

# Crime Scene Investigation with Bayesian Probabilistic Expert Systems

Marina Andrade\*, Manuel Alberto M. Ferreira

Department of Quantitative Methods, University Institute of Lisbon, Lisbon, 1649-026, Portugal

**Abstract** Criminal identification problems are examples of situations in which forensic approach the DNA profiles study is a common procedure. In order to deal with these problems it is needed an introduction to present and explain the various concepts involved, since distinct areas must be considered. Some problems are presented and the use of the object-oriented Bayesian networks, example of probabilistic expert systems, is shown.

**Keywords** Probabilistic Expert Systems, Bayesian Networks, DNA Profiles, Identification Problems

## 1. Introduction

The use of networks transporting probabilities began with the geneticist Sewall Wright in the beginning of the 20th century (1921). Since then their use had different forms in several areas like social sciences and economy – in which the used models are, in general, linear named Path Diagrams or Structural Equations Models (SEM), and in artificial intelligence – usually non-linear models named Bayesian networks also called Probabilistic Expert Systems (PES).

*Bayesian networks are graphical structures for representing the probabilistic relationships among a large number of variables and for doing probabilistic inference with those variables*, [1]. Before approaching the use of Bayesian networks to the interest problems some aspects of PES in connection with uncertainty problems must be studied, see for instance [2].

## 2. Objectives

A crime has been committed, and two persons were murdered,  $V_1$  and  $V_2$ . At the crime scene two different mixture traces were found:  $T_1$  in the toilet and  $T_2$  in the victims' car.  $S_2$  is a potential suspect.  $S_2$ 's DNA profile was measured and found to be compatible with the mixture traces.

Accepting that there was a fight during the assault that produced some material, it is obvious that the individual who perpetrated the crime could left some of his/her material in some but not in the whole traces. The non-DNA evidence indicates the possibility that two people were involved in the crime.

In [3] it is described a new approach to the problems mentioned in 1. The construction and use of Bayesian networks to analyse complex problems of forensic identification inference was initially done there followed by [4-6,7] among others.

The advances achieved in the forensic biology have certainly encouraged the interest in problems of forensic identification, also allowing a much more rigorous treatment of the problems in analysis. That is the case of problems of DNA mixtures – [6,7].

One of the complexities in the interpretation of the mixture traces is assigning the number of contributors to the mixture. In general, the trace suggests a lower bound for the total number of contributors but no upper bound. [5] gave a useful low upper bound on the number of contributors worth considering.

In what follows it is described a complex mixture case, related with the former crime scene, and presented the data to be considered in the analysis. After formulating the hypotheses the analysis is performed for one marker considering the information from one trace. Then the two traces are considered and finally the analysis is generalized considering two mixture traces and the three markers.

## 3. Methods

To summarize the evidence is presented in Table 1 the DNA profiles of the victims' and the suspect,  $S_2$ . In Table 2 the profiling results for the mixtures traces ( $T_1$  and  $T_2$ ), for the STR markers studied, respectively, and the allele frequencies for each marker are presented.

The traces contain biological material that must belong to some person other than the two victims.

The allele frequencies used in this work are the Portuguese population frequencies collected in the worldwide database "The Distribution of Human DNA-PCR Polymorphisms",

\* Corresponding author:  
marina.andrade@iscte.pt (Marina Andrade)  
Published online at <http://journal.sapub.org/am>  
Copyright © 2012 Scientific & Academic Publishing. All Rights Reserved

since the case mentioned took place in Portugal.

**Table 1.** Two victims and suspect DNA profiles

| Marker | $V1$ (f) | $V2$ (m) | $S2$  |
|--------|----------|----------|-------|
| TH01   | $D;E$    | $D;E$    | $B;C$ |
| FES    | $A;C$    | $C;C$    | $B;B$ |
| FGA    | $B;E$    | $B;C$    | $A;C$ |

It is natural to consider that the crime traces can contain DNA from up to three unknown contributors, in addition to the victims and/or the suspect.

**Table 2.** DNA mixture traces and allele frequencies<sup>1</sup>

|       | TH01         | FES       | FGA          |
|-------|--------------|-----------|--------------|
| $T1$  | $B; C; D; E$ | $A; B; C$ | $A; B; C; E$ |
| $T2$  | $B; C; D; E$ | $B; C$    | $A; B; C$    |
| $p_A$ | *            | 0.0129    | 0.0684       |
| $p_B$ | 0.1696       | 0.3287    | 0.1740       |
| $p_C$ | 0.1386       | 0.3664    | 0.1606       |
| $p_D$ | 0.1984       | *         | *            |
| $p_E$ | 0.2748       | *         | 0.0321       |

If the DNA from  $S_2$  is present in at least one of the traces this will place him at the crime scene and consequently as one of the possible perpetrators. Consideration of whether or not the suspect was a contributor to any of the mixture traces will give a measure of the evidence strength.

### Hypotheses

The court has to determine if the suspect is or is not guilty. These are described as the level III, or offence, propositions. However the forensic scientist does not typically address such propositions. In this case it appears more appropriate to address source level propositions.

Hypotheses to be addressed:

$H_1$ :  $S_2$  is one of the contributors to  $T_1$  but not  $T_2$ .

$H_2$ :  $S_2$  is one of the contributors to  $T_2$  but not  $T_1$ .

$H_3$ :  $S_2$  is one of the contributors to both  $T_1$  and  $T_2$ .

$H_4$ :  $S_2$  did not contribute to trace  $T_1$  or  $T_2$ .

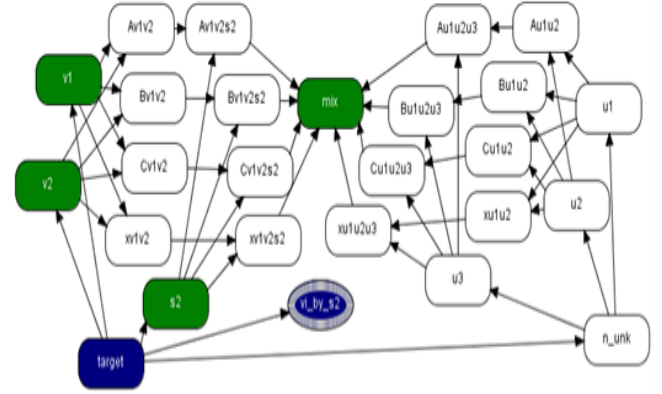
What interests is to measure:

$P(S_2 \text{ contr. to at least one of the traces} | \xi)$  where  $\xi$  is the vector comprising the profiles observed of the traces found at the crime scene: the victims' and the suspect profiles. This is equivalent to

$$P(H_1 \cap H_2 \cap H_3 | \xi) = 1 - P(H_4 | \xi).$$

### One mixture trace and a single marker

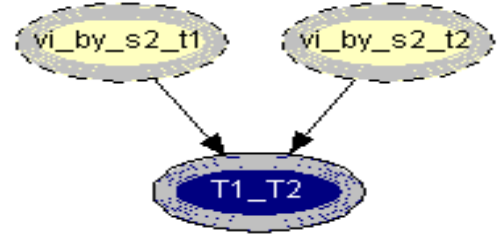
The network for one trace and a single marker follows[7], Figure 4 section 3.2, an OOBN version considering up to three unknown contributors Figure 1, marker network. Here it is presented the network for the marker, FES<sup>2</sup>.



**Figure 1.** Marker network

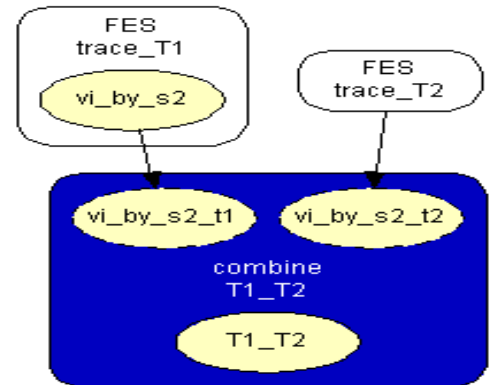
### Two mixture traces and a single marker

As described above there were two different traces at the crime scene. So it is necessary to combine the information from both traces. To do so define an instance **combine**, Figure 2. This instance has as parents the output nodes  $vi\_by\_s2$  of the instance **marker** for trace  $T_1$  and trace  $T_2$ . The node  $T_1\_T_2$  combines the results obtained in the parent instances for node  $vi\_by\_s2$  expressing the result values of the one-to-one correspondence with the eight joint configurations of its parents nodes for the considered marker.



**Figure 2.** Combine network

Therefore, the node  $T_1\_T_2$  takes values 0, 1, 2, 3 corresponding to the hypothesis  $H_4$ ,  $H_1$ ,  $H_2$  and  $H_3$ , respectively.  $T_1\_T_2$  is 0 if  $vi\_by\_s2$  is less than 4 in  $T_1$  and  $T_2$ ; assumes value 1 if  $vi\_by\_s2$  is equal to 4 or more in  $T_1$  and less than 4 in  $T_2$ ; takes value 2 if  $vi\_by\_s2$  is less than 4 in  $T_1$  and equal to 4 or more in  $T_2$ ; and is 3 if  $vi\_by\_s2$  is equal to 4 or more in both  $T_1$  and  $T_2$ . In the start a uniform prior distribution for node  $T_1\_T_2$  is assumed.



**Figure 3.** Combine\_  $T_1\_T_2$  network

<sup>1</sup> The use of \* refers values that are of no concern in the analysis.

<sup>2</sup>The marker networks differ only in the number of alleles to consider, whether it is the space of states of the nodes referring the alleles or in the presence of one more allele to consider in the network. Since Hugin software does not allow modification of the state of a node in order to reuse a network, for markers TH01 and FGA a codification in the space of states of the node *gene* was performed and put it in accordance with the alleles of each marker under consideration so that it could used the same network.

Now it is possible to put the networks for each trace together and compute the interest information, Figure 3. The instances **FES trace\_t1** and **FES trace\_t2** are of class **marker** in which all the individuals in any of the networks have the same structure (**individual**). Its differentiation is made when the evidence is inserted.

## 4. Results

When combining the two traces, in order to obtain a measure of the evidential weight associated with the possible presence of genetical material from the suspect in the traces found at the crime scene, the results listed in the Tables below are got. For marker FES with different mixture traces it is obtained:

**Table 3.** Results of the node  $vi\_by\_s2$

| $S_2, V_2, V_1$ | trace $T_1$ | trace $T_2$ |
|-----------------|-------------|-------------|
| 0 (FFF)         | 0.0048      | 0.1470      |
| 1 (FFT)         | 0.1334      | 0.0000      |
| 2 (FTF)         | 0.0068      | 0.1791      |
| 3 (FTT)         | 0.1334      | 0.0000      |
| 4 (TFF)         | 0.0072      | 0.1881      |
| 5 (TFT)         | 0.3526      | 0.0000      |
| 6 (TTF)         | 0.0092      | 0.4857      |
| 7 (TTT)         | 0.3526      | 0.0000      |

Where the state 0 corresponds to  $s2\_in\_mix? = False$ ,  $v2\_in\_mix? = False$  and  $v1\_in\_mix? = False$  (FFF), and for simplicity the state 0 is read as  $S_2; V_2; V_1 = FFF$ .

In Table 4 it is shown the combined information for the two traces for marker FES.

**Table 4.** Results for the node  $T1\_T2$

|       |        |
|-------|--------|
| $H_1$ | 0.2353 |
| $H_2$ | 0.1876 |
| $H_3$ | 0.4862 |
| $H_4$ | 0.0908 |

Thus,

$$P(S_2 \text{ contr. to at least one of the traces} | \xi) \cong 0.91.$$

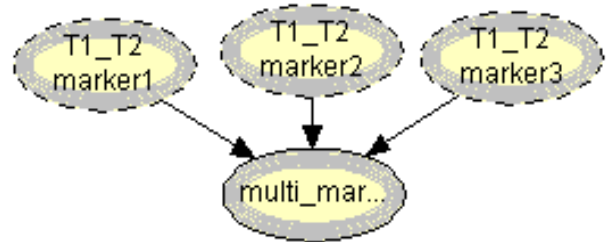
### Generalization: two mixture traces and three markers

Given the results obtained for one marker it is necessary to extend the reasoning in order to consider the information for the three markers, FES, TH01 and FGA.

The instances **combine  $T_1\_T_2$**  express the results for each marker accounting for the information for the two traces. The node  $T_1\_T_2$  in each of these instances computes the results for each marker. The respective tables, similar to Table 4, can be extracted for the other two markers.

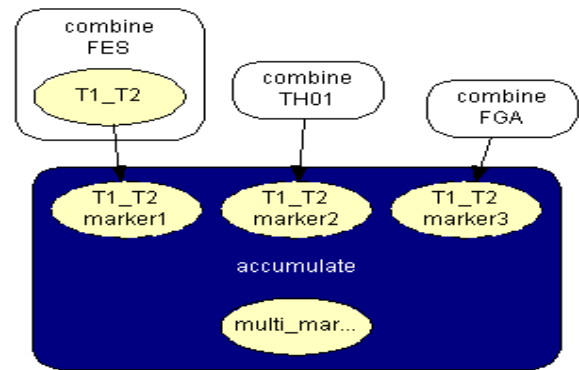
The instance **accumulate** having as inputs the output nodes of the instances **combine  $T_1\_T_2$** , with the results of each marker, incorporates the information for the two traces obtained separately, Figure 4. The node **multi\_markers** combines the information from the different instances **combine  $T_1\_T_2$** , i.e., **multi\_markers** gives the results synthesizing the results of  $T_1\_T_2$  for the three markers. The node **multi\_markers** with states 0, 1, 2 and 3 assumes the state 0 if

all the input nodes are 0. Takes value 1 if all the input nodes are 1 or at least one of the input nodes has state 1 and the others have the state 0<sup>3</sup>. The node **multi\_markers** is 2 if all the input nodes have state 2 or this state 2 is combined between the states 0 and 2 of the input nodes. The node assumes state 3 if all the input nodes have state 3 or if the inputs are combining state 0, state 1 and state 2.



**Figure 4.** Accumulate network

Joining the networks for the three markers, each of which accounts for the two traces, it is obtained the **accumulate\_three\_markers** network, Figure 5.



**Figure 5.** Accumulate three markers network

Tables 5 and 6 display the results for the marker FGA and TH01 and the cumulative result for all three markers, re-scaled to sum up to 1. This aims at the question of interest.

**Table 5.** Results for the eight configurations for markers FGA and TH01

| $S_2, V_2, V_1$ | trace $T_1$ | trace $T_2$ | trace $T_1$ | trace $T_2$ |
|-----------------|-------------|-------------|-------------|-------------|
| 0 (FFF)         | 0.0010      | 0.0084      | 0.0134      | 0.0134      |
| 1 (FFT)         | 0.0150      | 0.0000      | 0.0342      | 0.0342      |
| 2 (FTF)         | 0.0037      | 0.0476      | 0.0342      | 0.0342      |
| 3 (FTT)         | 0.0290      | 0.0000      | 0.0342      | 0.0342      |
| 4 (TFF)         | 0.0079      | 0.0977      | 0.0599      | 0.0599      |
| 5 (TFT)         | 0.4644      | 0.0000      | 0.2748      | 0.2748      |
| 6 (TTF)         | 0.0146      | 0.8463      | 0.2748      | 0.2748      |
| 7 (TTT)         | 0.4644      | 0.0000      | 0.2748      | 0.2748      |

**Table 6.** Results for the node  $T1\_T2$  for markers FGA and TH01

<sup>3</sup> e.g., **multi\_markers**=1 if  $T_1\_T_2 = 1$  for **marker1**, **marker2** and **marker3**; or  $T_1\_T_2 = 1$  for **marker1** and **marker2** and  $T_1\_T_2 = 0$  for **marker3**; or  $T_1\_T_2 = 1$  for **marker1** and **marker3** and  $T_1\_T_2 = 0$  for **marker2**; or  $T_1\_T_2 = 1$  for **marker2** and **marker3** and  $T_1\_T_2 = 0$  for **marker1**; or  $T_1\_T_2 = 1$  for **marker1** and  $T_1\_T_2 = 0$  for **marker2** and **marker3**; or  $T_1\_T_2 = 1$  for **marker2** and  $T_1\_T_2 = 0$  for **marker1** and **marker3**; or  $T_1\_T_2 = 1$  for **marker3** and  $T_1\_T_2 = 0$  for **marker1** and **marker2**.

|       |          |
|-------|----------|
| $H_1$ | 0.002114 |
| $H_2$ | 0.001568 |
| $H_3$ | 0.996313 |
| $H_4$ | 0.000003 |

Therefore,

$$P(S_{2\text{contr. to at least one of the traces}} | \xi) \cong 0.9999$$

## 5. Discussion

When the whole information for the two traces on the three markers is taken into account a very significant value for the interest quantity is obtained.

The use of DNA evidence analysis is commonly accepted nowadays in the whole courts. However, the presentation, interpretation and evaluation of this type of evidence sometimes raise some problems. And it is far the day when a total incorporation of this kind of evidence is achieved, although in some cases it has been decisive for the conviction or absolution of the individuals. This is already a good support for justice.

## 6. Conclusions

The statistical treatment of criminal evidence has raised new challenges to those that have to decide, in the basis of the presented results. Independently of the methodology used, the great difficulty inhabits in the interpretation of the evidence, which is summarized in a number – what does that value means?

In the most complex problems, as the mentioned ones, the use of Bayesian networks for the analysis and interpretation of the evidence can be of great help. In a Bayesian network the complex inter-relations between the variables are transformed into modular units.

This technology – which use is everyday more common in different areas – supplies, as a support to the decision, a number. It does not give the decision; it is a decision support instrument. Consequently it is important that the legal system knows how to evaluate and interpret correctly the information contained in it. However, there is still much to do.

## ACKNOWLEDGEMENTS

This work was financially supported by FCT through the Strategic Project PEst-OE/EGE/UI0315/2011.

## REFERENCES

- [1] R. E. Neapolitan, Learning Bayesian networks, Pearson Prentice Hall, 2004
- [2] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, D. J. Siegelhalter, Probabilistic expert systems, Springer, New York, 1999
- [3] A. P. Dawid, J. Mortera, V. L. Pascali and D. W. van Boxel, 2002, Probabilistic expert systems for forensic inference from genetic markers, Scandinavian Journal of Statistics, 29, 577-595
- [4] I. W. Evett, P. D. Gill, G. Jackson, J. Whitaker and C. Champod, 2002, Interpreting small quantities of DNA: the hierarchy of propositions and the use of Bayesian networks, Journal of Forensic Sciences, 47, 520-530
- [5] S. L. Lauritzen and J. Mortera, 2002, Bounding the number of contributors to mix DNA stains, Forensic Science International, 130, 125-126
- [6] J. Mortera, 2003, Analysis of DNA mixture using probabilistic expert systems, In: Green, P. J., Hjort, N. L., Richardson, S. (Eds.), Highly Structured Stochastic Systems. Oxford University Press
- [7] J. Mortera, A. P. Dawid and S. L. Lauritzen, 2003, Probabilistic expert systems for DNA mixture profiling, Theoretical Population Biology, 63, 191-205
- [8] M. Andrade, 2010, A note on foundations of probability, Journal of Mathematics and Technology, 1 (1), 96-98
- [9] M. Andrade and M. A. M. Ferreira, 2007, Mixture traces: comparison of several hypotheses, 56th Session of the ISI-ISI
- [10] M. Andrade and M. A. M. Ferreira, 2009, Criminal and Civil Identification with DNA Databases Using Bayesian Networks", International Journal of Security-CSC Journals, 3 (4), 65-74
- [11] M. Andrade and M. A. M. Ferreira, 2009, Bayesian networks in forensic identification problems, Aplimat-Journal of Applied Mathematics, 2 (3), 13-30
- [12] M. Andrade and M. A. M. Ferreira, 2010, Solving civil identification cases with DNA profiles databases using Bayesian networks", Journal of Mathematics and Technology, 1 (2), 37-40
- [13] M. Andrade and M. A. M. Ferreira, 2011, Some considerations about forensic DNA evidences", International Journal of Academic Research, 3 (1,Part I), 7-10
- [14] M. Andrade and M. A. M. Ferreira, 2011, Evidence evaluation in DNA mixture traces possibly resulting from two victims and two suspects, *PJQM-Portuguese Journal of Quantitative Methods*, 2 (1), 99-103
- [15] B. Budowle, J. Ge, R. Chakraborty and H. Gill-King, 2011, Use of prior odds for missing persons identifications, Investigative Genetics, 2 (1):15
- [16] M. A.M. Ferreira and M. Andrade, 2009, A note on Dawid Wolfe Steadman, Bradley J. Adams, and Lyle W. Konigsberg, "...", International Journal of Academic Research, 1 (2), 23-26