CIVIL IDENTIFICATION WITH DNA PROFILES DATABASES USING BAYESIAN NETWORKS

Marina Andrade[1], Manuel Alberto M. Ferreira[2]

[1]Instituto Universitário de Lisboa (ISCTE-IUL), Information Sciences, Technologies and Architecture Research Center (ISTAR-IUL) and Business Research Unit (BRU-IUL), ***PORTUGAL***
[2]Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL) and Information Sciences, Technologies and Architecture Research Center (ISTAR-IUL), ***PORTUGAL***

E-mails: marina.andrade@iscte.pt, manuel.ferreira@iscte.pt

ABSTRACT

In forensic identification problems it is common the use of DNA profiles. The first bases of DNA profiling data emerged in England in 1995 having raised new challenges in the context of forensic identification. In Portugal the legislation that allowed the creation of a DNA profile database was set in 2008. In this work it is intended to discuss how to properly use the databases in the recognition of bodies problem that compiles information about missing individuals which are known to belong to one or more families whose members have reported the disappearance of relatives.

Keywords: *DNA profiles, Identification problems, Bayesian networks*

INTRODUCTION

The use of networks that "carry" probabilities began with the geneticist Sewall Wright in 1921. Since then its use coated in different ways in various fields such as social sciences and economics. Here the models used are generally linear being examples the "Path Diagrams" or "Structural Equation Models (SEM)". In artificial intelligence are usually used nonlinear models called "Bayesian networks" also called "Probabilistic Expert Systems (PES)".

In this work the aim is to approach civil identification problems: the recognition of bodies which compiles information about missing individuals who are known to belong to one or more families whose members have reported the disappearance of family members using the DNA database and information about the genetic profile of the bodies found. In the processing of data, we resort to "object-oriented" Bayesian networks, which are an example of PES, using the Hugin software.

Portuguese law 5/2008 establishes the principles for creating and maintaining a database of DNA profiles for identification purposes. And regulates the collection, processing and storage of human cell samples, their collection and

analysis of DNA profiling, the method for comparing DNA profiles resulting samples, and processing and storage of information in a computer file.

It is assumed that the database consists of a file containing samples delinquent information doomed to 3 or more years in prison - α ;a file containing information volunteers samples - β ; a file containing information about "problem samples" or "reference samples" of bodies or parts of bodies or objects or places where authorities collected samples - γ .

Unfortunately, the project to create a DNA profiles database for identification purposes has been abandoned by the Portuguese Government.

CIVIL IDENTIFICATION USING DNA PROFILIES DATABASES-ONE MISSING AND A VOLUNTEER

It is reported the disappearance of an individual. A body is found. The hypotheses are: HA: The body found is that of the missing individual X vs HD: The body found is any other individual than that of X. In general, is not admitted the death of relatives and rejects the hypothesis that establishes the loss of the missing relative. A volunteer provides genetic material to be used to test a partial match. The evidence is - $E = (C_{Bf}, C_{vol})$ - the genetic trait of the body found, C_{BF} , and the volunteer, C_{vol} . The odds *a posteriori* (ratio of *a posteriori* probability) is

$$\frac{P(H_A \mid E, vol \in \beta, \gamma)}{P(H_D \mid E, vol \in \beta, \gamma)} = \frac{P(E \mid H_A, vol \in \beta, \gamma)}{P(E \mid H_D, vol \in \beta, \gamma)} \frac{P(H_A \mid vol \in \beta, \gamma)}{P(H_D \mid vol \in \beta, \gamma)}.$$

Assuming that $P(H_A \mid vol \in \beta, \gamma) = P(H_D \mid vol \in \beta, \gamma)$, an usual proceeding,

$$\frac{P(H_A \mid E, vol \in \beta, \gamma)}{P(H_D \mid E, vol \in \beta, \gamma)} = \underbrace{\frac{P(C_{BF}, C_{vol} \mid H_A, vol \in \beta, \gamma)}{P(C_{BF}, C_{vol} \mid H_D, vol \in \beta, \gamma)}}_{likelihood-ratio}$$

$$\frac{P(H_A \mid E, vol \in \beta, \gamma)}{P(H_D \mid E, vol \in \beta, \gamma)} = \frac{P(C_{BF} \mid C_{vol}, H_A, vol \in \beta, \gamma)}{P(C_{BF}, \mid C_{vol}, H_D, vol \in \beta, \gamma)} \underbrace{\frac{P(C_{vol} \mid H_A, vol \in \beta, \gamma)}{P(C_{vol} \mid H_D, vol \in \beta, \gamma)}}_{=1}.$$

Genetic voluntary feature is not in the conditioning events. Whether voluntary or not related to the individual whose body was found it does not provide information to the uncertainty about its genotype.

Depending on the volunteer and his family relationship with the missing person just can observe a partial match with a volunteer. Furthermore it is important to check for a match between C_{BF} and some of the "sample problem" in γ . Assuming no coincidence of C_{BF} and any γ sample, the likelihood ratio may be written as:

$$\frac{P(C_{BF} \mid C_{vol}, H_A)}{P(C_{BF} \mid H_D)}.$$

If only one volunteer is considered the likelihood ratio can be calculated using a Bayesian network. Suppose you have a missing person, an old man or woman, and a son or daughter who announces the disappearance and voluntarily provides its own genetic profile. The likelihood ratio can be calculated using the following chain:

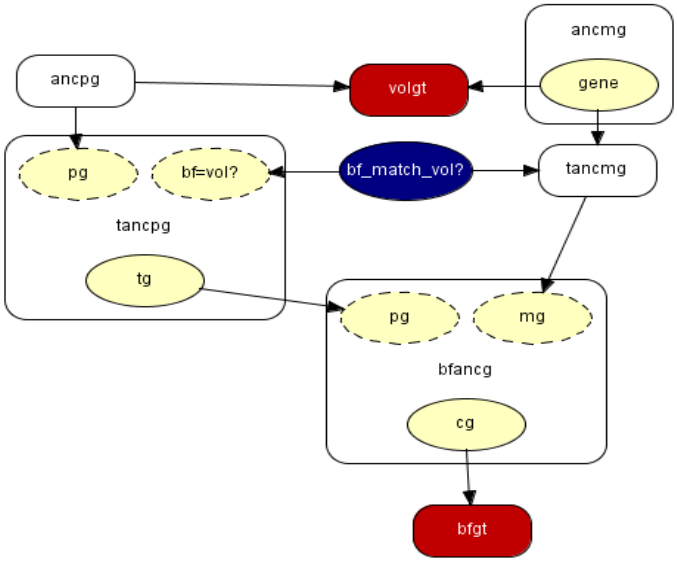

Figure 1. *Network for the civil identification with one volunteer, son or daughter*

For a case with a volunteer, but now for example a brother or sister of a missing person who is being sought, the likelihood ratio can be calculated using the following Bayesian network:

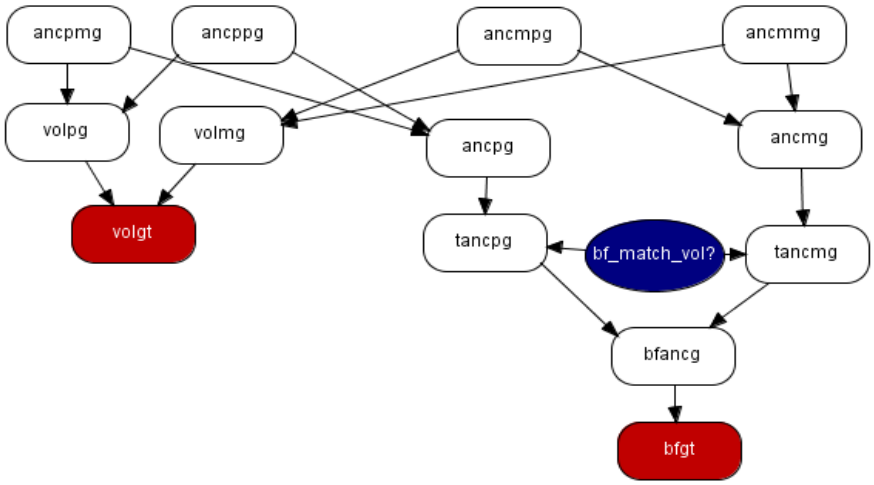

Figure 2. *Network for the civil identification with one volunteer, brother or sister*

The odds *a posteriori* is, as it was seen, the likelihood ratio, which allows for benchmarking between HA and HD.

In civil identification problem, the identification evidence may involve more than one individual. It is a case already considered, see for example (Andrade and Ferreira, 2010a).

REFERENCES

1) D. Balding (2002). The DNA Database Controversy. Biometrics, 58 (1), 241-244.
2) F. Corte-Real (2004). Forensic DNA Databases. Forensic Science International, S143-S144.
3) M. Andrade, M. A. M. Ferreira (2007). Analysis of a DNA mixture sample using object-oriented Bayesian networks. 6th International Conference APLIMAT 2007; Bratislava; Slovakia,295-305
4) M. Andrade, M. A. M. Ferreira (2009). Criminal and Civil Identification with DNA Databases Using Bayesian Networks. International Journal of Security, 3 (4), 65-74.
5) M. Andrade, M. A. M. Ferreira (2010). Evaluation of paternities with less usual data using Bayesian networks. Proceedings - 2010 3rd International Conference on Biomedical Engineering and Informatics, BMEI 2010, Yantai, China. Volume 6, 2010, Article number 5639678, Pages 2475-2477. DOI: 10.1109/BMEI.2010.5639678
6) M. Andrade, M. A.M. Ferreira (2010a). Civil Identification Problems with Bayesian Networks Using Official DNA Databases. APLIMAT – Journal of Applied Mathematics, 3 (3), 155-162.
7) M. Andrade, M. A.M. Ferreira (2013). Simple Maternity Search with Bayesian Networks. 12th Conference on Applied Mathematics, APLIMAT 2013, Proceedings. Bratislava, Slovakia.
8) M. Andrade, M. A.M. Ferreira (2014). Bayesian Networks use in simple maternity problems. Applied Mathematical Sciences, 8 (137-140), 6963-6967. DOI:10.12988/ams.2014.48617