

## THE IDENTIFICATION OF MISSING PERSONS MAKING USE OF DNA PROFILES

**Fabio Corradi**<sup>1</sup>

*Department of Statistics, Computer Science, Applications “G.Parenti”, University of Florence, Florence, Italy*

**Abstract.** *This paper tackles the issue of identifying missing persons through probability and decision. Evidence is represented by measurements of DNA loci on an Unknown individual and on familial donors requiring to track down their missing relatives. The inferential problem is solved by reviewing, extending and commenting two different approaches recently appeared in the literature. Stemming from a more general contribution we recently provided we explain how to achieve a short-list of individuals to be further scrutinized with a view to finally ascertain if one of them is the recovered Unknown.*

**Keywords:** *Missing individuals, Database search, Bayesian decision theory, Bayesian networks.*

### 1. INTRODUCTION

According to official records (Commissario di Governo, 2015), in Italy, at the end of 2014, the number of missing persons amounted to 29.234 people, which we consider an impressive figure for a small Country as Italy.

To succeed in tracking them down depends on the causes of disappearance, on the most recent time the missing persons were seen alive and on many other circumstances. Often, the only possibility to recover them is through the joint consideration of the DNA profile of the *Unknown*( $U$ ), an individual or a corpse with no identity paper, and the genetic evidence of relatives of missing persons.

In 2005, the Prüm Convention, sometimes cited as Schengen III Agreement, signed by seven European countries, gave a considerable boost to the creation of data bases (DB) of DNA profiles and to the exchange of information. Italy accepted the Prüm Convention some years later, and Law n. 85, 2009 made it legal to feed the National DNA data base with biological samples of the relatives of missing persons and of those individuals or corpses whose identity was not otherwise ascertained.

---

<sup>1</sup> corradi@disia.unifi.it

At the time of writing, unfortunately, the Italian National DNA data-base has not come into force although guidelines for law implementation should be issued soon. This is the reason why, it seems timely to reconsider the probabilistic assessment of identification propositions and to investigate alternative methods to achieve a short list of missing persons, thus significantly limiting the number of possible candidates to identification.

## 2. BASIC DEFINITIONS AND KINSHIP ANALYSIS

DNA traits are measurements on several specific locations of the DNA called *loci*. For an individual, at each locus, we observe a *genotype*  $x = \{a_t, a_u\}$ ,  $t \leq u$ , i.e. two *alleles*,  $a$ , inherited from the parents. Generic alleles  $a_t, a_u$  are the determinations of a discrete random variable  $A = \{a_1, \dots, a_m\}$  with sample space varying among loci. To keep notation as simple as possible, we indicate with  $x$  all the genetic information available, formed by the observation on several loci, for the individual/s specified in the subscript.

For an individual, considered as a member of a population, i.e. not recognized to belong to a specific family, the probability to observe the DNA profile is evaluated by a population model and its parameters,  $\theta$ . We do not give details about the choice of the model and the statistical procedure assessing its parameters. We assume only that a reasonable choice can be done in specific circumstances.

The possibility to consider in isolation hypotheses whether the Unknown, whose genetic traits are  $x_U$ , is the missing member ( $M$ ) of a family, is called kinship analysis and relies on the heritability of the DNA traits. Following the first Mendelian law, and taking into account possible mutations, the probability that  $U$  is the specific missing individual  $M$  claimed by a family can be evaluated conditionally to the genetic information,  $x_f$ , coming from members of the familial pedigree. Let's define the following hypothesis,  $H = \{H_f, H_g\}$ .

- $H_f$ : the Unknown is the missing person claimed by a family.
- $H_g$ : the Unknown is a *generic* unidentified member of the population of missing persons.

Results are typically provided by the likelihood ratio supporting  $H_f$

$$LR = \frac{p(x_U|x_f, H_f)}{p(x_U|H_g)}, \quad (1)$$

where  $p(x_U|x_f, H_f)$  is the probability to observe  $x_U$  according to the probability of the genetic traits of  $M$  given  $x_f$  and  $p(x_U|H_g)$  is the probability to observe the genetic traits of  $U$  in a reference population from which, alternatively to the considered family, he/she could come from.

Furthermore, owing to the Bayes theorem and also introducing prior probabilities on  $H$ , we can directly derive their posterior probabilities. Unconventionally, by using (1), we have

$$p(H_f|x_U, x_f) = LR \frac{p(H_f)}{p(H_g)} / (1 + LR \frac{p(H_f)}{p(H_g)}).$$

**Remark 1:**  $p(x_U|x_f, H_f)$  can be evaluated by analytical probabilistic computations or by representing the pedigree of the claiming family by a Bayesian network (Dawid et al., 2002). After the instantiation of the relatives' nodes providing their DNA traits, the probability of  $x_U$  can be read on the marginal posterior of  $x_M$ . Then, some specific software like Familias and DNView directly provide the solution allowing for some different population and segregation models and their parameters. See Dråbek (2009) for a comparative illustration.

**Remark 2:** The well-known Essen-Moller paternity index, is a LR, where the alleged father is the Unknown, eventually identified as the true father of a baby observed with his/her mother on their genetic traits. Afterwards kinship analysis evolved (Egeland et al., 2006) and studied many kinds of relationships, also considering alternative well definite kinship relations within the same family, contributing to solve questions like: "Are we full or half brothers?"

### 3. FROM KINSHIP ANALYSIS TO DATA BASE SEARCH

There are three perspectives to consider the identification of individuals by using a data base of DNA profiles. We hereafter give a brief description of the topic emphasizing the most important features.

- **Criminal investigations.** The trace of an Unknown is found in a place of interest (often a crime scene) and is compared with the DNA profiles contained in a data base of people somehow in touch with the judicial system. At its simplest level, the comparison is limited to find if one of the DB members matches the Unknown (more than one match is extremely unlikely). There was a fierce controversy on how to evaluate the LR in case that a match is found. Hyper simplifying, the matter was if the LR would be smaller (Stockmarr, 1999) or bigger (Dawid, 2001) compared with the LR of type (1) obtained by not considering as evidence all the other no-match results. Now is largely acknowledged the contribution of the no-matching evidence and we find it correct. In some countries, and for some specific investigative purposes, the data base search was enlarged through the introduction of identification hypotheses involving the relatives of the DB members, see Cavallini and Corradi (2005) and Slooten and Meester (2014). Almost always, in this circumstance, it is not known if the members

of the DB have actually the relative/s for whom the search is established, so a further level of uncertainty must be introduced.

- **Search of missing persons.** The scheme is similar to the criminal identification DB search, but some simplifications arise. Since a biological sample of the missing person is rarely available, usually uncertainty on the missing person genetic profile is conditioned to familial donors' DNA evidence. In some cases, familial inference concentrates this probability in a spiky distribution, a limiting case is when two homozygous parents look for their missing sibling, which is equivalent, excluding mutations, to the availability of the sibling's DNA profile. Also the relationships among the familial donors and the missing person are faithfully elicited since relatives clearly understand the importance of the reliability of this information. Often the number of missing individuals in an area is uncertain, prone to be underestimated and overwhelms the number of claimed individuals, i.e. we do not have any familial information about several missing individuals.
- **Mass fatality incidents (MFI).** This is the most challenging exercise. Several Unknowns, hundreds in air crash episodes, many more in case of natural disasters, must be identified. The difficulty is to perform the simultaneous identification of all the unknowns (the victims) considering all the possible ways in which they can be allocated among the claiming families which often require the identification of more than a missing person, see Corradi (2010) for more details.

#### 4. INFERENCE FOR THE SEARCH OF MISSING PERSONS

There are two main contributions to cope with the inference on identification hypotheses concerning a DB search extended to relatives of the DB members. One approach, (Cavallini and Corradi, 2005), solves the inference problem by using a Bayesian network, exploiting some reasonable conditional independence assumptions to get efficient computations. Another one, (Slooten and Meester, 2014), obtains the required inference analytically by combining the likelihood ratios derived by separate kinship analyses illustrated in Section 2, to obtain some posterior probabilities of interesting events and likelihood ratios. Since they both were originally formulated for criminal investigations, in Sections 4.1 and 4.2 we provide some adaptation to cope with the search of missing persons.

##### 4.1 THE BAYESIAN NETWORK APPROACH

Let  $N$  the number of missing persons in the area we are concerned and  $n$  the families,  $\mathcal{F} = \{f_1, \dots, f_n\}$ , providing their genetic evidence  $x_f = \{x_{f_1}, \dots, x_{f_n}\}$ .

$\mathcal{M} = \{M_1, \dots, M_n\}$  is the set of missing relatives whose unobserved genetic traits are probabilistically described by the collection of random variables  $X_M = \{X_{M_1}, \dots, X_{M_n}\}$ . The identification hypothesis  $H = \{1, \dots, n, r\}$ , alternatively considers the Unknown as the missing relative of one of the  $n$  families or, if  $H = r$  (*rest*) one of the  $N - n$  unclaimed missing individuals, i.e.  $U$  is regarded as a generic member of the unclaimed missing population. Let's introduce the following assumptions.

*Assumption 1.* For all individuals, their genetic characteristics, considered at different loci, are mutually independent given the population parameters.

*Assumption 2.*  $X_f \perp\!\!\!\perp H$ , i.e. familial relations and genetic traits of the families' members are independent of the identification hypothesis.

*Assumption 3.*  $x_U = x_{m_i} | H = i, i \in \{1, \dots, n\}$ . If the Unknown is identified as the missing member of  $i$ th family, the probability to observe his/her characteristics only depends on  $X_{m_i}$ .

*Assumption 4.*  $p(x_U | H = r) = f(\theta)$ . If the unknown is assumed to be a generic member of the population, the probability to observe his/her characteristics only depends on the population parameters.

These assumptions justify the first graphical representation of the DB search missing persons problem in Figure 1 through its Directed Acyclic Graph (DAG). This representation is formally correct but largely inefficient since no conditional independence among families is revealed and the dimension of the conditional probability table of  $x_U$  increases with the number of claiming families, making the inference practically infeasible.

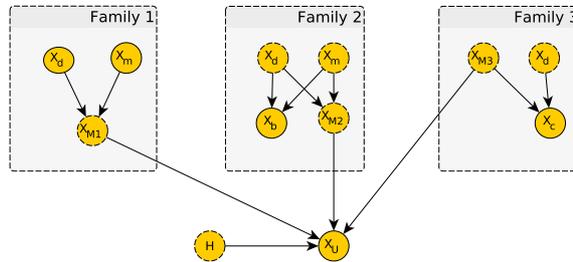
A more efficient but equivalent representation can be obtained by introducing a set of mediating variables  $Z = \{Z_1, \dots, Z_n\}$  between  $x_U$  and the  $X_{M_i}, i = 1, \dots, n$ , to induce these latter to *divorce*, that is the number of incident arcs converging to  $x_u$  are distributed to the  $z$ s.

$$p(x_U = (a_i, a_j) | z_i) = \begin{cases} 1, & \text{if } z_i = (a_i, a_j), \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

The  $Z$ s are simply a copy of the  $x_U$  so that this latter can be removed if the  $Z_i$  are instantiated with the genetic traits of the Unknown in each family. Furthermore, by *Assumption 2*, this means that computation can be performed separately in each family, conditionally on  $H$ .

Then, to immediately evaluate the probability the Unknown is the missing person in each of the  $n$  family, versus he/she is someone else, we introduce a set of  $n$  propositions,  $H_i = \{i, \bar{i}\}$ , deterministically related to  $H$  through

$$p(H_i = i | H) = \begin{cases} 1, & \text{if } H = i, \\ 0 & \text{otherwise.} \end{cases}$$

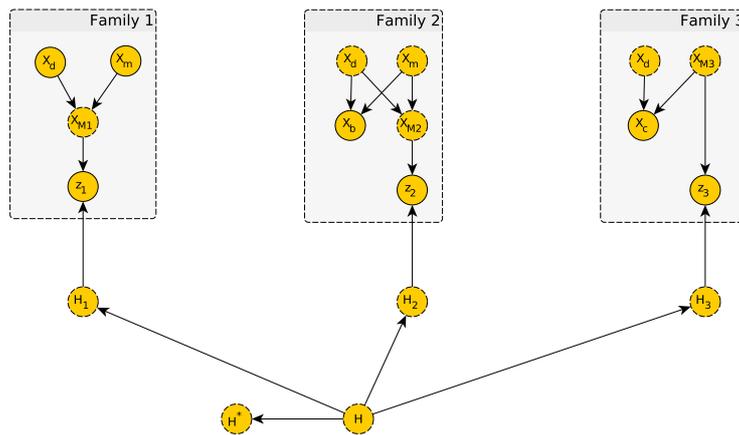


**Figure 1: DAG representation of a DB search problem according to Assumptions 1-4 for the case detailed in Section 4.3. Family 1: two parents ( $d,m$ ) are looking for their missing son,  $M_1$ . Family 2: a man ( $b$ ) is looking for his missing brother  $M_2$ . Family 3: a man ( $C$ ) is looking for his missing mother  $M_3$ . Continuous and dotted lines respectively indicate observed and unobserved genotypes.**

Finally, to compare directly the overall probability that the Unknown is one of the missing individuals searched by the families in the DB, versus he/she cannot be identified at a certain state of knowledge, we introduce a new hypothesis random variable  $H^* = \{DB, r\}$ , summarizing the results obtained by  $H$ , according to the following deterministic relation

$$p(H^* = DB|H) = \begin{cases} 1, & \text{if } H = i, \dots, n, \\ 0 & \text{if } H = r. \end{cases}$$

The manipulations proposed by adding the  $Z$ s, removing  $x_U$  and adding the  $H_i$ s and  $H^*$  are illustrated in Figure 2.



**Figure 2: DAG representation of a DB search problem augmented by variables  $Z$ s, having removed  $x_U$  and after the introduction of the families' detailed identification variables ( $H_i$ ) and the identification variable  $H$  summarizing both the hypotheses the Unknown is identified by the families in the DB or not.**

**Remark 3:** The  $H_i$ s and  $H^*$  are not essential to solve the inferential problem but, since dichotomous, they allow to simply derive meaningful likelihood ratios by asking to the net the probability of the evidence after the instantiation of each of their two states. By  $H_i$  we can derive the support provided by the data to the hypothesis that the Unknown is the missing individual searched by the  $i$ th family, versus he/she is some other claimed or unclaimed missing person. By  $H^*$  we easily derive the likelihood ratio supporting the hypothesis the Unknown is one of the people searched by the families in  $\mathcal{F}$  versus he/she is one of the unclaimed missing persons. Figures 1 and 2 show genotype networks which are not very suitable for computational purposes. Actually computations are performed through allele networks, which are definitely more efficient as demonstrated by Lauritzen and Sheehan (2003).

## 4.2 LIKELIHOOD RATIOS APPROACH

This approach, due to Slooten and Meester (2014), considers LRs defined in (1) and evaluated for each of the families, as building blocks to analytically derive some intriguing likelihood ratios and some relevant events' posterior probability. The possibility to use simple LRs to work with identification hypotheses related to a DB search has been introduced in the literature by Chung et al. (2010), for crime samples in form of mixtures, and by Corradi (2010) in case of MFI.

We now provide a summary of the main results about the evaluation of some events of interest, obtained

- a) analytically, according to Slooten and Meester (2014);
- b) using the BN represented in Figure 2, due to Cavallini and Corradi (2005), after instantiation with  $x_f$ ,  $z_i = x_U$  and evidence propagation.

Let  $\mathbf{r} = \{r_0, r_1, \dots, r_n\}$  be the likelihood ratios (1) evaluated for each of the families in  $\mathcal{F}$ , indicating with  $r_0 = 1$  the LR obtained if no familial evidence is available.

1. Posterior identification probability for each of the claimed missing persons:

a)

$$p(H_j = j | x_U, x_f) = \frac{r_j p(H_j = j)}{\sum_{i=0}^n r_i p(H_i = i)}. \quad (3)$$

b) Directly provided by the  $H_i$  nodes.

2. Likelihood ratio supporting the hypothesis  $U \in \mathcal{M}$  versus  $U$  is in the *rest* of the missing population.

a)

$$LR(U \in \mathcal{M}) = \frac{p(x_U, x_f | U \in \mathcal{M})}{p(x_U, x_f | U \notin \mathcal{M})} = \frac{\sum_{i=1}^n r_i p(H_i = i)}{p(H = r)}. \quad (4)$$

b) First compute the ratio of the probability of the overall evidence after having instantiated  $H^*$  to DB and to  $r$ , respectively. The result must be multiplied by  $p(H^* = r)/P(H^* = DB)$ .

3. Likelihood ratio supporting the hypothesis  $U$  is  $M_j$  versus  $U$  is one of the other claimed or unclaimed missing individuals.

a)

$$\begin{aligned} LR(U = M_j) &= \frac{p(x_U, x_f | U = M_j)}{p(x_U, x_f | U \neq M_j)} \\ &= \frac{r_j(1 - p(H_j = j))}{\sum_{i=1, i \neq j}^n r_i p(H = i) + p(H = r)}. \end{aligned} \quad (5)$$

b) First compute the ratio of the probability of evidence after having instantiated  $H_j$  to  $j$  and  $\bar{j}$ , respectively. The result must be multiplied by  $p(H_j = \bar{j})/P(H_j = j)$ .

**Remark 4:** The coincidence of the results provided by the considered approaches is reassuring. Notwithstanding the following question arises: “Is it better to use the conventional analytic derivations or to obtain results by a Bayesian Network?”. Since both the likelihood ratios  $\mathbf{r}$  and the expressions (3)-(5) are very easy to evaluate, the analytic approach seems to address the problem in a very simple way. However, when mixed populations, silent alleles, mutations, or the uncertainty on alleles probabilities are included in the evaluation, the BN approach seems more suitable to cope with all these complicated features as testified by recent literature addressing these topics. Historically, in forensic science, as noted by Taroni et al. (2014), analytical solutions have anticipated the representation of the same problem through Bayesian Networks. For the data base search problem, curiously, the BN solution was provided about 10 years before the analytic answer.

### 4.3 EXAMPLE

Now we consider in details a toy example of the DB search as represented in Figures 1 and 2.

*In the area in which we are operating,  $N = 100$  people were declared missing. Three families,  $n = 3$ , are looking for their missing relatives.*

- Family 1: two parents,  $d$  and  $m$ , are looking for their missing son,  $M_1$ .
- Family 2: a man,  $b$ , is looking for his missing brother,  $M_2$ .
- Family 3: a child,  $c$ , is looking for his missing mother,  $M_3$ .

We consider a single diallelic locus which exhibits allele  $A$  with probability 0.1 and  $B$  with probability 0.90. No other clue is available so that  $p(H_i) = \frac{1}{100}$ ,  $\forall i$ ,  $p(H = r) = 0.97$ . The body of an Unknown is recovered. We are extremely lucky because  $x_U = \{A, A\}$  i.e.  $U$  is homozygous for the less common allele  $A$ .

Data about familial donors,  $x_f$  and the evaluation of the probabilities of some relevant events and  $LR$ , as illustrated in Section 4.2, are in Table 1.

**Table 1: Data about familial donors,  $x_f$  and some results of the data base search based on the families described in Section (4.3) and illustrated in Figures 1 and 2.**

$F$	Missing	$x_f$	$p(H = i x_U, x_f)$	$LR(U = M_i)$
$f_1$	$M_1 = \text{son}$	$x_d = x_m = AA$	0.492	96.022
$f_2$	$M_2 = \text{brother}$	$x_b = AB$	0.017	1.663
$f_3$	$M_3 = \text{mother}$	$x_c = AB$	0.014	1.390

**Remark 5:** Apparently, results in Table 1 are not easy to interpret: the likelihood ratio favouring the identification of  $U$  as the missing person  $M_1$  is high (96.022) but the posterior probability of identifying  $U$  as  $M_1$  is less than 50%. At the same time the value of the likelihood ratio favouring the presence of  $U$  among the missing persons claimed by the families, computed using (4), amounts to 32, i.e. is much smaller than  $LR(U = M_1) = 96.022$ . A first answer is that, whereas the genetic evidence is very favourable to identify  $U$  as  $M_1$  ( $M_1$  is completely determined by the homozygous relatives resulting himself homozygous for the rare allele  $A$  exactly as  $U$ ) the large fraction of missing individuals for which we do not have any familiar information (97%) makes the probability of identification smaller than 50%. This figure is not so large and is due to the competing Families 2 and 3, each one exhibit the rare allele  $A$  so increasing the probability of  $U$  to be  $M_2$  or  $M_3$ . Of course the question is: “Do we have to believe in the strong support provided by the likelihood ratios or do we have to follow indications coming from the posterior probabilities of identification?”. In either way we need to formalize a rule to achieve to a short list of candidates  $\mathcal{D}$ , spanning from  $\emptyset$  to  $\mathcal{M}$ , to be verified by means of other clues and further efforts.

## 5. SEARCH THE MISSING INDIVIDUALS: THE SHORT LIST

There are only few contributions on what to do after the inference based on a data base search has been obtained.

Gittelsohn et al. (2012) considered the problem for criminal investigations. There, the most favourable evidence consists in a single match between the crime sample and the profile of one of the database members. In such circumstance the authors formally integrate a correct representation of the probabilistic part of the DB with the decision of identifying the only DB member matching the crime sample, taking into account the consequence of every action.

Also Slooten and Meester (2014) considered the problem of restricting the list of candidates to identification for criminal investigation purposes and extend the search, if no match were obtained, to some relatives of the DB members. They proposed two methods. That one, named *Profile-centred*, cannot be reformulated for the search of missing persons since it requires each missing person to be in the same relation with the relative providing the DNA evidence. The other, called *Conditional method*, can be extended to the search of missing individuals since it is based on a vector  $\mathbf{r}$  of LR as in (1), evaluated for each family according to the specific pedigree. From this setting they define as optimal a short list  $\mathcal{D}^k$  formed by the smallest set of size  $k$  of missing individuals with the highest product  $r_j p(H = j)$ , such that

$$\sum_{j=1}^k r_j p(H = j) \geq \alpha \sum_{i=1}^n r_i p(H = i).$$

The set  $\mathcal{D}^k$  is considered optimal by the authors because, if  $U \subset \mathcal{M}$  then, obviously,  $\mathcal{D}^k$  has probability to identify the Unknown equal to or greater than  $\alpha$ . The *property* obviously depends on the assumption that  $U$  is among the  $n$  searched missing persons, and, in our opinion, the assumption is likely to be overlooked.

Example (cont'd) Consider again the example of Section 4.3 and assume  $\alpha = 0.95$ . The optimal set is  $\mathcal{D}^2 = \{1, 2\}$  but the probability that the Unknown is one of the indicated missing individuals is actually 0.509 and not greater than 0.95, as a superficial interpretation of the conditional method would suggest.

Boreale and Corradi (2016), stemming from the tradition of the Bayesian decision theory, recently dealt with the problem of an optimal search in a general setting implying a reward, if the search is successful, and a fixed cost to be sustained for each item indicated as the possible secret. This approach seems suitable to be applied to the identification issue. In the case at hand the secret is the missing person providing the correct identification of the Unknown. Since after inference, many of missing persons claimed by the families might increase the probability to be the Unknown: if further verifications were cost-free, there should

be no reason to disregard some of them from a list of interesting candidates. But verifications are painful and expensive and this cost has to be taken into account. This justifies the creation of a short list.

Consider the introduction of a reward  $a$  if the Unknown is identified as one of the missing persons claimed by the families, and a fixed cost for each missing person included in the short list, accounting for the efforts required to verify the hint. If the short list is empty, no action is undertaken and both terms equal zero. The reward measures the degree of interest in the Unknown identification and can be conveniently expressed in cost units, i.e the identification is considered affordable if it is achieved in at most  $a$  attempts.

Let  $\mathcal{D}$  be a subset of  $H \setminus r$ . The net gain to explore  $\mathcal{D}$ , if the correct identification of the Unknown is  $H = i$ , is

$$g(H = i, \mathcal{D}) = a \cdot 1_{[H=i \in \mathcal{D}]} - |\mathcal{D}|.$$

For any set  $\mathcal{D}$ , the expected gain, after learning on  $H$  is established by

$$\begin{aligned} G(H, \mathcal{D}) &= \sum_{i=1}^n p(H = i|x_U, x_f) a \cdot 1_{[H=i \in \mathcal{D}]} - |\mathcal{D}| \\ &= \sum_{H=i \in \mathcal{D}} (p(H = i|x_U, x_f) a - 1). \end{aligned} \tag{6}$$

The Bayes action maximizing (6) is the set  $\mathcal{D}^*$

$$\mathcal{D}^* = \{i : p(H = i|x_U, x_f) \geq 1/a\}.$$

PROOF See Lemma 2 of Boreale and Corradi (2016).

If no other clue but  $x_U$  is available, it makes sense to pose

$$p(H = i) = \begin{cases} \frac{1}{N}, & \text{if } i = 1, \dots, n, \\ \frac{N-n}{N} & \text{if } H = r, \end{cases}$$

so that, by evaluating  $p(H = i|x_U, x_f)$  via Bayes theorem,  $\mathcal{D}^*$  becomes

$$\mathcal{D}^* \{ : p(x_U|x_f, H = i) \geq \frac{1}{a} (\sum_{i=1}^n p(x_U|x_f, H = i) + (N - n)p(x_U|H = r)) \}.$$

The expression makes clear that the inclusion of elements in  $\mathcal{D}^*$  is favoured by high values of  $a$ . On the other hand, if  $p(x_U|H = r)$  is high, i.e. the DNA traits of the Unknown are fairly common in the population, and/or if  $\frac{n}{N}$  is pretty small, those circumstances make it hard for a family to have the missing relative included in  $\mathcal{D}^*$ .

Example (cont'd). Consider again the example of Section 4.3. What is the optimal action after  $x_U$  is observed? By observing the last column on the right of Table 1,  $\mathcal{D}^*$  is evaluated according to some possible  $a$ 's values. Results are in Table 2.

**Table 2: Missing relatives included in  $\mathcal{D}^*$  to verify the results of the the DB search according to the example described in Section 4.3 for some different values of  $a$ .**

$a$	1	2	3	...	12	...	15
$\mathcal{D}^*$	$\emptyset$	$\emptyset$	1	...	1, 2	...	1, 2, 3

**Remark 6:** The inclusion in  $D^*$  of the extremely promising missing person in Family 1 is not straightforward now: he/she is included only if the worth of identification is, at least, as valuable as the cost of three verifications among families. The results takes into account the high  $LR(U = M_1)$  favouring the possibility that  $U$  is  $M_1$  but also considers the very small fraction (3%) of families asking for identification.  $M_2$  is not immediately included in the short list as it happens if the *Conditional method* is applied.  $M_2$  will be scrutinized only if the worth of identifying a missing person is considered at least 12 times the cost required to refine the search in a family.

## 6. CONCLUSIONS

In this work we have reconsidered and compared some different contributions to the identification problem using a data base of DNA profiles. Emphasis has been put on the missing person problem, tailoring to this specific topic the solution of the inference problem proposed by Cavallini and Corradi (2005) and Slooten and Meester (2014). We also reflected on the issue of restricting in the most promising missing persons worth further identification efforts. The solution stems from the Bayesian decision theory and following a recent contribution of Boreale and Corradi (2016). We suggested it as a solution for a simple structure of reward and costs related to the possible actions and consequences.

Next step will be to implement the proposal into a belief network, integrating the probabilistic and the decision sides of the approaches into a single coherent framework. Further efforts are also required to specify a suitable model to define prior probability of identification usually related to some case-specific observables as are the distance between where the Unknown was found and the place of the missing persons and other circumstances related to how the Unknown was recovered.

## REFERENCES

- Boreale, M. and Corradi, F. (2016). Searching secrets rationally. In *International Journal of Approximate Reasoning*. 69: 133-146.
- Cavallini, D. and Corradi, F. (2005). Forensic identification of relatives of individuals included in a database of DNA profiles. In *Biometrika*. 93(3): 525-536.
- Chung, Y.K., Hu, Y.Q. and Fung, W.K. (2010). Evaluation of DNA mixtures from database search. In *Biometrics*. 66: 233-238.
- Commissario straordinario del Governo per le persone scomparse (2015). *XIII Relazione semestrale*, Direzione Centrale della Polizia Criminale, Roma: 1-45.
- Corradi, F. (2010). Mass fatality incident identification based on nuclear DNA evidence. In *Journal of Machine Learning Research*. 9: 105-112.
- Dawid, A.P. (2001). Comment on Stockmar's Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search. In *Biometrics*. 57: 976-978.
- Dawid, A.P., Mortera, J., Pascali, V.L. and Van Boxel, D. (2002). Probabilistic expert systems for forensic inference from genetic markers. In *Scandinavian Journal of Statistics*. 29: 577-595.
- Dràbek, J. (2009). Validation of software for calculating the likelihood ratio for parentage and kinship. In *Forensic Science International: Genetics*. 3(2): 112-118.
- Egeland, T., Kulle, B. and Andreassen, R. (2006). Essen-Møller and identification based on DNA. In *Chance*. 19(2): 27-31.
- Gittelsohn, S., Biederman, A., Bozza, S. and Taroni, F. (2012). The database search problem: A question of rational decision making. In *Forensic Science International*. 222: 186-199.
- Lauritzen, S.L. and Sheehan, N.A. (2003). Graphical models for genetic analyses. In *Statistical Science*. 18: 489-514.
- Slooten, K. and Meester, R. (2014). Probabilistic strategies for familial DNA searching. In *Journal of the Royal Statistical Society C*. 63(3): 361-384.
- Stockmarr, A. (1999). Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search. In *Biometrics*. 55: 671-677.
- Taroni, F., Biederman, A., Bozza, S., Garbolino, P. and Aitken, C. (2014). *Bayesian Network for Probabilistic Inference and Decision Analysis in Forensic Science*. John Wiley & Sons.