Journal of Applied Mathematics

volume IV (2011), number III

PATERNITIES SEARCH WITH OBJECT-ORIENTED BAYESIAN NETWORKS

ANDRADE Marina, (PT), FERREIRA, Manuel Alberto M., (PT)

Abstract. Paternity dispute problems are examples of situations in which forensic approach the DNA profiles study is a common procedure. To implement this approach an efficient tool are the object-oriented Bayesian networks (OOBN). Along this paper are presented the various OOBN adequate to solve the simple paternity dispute and more complex paternity dispute problems with incomplete DNA profiles data about the putative father such as: only putative grandfather information, only putative uncle information, only putative father 's uncle information and only simultaneously putative uncle and putative father's uncle information. Here it is exhibited an algebraic treatment, for the simple problem and with those the use of the object-oriented Bayesian networks is shown. Then the most complex kind of problems that may occur is presented. Although these are not the most common cases there is notice of its occurrence at least in Portuguese courts.

Key words: Bayesian networks, DNA profiles, paternity dispute problems.

Mathematics Subject Classification: Primary 62C10; Secondary 62P99.

1 Introduction

The use of Bayesian networks in forensic identification problems has raised more and more attention, even for the social impact of these problems. It is usually recognized that the paternity dispute problems approach using Bayesian networks began with the works of Dawid et al. (2002) and Lauritzen (2003). In Andrade (2007) the use of this tool in paternity dispute and criminal cases is discussed. Some of the paternity dispute cases discussed here, although not the more frequent in courts, already occurred. And given its specificity justify the discussion and the use of Bayesian networks in the computation of a measure of the available evidence.

In the developed countries the application of the forensic identification statistics approach has grown significantly. The use of DNA evidence in forensic identification problems tries essentially

to look for answers to the logical and computational challenges that may occur in more complex situations such as, for instance, incomplete data.

The OOBN adequate to solve the simple paternity dispute is presented first jointly with the alternative algebraic treatment for checking purpose. Then the OOBN for the more complex paternity dispute problems with incomplete DNA profiles data about the putative father such as:

- only putative grandfather information,
- only putative uncle information,
- only putative father 's uncle information,
- only simultaneously putative uncle and putative father's uncle information

are shown. In these cases an algebraic treatment is out of question being the computational procedure imperative.

2 Simple Paternity Dispute

In a disputed paternity decision problem there are formally two challenging hypotheses (prosecution and defense):

H_P: The true father is the putative father.

vs

H_D: The true father is another individual randomly drawn from the population, and not genetically related with the mother or the putative father.

The court has to decide about the paternity of the child, and so, after Bayes' Law

$$\frac{P(H_P \mid E)}{P(H_D \mid E)} = \frac{P(E \mid H_P)}{P(E \mid H_D)} \times \frac{P(H_P)}{P(H_D)} \qquad (1),$$

with E the vector containing the available evidence, genetic information of the mother (*mgt*), of the child (*cgt*) and of the putative father (*pfgt*), being the algebraic approach simple.

It is needed to assess the likelihood function over the hypotheses as to the true father, i.e., to evaluate the likelihood ratio:

$$LR = \frac{P(E \mid H_P)}{P(E \mid H_D)} \qquad (2).$$

Naturally the court has to answer to the truly paternity of the child. So it has to evaluate the ratio of the hypotheses in dispute. Admitting that $P(H_P) = P(H_D)$ then (1) becomes

$$\frac{P(H_P \mid E)}{P(H_D \mid E)} = \frac{P(E \mid H_P)}{P(E \mid H_D)} \qquad (3).$$

In fact, knowing that the markers are in different chromosomes (*linkage equilibrium*) and assuming random mating (*Hardy-Weinberg equilibrium*) there is independence between and within markers. Thus, it is possible to obtain the *LR* for each marker separately and finally multiply the values to determine the overall likelihood ratio based on the data available for all markers.

To determine algebraically the probability of the triplet *E*, under the two hypotheses, it is reasonable to consider that before knowing any data on the child it is reasonable to assume that the identity of the true father is independent of the mother's and the putative father's. And supported on that, it is easily seen that it is possible to determine the conditional probability of the child's genotype, given the other two available genotypes. Thus, to determine $P(E \mid H_P)$ one has only to apply Mendel's laws. But the calculus of $P(E \mid H_D)$ necessarily demands the knowledge of the population allele frequencies for the considered markers.

If for a certain marker the triplet $E = (mgt, cgt, pfgt)$ is $E = ((A,B);(B,B);(A,B))$, and p_A and p_B are the population allele frequencies then

$$\begin{aligned} P(E \mid H_P) &= P[(mgt; cgt; pfgt) \mid (mgt; pfgt)] \\ &= P[cgt \mid (mgt; pfgt)] \\ &= 0.5 \times 0.5 \end{aligned}$$

and

$$\begin{aligned} P(E \mid H_D) &= P[(mgt; cgt; pfgt) \mid (mgt; rgt)] \\ &= P[cgt \mid (mgt; rgt)] \\ &= 0.5 \times p_B \end{aligned}$$

where *rgt* assigns the genotype of a random individual of the population, not related to the mother or the putative father.

Therefore,

$$LR = \frac{0.5}{p_B}.$$

The considered problem is, as shown, easily algebraically solved. It is used to illustrate the simplicity and the advantages of this tool in more complex situations. Given the freedom of choice for the variables to include in the graphical approach, different representations can be obtained. Some of them simpler than others. To get a 'good' representation is very important to the efficiency and the viability of the computational routines. These are extremely sensible to the organization of the graphical structure. The first step consists on the identification and definition of the nodes for all the variables of interest to the problem.

Then the graphical representation can be obtained. According to Dawid *et al*. (2002), *in order to maximize the efficiency of the calculations as well as the logical clarity of the representation we chose to disaggregate each individual's genotype into its constituent, unobserved, paternally and maternally inherited genes*.

Figure 1 exhibits the OOBN for a paternity case as the discussed above considering a single marker. Each node (instance) in the network represents itself a Bayesian network.

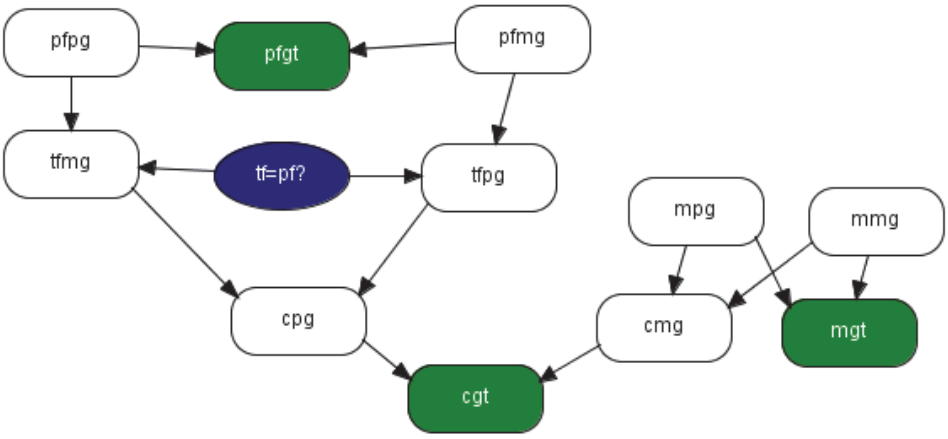

Figure 1: **Simple paternity network.**

In this simple paternity case instances **pfmg**, **pfpg**, **mpg** and **mmg** are all of class **founder**, a single node *gene,* having for its space of states all the possible alleles that can be presented for the specific case, and the correspondent population gene frequencies. Instances **mgt**, **cgt** and **pfgt** are of class **genotype**, an unordered pair of alleles inherited from paternal, *pg*, and maternal, *mg*, genes, here represented by *gtmin:=min*{*pg*, *mg*} and *gtmax:=max*{*pg*, *mg*}, where *pg* and *mg* are input nodes identical to the *gene* node of **founder**. Instances **tfmg** and **tfpg** are of class **whom**, describing the true father's allele origin. If *tf=pf?* has true for value then the true father's allele, *tfg*, will be identical with the putative father's, *pfg*, otherwise the true father's allele is randomly chosen from another man in the population. And **cpg** and **cmg** instances are of class **inherit**, modelling the Mendel's inheritance in which the child's allele is chosen at random from the two parents, *pg* and *mg*, here as the sequence of the observed outcome of a fair coin toss.

For illustration according to Dawid *et al*. (2002), the data for marker FES are child genotype *cgt* = {*B, B*}, mother's genotype *mgt* = {*A, B*} and putative father's genotype = {*A, B*}. The population allele frequencies are $p_A = 0.28425$ and $p_B = 0.25942$.

After specifying the network, put it to run and then insert the evidence. Considering equal prior probabilities for the query node representing the hypotheses, the likelihood is got after inserting the evidence. The likelihood ratio, based on the data for this marker, is obtained from the marginal posterior distribution of the query node. Thus, $P(tf = pf\,? := true \mid E) = 0.6584$ and $P(tf = pf\,? := false \mid E) = 0.3416$, and $LR = 1.9274$, being these results in agreement with the algebraic approach (note that $0.5/0.25942 \cong 1.9274$).

3 Paternities search in more uncommon situations

When the data E are not in the form $(mgt, cgt, pfgt)$ it is not possible to determine in algebraic form the likelihood function for the various hypotheses, i.e. to determine the weight of the genetic connection of the child with the putative father ancestor(s). The use of Bayesian networks allow to overcome these problems. These networks are a good tool to compute the likelihood functions.

Forwarding and backwarding the information a measure of the "strength" of the information available in each case is obtained.

In the sequence the networks for the uncommon cases described in the introduction are presented each one together with a numerical example.

The data considered are the same for the whole cases and are in Table 1 where five different markers are considered and the respective genotypes for the mother, the child, the grandfather, the uncle and the grandfather brother, where * indicates rare alleles, and (*a*) signs alleles considered as good discriminate markers, with more than 10 alleles in each marker.

Table 1:

Marker	*mgt*	*cgt*	*gfgt*	*ungt*	*gfbgt*
D3S1358	16, 18	13*, 16	13*, 17	13*, 16	13*, 15
VWA	16, 17	13*, 16	13*, 16	16, 18	13*, 15
D16S539	11, 12	12, 12	9, 12	10, 12	12, 13
D8S1179	12, 13	13, 17*	14, 17*	14, 15	12, 17*
D21S11(a)	29, 31.2	29, 31.2	29, 31.2	28, 31.2	29, 30

Genetic profiles

In Table 2 the respective allelic frequencies are presented: p_i is the i allele frequency in the population.

Marker	**Frequencies**				
D3S1358	p_{13}	p_{15}	p_{16}	p_{17}	p_{18}
	0.0032	0.2611	0.2477	0.2065	0.1606
VWA	p_{13}	p_{15}	p_{16}	p_{17}	p_{18}
	0.0023	0.1216	0.2300	0.2649	0.1859
D16S539	p_{9}	p_{10}	p_{11}	p_{12}	p_{13}
	0.1431	0.0545	0.3009	0.2876	0.1654
D8S1179	p_{12}	p_{13}	p_{14}	p_{15}	p_{17}
	0.1351	0.3028	0.2178	0.1223	0.0031
D21S11(a)	p_{28}	p_{29}	p_{30}	$p_{31.2}$	
	0.1674	0.2136	0.2437	0.1138	

Table 2: **Allele frequencies**

The allelic frequencies used were collected in
www.uni.duesseldorf.de/WWW/MedFak/Serology/dna.htm
for Portugal (Azores and Madeira archipelagos not included).

4 Only putative grandfather information

Bayesian networks for more complex problems can be built out of the same fundamental local modules that we have already described for the simple paternity dispute problem, Dawid et al. (2002).

The object-oriented Bayesian network for the "only putative grandfather information" case is shown in Figure 2. Note, for example, the node gfgt (grandfather genotype) and the respective connections with the other nodes.

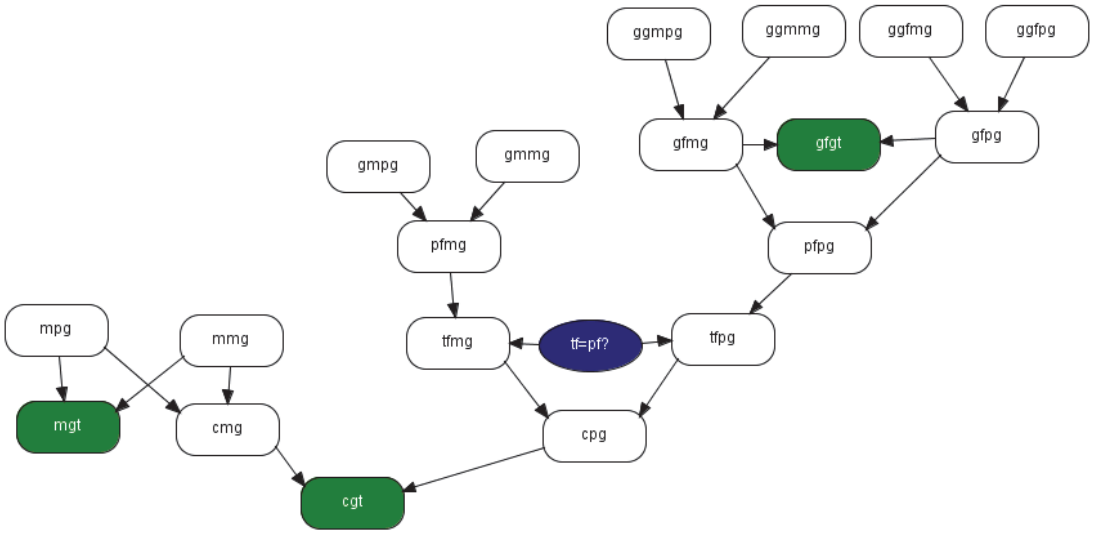

Figure 2**: Only putative grandfather network**

The results obtained are in Table 3. In the last column **Rescaled** – corrected so that the sum of the entries is equal to 1 – is presented the result for the 5 markers. Since the markers are independent the final result is obtained by multiplying the result obtained for each marker.

	D3S1358	**VWA**	**D16S539**	**D8S1179**	**D21S11**	**Rescaled**
$P(H_P\|E)$	0.9874	0.9909	0.5779	0.9878	0.6255	0,999999
$P(H_D\|E)$	0.0126	0.0091	0.4221	0.0122	0.3745	6,33E-07

Table 3: **Analysis results with only putative grandfather information**

5 Only putative uncle information

The object-oriented Bayesian network for the "only putative uncle information" case is shown in Figure 3.

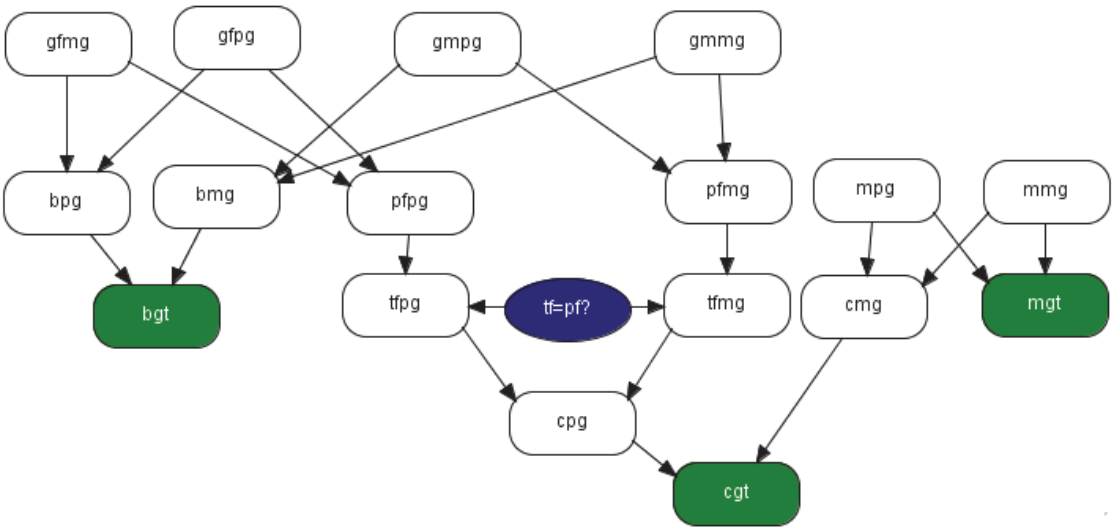

Figure 3**: Only putative uncle network**

The results obtained are in Table 4 following the same methodology as in section 5.

	D3S1358	**VWA**	**D16S539**	**D8S1179**	**D21S11**	**Rescaled**
$P(H_P\|E)$	0.9874	0.3333	0.5779	0.3333	0.5582	0,97133
$P(H_D\|E)$	0.0126	0.6667	0.4221	0.6667	0.4418	0,02867

Table 4: **Analysis results with only putative uncle information**

6 Only putative father ‘s uncle information

In Figure 4 the object-oriented Bayesian network for the “only putative father’s uncle information” case is shown.

It is a network more complex than the former ones owing to the further parentage relationship considered, that implies more complex genetic connections.

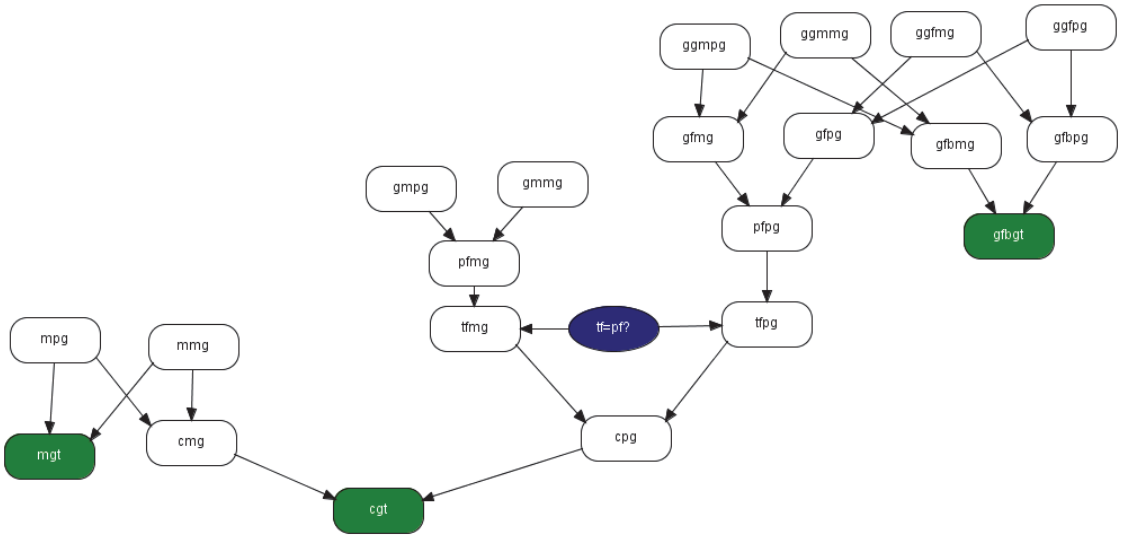

Figure 4**: Only putative father 's uncle network**

The results obtained are in Table 5.

	D3S1358	**VWA**	**D16S539**	**D8S1179**	**D21S11**	**Rescaled**
$P(H_P\|E)$	0.9755	0.9822	0.5423	0.9762	0.5309	0,999992
$P(H_D\|E)$	0.0245	0.0178	0.4577	0.0238	0.4691	8,28E-06

Table 5: **Analysis results with only putative father 's uncle information**

7 Only simultaneously putative uncle and putative father's uncle information

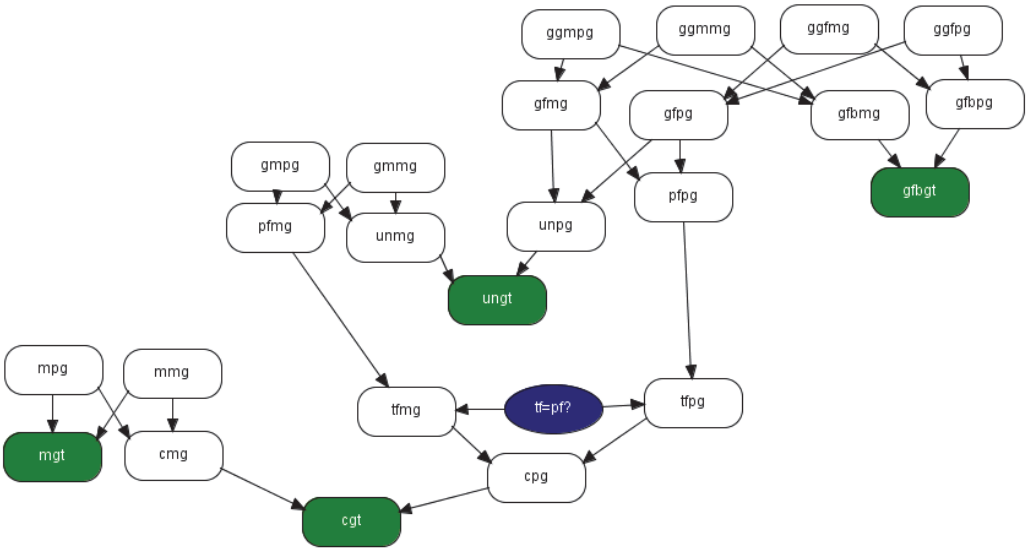

Figure 5**: Only simultaneously putative uncle and putative father's uncle**

The "only simultaneously putative uncle and putative father's uncle information" case network is the last one presented (in Figure 5) and the results are presented in Table 6.

	D3S1358	VWA	D16S539	D8S1179	D21S11	Rescaled
$P(H_P\|E)$	0.9875	0.9650	0.5764	0.9536	0.5707	0,999988
$P(H_D\|E)$	0.0125	0.0350	0.4236	0.0464	0.4293	1,23E-05

Table 6: **Analysis results with only simultaneously putative uncle and putative father's uncle information**

8 Conclusions

The paternities search in more uncommon cases demands the calculation of probabilities in the context of numerous and complex successive uses of Bayes Law. This situation is impossible to be treated algebraically. It was shown that the object-oriented Bayesian networks are a very powerful tool, very simple to use, that allows the referred calculations in an efficient way.

The major problem is to build the network taking in account the various and complex connections that may occur in parentage relationships. Then the use of an adequate software as Hugin or SPSS makes easy to apply it in practical cases. In this work Hugin was the chosen.

Inspecting the tables of results one can note that, as expected, rare alleles shared lead to greater probabilities of true paternity. On the contrary, more frequent alleles shared lead to lesser probabilities.

With the particular data used the final probabilities for true paternity were in general great.

References

[1] ABRANTES, D., PONTES, M. L., PINHEIRO, M. F., ANDRADE, M. and FERREIRA, M. A. M.: *Towards a systematic probabilistic evaluation of parentage casework in forensic genetics: A modest attempt to define a general standardized approach to simple and complex cases*. Forensic Science International: Genetics Supplement Series 1, pp. 635-637, 2008.

[2] ANDRADE, M.: *A Estatística Bayesiana na Identificação Forense – análise e avaliação de vestígios de DNA com redes Bayesianas*. PhD Thesis, ISCTE, Lisboa, 2007.

[3] ANDRADE, M.: *A Note on Foundations of Probability*. Journal of Mathematics and Technology, vol.. 1 (1), pp 96-98, 2010.

[4] ANDRADE, M., FERREIRA, M. A. M. and FILIPE, J. A.: *Evidence evaluation in DNA mixture traces*. Journal of Mathematics, Statistics and Allied Fields (Scientific Journals International-Published online), vol. 2 (2), 2008.

[5] ANDRADE, M., FERREIRA, M. A. M., FILIPE, J. A. and COELHO, M.: *Paternity dispute: is it important to be conservative?*. Aplimat – Journal of Applied Mathematics, vol. 1 (2), 2008.

[6] ANDRADE, M. and FERREIRA, M. A. M.: *Bayesian networks in forensic identification problems*. Aplimat - Journal of Applied Mathematics, vol. 2 (3), pp. 13-30, 2009.
[7] ANDRADE, M. and FERREIRA, M. A. M.: *Civil Identification Problems with Bayesian Networks Using Official DNA Databases*. Aplimat-Journal of Applied Mathematics, vol. 3 (3), pp. 155-162, 2010.
[8] ANDRADE, M. e FERREIRA, M. A. M.: *Solving civil identification cases with DNA profiles databases using Bayesian nerworks*. Journal of Mathematics and Technology, 1(2), pp. 37-40, 2010.
[9] ANDRADE, M. e FERREIRA, M. A. M.: *Evaluation of Paternities with less usual Data using Bayesian Networks*. *IEEE Xplore (BMEI 2010 IEEE Catalog Number CFP1093D-PRT, ISBN: 978-1-4244-6496-8)*, 2010.
[10] ANDRADE, M., FERREIRA, M. A. M.,ABRANTES, D., PONTES, M. L. e PINHEIRO, M. F.: *Object-oriented Bayesian Networks in the evaluation of paternities in less usual environments*. Journal of Mathematics and Technology, 1(1), pp. 161-164, 2010.
[11] DAWID, A. P., MORTERA, J., PASCALI, V. L. and van BOXEL, D. W.: *Probabilistic expert systems for forensic inference from genetic markers.* Scandinavian Journal of Statistics vol. 29, pp. 577-595, 2002.
[12] FERREIRA, M. A. M. and ANDRADE, M.: *A note on Dawnie Wolfe Steadman, Bradley J. Adams, and Lyle W. Konigsberg, Statistical Basis for Positive Identification in Forensic Anthropology. American Journal of Physical Anthropology 131: 15-26 (2006)*. International Journal of Academic Research, vol. 1 (2), pp. 23-26, 2009.
[13] LAURITZEN, S. L.: *Bayesian networks for forensic identification Problems*. Tutorial 19th Conference on Uncertainty in Artificial Intelligence, Mexico, 2003.

Current address

Marina Andrade, Professor Auxiliar
ISCTE – Lisbon University Institute
UNIDE - IUL
Av. Das Forças armadas
1649-026 Lisboa
Telefone: + 351 21 790 34 05
Fax: + 351 21 790 39 41
e-mail: marina.andrade@iscte.pt

Manuel Alberto M. Ferreira, Professor Catedrático
ISCTE – Lisbon University Institute
UNIDE - IUL
Av. Das Forças armadas
1649-026 Lisboa
telefone: + 351 21 790 37 03
fax: + 351 21 790 39 41
e-mail: manuel.ferreira@iscte.pt