.

To extend the analysis to consider more than one independent locus, let $\mathbb{F}_l$ the family class for the l th locus. For the j th family the relationship between the realization of $\mathbb{F}_l$ classes is expressed by reference links connecting $\tilde{H}_j$ with every referenced node $\tilde{H}$ defined inside to each $\mathbb{F}_l(j)$. As an example, Fig. 2(e) depicts the object oriented BN for two loci.

4. Discussion

In this paper the BN technology is invoked to infer about the originator of a trace when there is no clue about its origin but 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 BN approach provides a striking solution establishing a hierarchy among well-identified objects, that is concise Bayesian networks representing repetitive parts of the problem.

References

- Cavallini D. and Corradi F. (2006) Forensic identification of relatives of individuals included in a database of DNA profiles, *Biometrika*, 93, 525–536.
- Cowell R.G., Dawid A.P., Lauritzen S.L. and Spiegelhalter D.J. (1999) *Probabilistic Networks and Expert Systems*, Springer-Verlag.
- Dawid A.P. and Mortera J. (1996) Coherent Analysis of Forensic Identification Evidence, *Journal of the Royal Statistical Society*, B58, 425–443.
- Dawid A.P., Mortera J., Pascali V.L. and Boxel D.V. (2002) Probabilistic Expert Systems for Forensic Inference from Genetic Markers, *Scandinavian Journal of Statistics*, 29, 577–595.
- Donnelly P. and Friedman R.D. (1999) DNA Database Searches and the Legal Consumption of Science Evidence, *Michigan Law Review*, 157, 931–984.
- Jensen F.V. (2001) *Bayesian Network and Decision Graphs*, Springer-Verlag.
- Lauritzen S.L. and Sheehan N.A. (2003) Graphical Models for Genetic Analyses, *Statistical Science*, 18, 489–514.