Journal of Applied Mathematics

volume III (2010), number III

CIVIL IDENTIFICATION PROBLEMS WITH BAYESIAN NETWORKS USING OFFICIAL DNA DATABASES

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

Abstract. In forensic identification problems the study of DNA profiles is often used. DNA databases began to be used in England in 1995 and gave rise to new challenges when used in identification problems. In Portugal the legislation for the construction of a DNA database file was defined in 2008. So, it is important to determine how to use it in an appropriate way. An important forensic identification problem is body identification. That is, in general, the identification of a body found, or more than one, using the information of missing persons belonging to one or more known families for which there may be information of family members who claimed the disappearance. Here it is intend to discuss how to use the database: the hypotheses of interest and the database use to determine the likelihood ratios, i.e., how to evaluate the evidence in different situations.

Key words. Bayesian networks, DNA profiles, civil identification problems

Mathematics Subject Classification: 62C10, 68M10**.**

1 Introduction

The reason of this work is to propose a methodology, and give the adequate tools, to use correctly a DNA profiles database in the problem of civil identification case if there is a partial match between the genetic characteristic of an individual whose body was found, one volunteer who claimed a family member disappearance and one sample in the DNA database.

So in section 2 the civil identification case under study is presented and discussed. In section 3 a Bayesian network that allows the efficient computation of the probabilities determinant to evaluate the hypothesis in comparison is presented. Still in section 3 real life examples, which clarify the exposition, are presented. In the end a short list of references about these kind of problems is given.

2 Civil identification

The use of DNA profiles in forensic identification problems has become, in the last years, an almost regular procedure in many and different situations. Frequent examples of civil identification problems are the case of a body identification, together with the information of a missing person belonging to a known family, or the identification of more than one body resultant of a disaster or an attempt, and even immigration cases in which it is important to establish family relations.

This work focuses on civil identification problems. The establishment and use of DNA database files for a great number of European countries worked as a motivation to study in more detail the mentioned problems and the use of these database files for identification. In the context of the civil identification it may be very useful when unidentified corpses appear and may be identified by comparison of their DNA profiles with family volunteer's profiles.

The Portuguese law n°5/2008 establishes the principles for creating and maintaining a database of DNA profiles for identification purposes, and regulates the collection, processing and conservation of samples of human cells, their analysis and collection of DNA profiles, the methodology for comparison of DNA profiles taken from the samples, and the processing and storage of information in a computer file.

Here it is assumed that the database is composed of a file containing information of samples from convicted offenders with 3 years of imprisonment or more - α; a file containing the information of samples of volunteers - β; a file containing information on the "problem samples" or "reference samples" from corpses, or parts of corpses, or things in places where the authorities collect samples - γ. In this work the interest is to study problems of civil identification, particularly if there is a partial match between the genetic characteristic of an individual whose body was found and one volunteer who claimed a family member disappearance and one sample in the file γ of database.

2.1 A partial match with the volunteer and one $\gamma - sample$

In a problem of civil identification where there is an individual claiming for a disappeared person and gives his/her genetic information, C_{vol}, to be compared with the genetic characteristic of a body found, it is important to check first if there is a match between the genetic characteristic of the individual whose body was found, C_{BF}, and any sample of the DNA file, $\gamma - sample$, which is named "problem samples".

Considering it is checked and there is a partial match between the genetic profile of the individual whose body was found and one sample in the file γ, the evidence now is $E = \left(C_{BF}, \gamma - sample, C_{vol}\right)$.

Regarding the problem it follows the establishment of the hypotheses of interest. The identification hypothesis (H_{ID}) versus the non identification hypothesis ($H_{not\ ID}$), as:

H_{ID}: It is possible to reach an identification of the individual whose body was found

vs

$H_{not\ ID}$: It is not possible to reach an identification of the individual whose body was found.

The first approach is to check the possibility of a partial match between the profile of the individual whose body was found, C_{BF}, the sample in the file γ, $\gamma - sample$, and the volunteer, C_{vol}. Thus, two different comparisons are made in order to obtain a measure either of the possible genetic relation between the individual whose body was found with the $\gamma - sample$ (*bf_match_gs?*), or of the possible genetic relation between the individual whose body was found and the volunteer (*bf_match_vol?*). These comparisons may have as an answer: *yes* or *no*.

Combining the states of each comparison (*yes, no*); (*no, yes*); (*yes, yes*) and (*no, no*) are the resulting pairs.

State: (*yes, no*) – defines the possibility of genetic relationship between the individual whose body was found and the $\gamma - sample$ but not the volunteer;

State: (*no, yes*) – defines the possibility of genetic relationship between the individual whose body was found and the volunteer but not the $\gamma - sample$;

State: (*yes*, *yes*) – defines the possibility of genetic relationship between the individual whose body was found and both the volunteer and the $\gamma - sample$;

State: (*no, no*) – defines the possibility of genetic relationship between the individual whose body was found neither with the volunteer nor with the $\gamma - sample$;

The first two states define the identification hypothesis, H_{ID}, and the last two define the non identification hypothesis, $H_{not\ ID}$. The state (*no, yes*) is a particular one that is the simple problem studied in Andrade and Ferreira (2009b). Each of the four possible states probabilities provide a measure for each event, and the four are pairwise incompatible.

After the probabilities computation it is important to have in mind the comparison between the state (*no, no*) versus the others; i.e., to evaluate the event the individual whose body was found is not genetically related either with the $\gamma - sample$ or the volunteer. This first step comparison intends to evaluate the situation "the genetic information of the individual whose body was found is not compatible with the other genetic information available" and "the genetic information of the individual whose body was found is compatible with at least one of the remaining genetic information", that is, compares the sets $\{(no, no)\}$ with $\{(no, yes), (yes, no), (yes, yes)\}$.

If $\{(no, no)\}$ is accepted the process ends and the body genetic information joins the file γ in the database. If $\{(no, no)\}$ is discarded next it is necessary to perform a comparison between $\{(no, yes), (yes, no)\}$ and $\{(yes, yes)\}$ events**.** If $\{(yes, yes)\}$ is accepted the process ends and police intelligence investigations must be done. If $\{(yes, yes)\}$ is discarded finally $\{(no, yes)\}$ and $\{(yes, no)\}$ must be compared. If $\{(no, yes)\}$ is accepted the conclusion is that the individual whose

body was found is a volunteer relative. If $\{(yes, no)\}$ is accepted the individual whose body was found is related with the $\gamma - sample$.

3 The Bayesian network

The comparisons described in the previous section are done through the respective probabilities events ratios: the likelihood ratios. In each comparison the hypothesis with the greatest probability is the accepted one. So it is imperative to compute the probabilities of the four states defined above. These computations demand the intermediary computation of a lot of conditional probabilities that are impossible to do with algebraic manipulations. In consequence these probabilities will be computed using the Bayesian network, see Andrade and Ferreira (2009a), in the Figure 1.

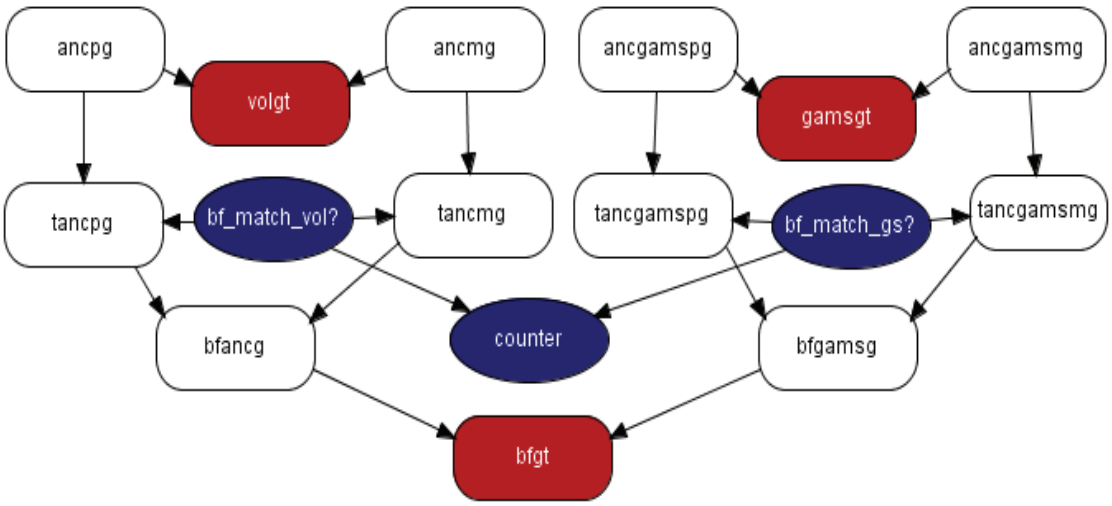

Figure 1**: Network for civil identification with one volunteer and one** $\gamma - sample$ **partial match case**.

The nodes **ancpg, ancmg, ancgampg** and **ancgammg** are of class founder, a network with only one node which states are the alleles in the problem and the respective frequencies in the population, and represent the volunteer's ancient paternal and maternal inheritance. Nodes **volgt, gamsgt** and **bfgt** are of class genotype: the volunteer, the $\gamma - sample$ and the body found genotypes.

Nodes **tancmg, tancpg, tancgamspg** and **tancgamsmg** specify whether the correspondent allele is or is not the same as the volunteer and the same as the $\gamma - sample$. If **bf_match_vol?** has true for value then the volunteer's allele will be identical with the body found allele, otherwise the allele is randomly chosen in the population and if **bf_match_gs?** has true for value then the $\gamma - sample$'s allele will be identical with the body found allele, otherwise the allele is randomly chosen in the population. The node **bfancg** and **bfgamsg** defines the Mendel inheritance in which the allele of the individual whose body was found is chosen at random from the ancient's paternal and maternal gene. Node **counter** counts the number of true states of the preceding nodes, accounting the results for the $\{(no, no)(no, yes), (yes, no), (yes, yes)\}$ possible events.

3.1 Examples

In order to exemplify the described methodology in Table 1 are presented the allele frequencies (real ones) for some genetic markers[1] and, for each marker, possible evidence profiles for the body found (C_{BF}), the $\gamma - sample$ and the volunteer (C_{vol}).

Marker	Allele Frequencies				$\{(C_{BF}),(\gamma - sample),(C_{vol})\}$
FGA	p_{21}	p_{22}	p_{23}	p_{24}	$\{(21,24),(21,21),(22,24)\}$
	0.1750	0.1950	0.1550	0.1750	
D21S11	p_{28}	p_{29}	p_{30}	$p_{31.2}$	$\{(29,30),(30,30),(29,31.2)\}$
	0.1674	0.2136	0.2437	0.1138	
F13A1	p_5	p_6	p_7	p_8	$\{(6,7),(7,8),(5,6)\}$
	0.1985	0.2890	0.3377	0.0112	
SE33	$p_{22.2}$	$p_{25.2}$	$p_{27.2}$	$p_{28.2}$	$\{(22.2,27.2),(22.2,28.2),(27.2,25.2)\}$
	0.1043	0.0764	0.1458	0.0695	
TH01	p_6	p_7	p_9	$p_{9.3}$	$\{(7,9),(9,9.3),(6,7)\}$
	0.2044	0.1696	0.1984	0.2748	
TPOX	p_8	p_9	p_{10}	p_{11}	$\{(8,11),(8,10),(9,11)\}$
	0.5053	0.0974	0.0647	0.2893	
VWA31	p_{15}	p_{16}	p_{17}	p_{18}	$\{(16,17),(17,17),(16,18)\}$
	0.1216	0.2300	0.2649	0.1859	

Table1**: Allele frequencies and genetic profiles.**

In Table 2 the state probabilities (the node counter states, see Figure 1) are presented.

States	**FGA**	**D21S11**	**F13A1**	**SE33**	**TH01**	**TPOX**	**VWA31**
(*no, no*)	0.0744	0.1108	0.2574	0.0751	0.1343	0.3112	0.1108
(*no, yes*)	0.1063	0.1296	0.2226	0.1287	0.1978	0.2688	0.2251
(*yes, no*)	0.2124	0.2274	0.1904	0.1797	0.1692	0.1539	0.2092
(*yes, yes*)	0.6069	0.5322	0.3296	0.6165	0.4987	0.2661	0.4548

Table 2: **State probabilities.**

[1] http://www.uni-duesseldorf.de/WWW/MedFak/Serology/dna.html

And in Table 3 the decisions, consequence of the procedures proposed in section 2.1, are presented for each example evidence profile.

Evidence Profiles	Decision
$\{(21,24),(21,21),(22,24)\}$	Police intelligence investigations must be done
$\{(29,30),(30,30),(29,31.2)\}$	Police intelligence investigations must be done
$\{(6,7),(7,8),(5,6)\}$	The individual whose body was found is a volunteer relative
$\{(22.2,27.2),(22.2,28.2),(27.2,25.2)\}$	Police intelligence investigations must be done
$\{(7,9),(9,9.3),(6,7)\}$	Police intelligence investigations must be done
$\{(8,11),(8,10),(9,11)\}$	The individual whose body was found is a volunteer relative
$\{(16,17),(17,17),(16,18)\}$	Police intelligence investigations must be done

Table 3: **Decisions for each evidence profile**

4 Conclusions

Performing the sequence of three hypothesis tests proposed with the probabilities computed through the Bayesian network, built specifically for a civil identification problem in which there is a partial match between an individual whose body was found, a volunteer who claimed a relative disappearance supplying his/her own genetic information and a DNA database file sample existent, it is possible to decide first if an identification is possible or not; second if an effective identification is possible or not; third to make the identification. The comparison between the hypotheses in each test is made through its probabilities values computing the respective likelihood ratio. The accepted hypothesis is the one that corresponds to the greatest probability event.

So with a procedure technically simple, since it was defined the Bayesian network and the chain of hypothesis tests it is possible to make an adequate and correct use of a DNA database.

And as the examples illustrate, the procedure leads almost surely to a decision: whether it is to close the case identifying the individual, or concluding that it is not possible any identification, or to go on with the police investigations.

Acknowledgement

The authors are members of the UNIDE/ISCTE research group StatMath/ISCTE which support they gratefully thank

References

[1.] ANDRADE, M., FERREIRA, M. A. M.: "*Bayesian networks in forensic identification problems*". Aplimat - Journal of Applied Mathematics. Volume 2, number 3, 13-30, 2009.
[2.] ANDRADE, M. and FERREIRA, M. A. M.: *Criminal and Civil Identification with DNA Databases Using Bayesian Networks*. International Journal of Security. Volume 3, issue 4, 65-74, 2009.
[3.] ANDRADE, M., FERREIRA, M. A. M., FILIPE, J. A.: *"Evidence evaluation in DNA mixture traces"*. Journal of Mathematics and Allied Fields (Scientific Journals International-Published online). Volume 2, issue 2, 2008.
[4.] ANDRADE, M., FERREIRA, M. A. M., FILIPE, J. A., COELHO, M.: *"Paternity dispute: is it important to be conservative?"*. Aplimat – Journal of Applied Mathematics. Volume 1, number 2, 2008
[5.] BALDING, David J.: *"The DNA database controversy"*. Biometrics, 58(1):241-244, 2002
[6.] CORTE-REAL, F.: *"Forensic DNA databases"*. Forensic Science International, 146s:s143-s144, 2004.
[7.] COWELL, R. G., DAWID, A. P., LAURITZEN, S. L., SPIEGELHALTER, D. J.: *"Probabilistic Expert Systems"*, Springer, New York, 1999.
[8.] DAWID, A. P., MORTERA, J., PASCALI, V. L. Van BOXEL, D. W.: *"Probabilistic expert systems for forensic inference from genetic markers"*. Scandinavian Journal of Statistics, 29:577-595, 2002.
[9.] EVETT, I., WEIR, B. S.: *"Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists"*, Sinauer Associates, Inc., 1998.
[10.] GUILLÉN, M., LAREU, M. V., PESTONI, C., SALAS, A., CARRECEDO, A.: *"Ethical-legal problems of DNA databases in criminal investigation"*. Journal of Medical Ethics, 26:266-271, 2000.
[11.] MARTIN, P.: *"National DNA databases – practice and practability. A forum for discussion"*. In International Congress Series 1261, 1-8, 2004.
[12.] NEAPOLITAN, R. E.: *"Learning Bayesian networks"* , Pearson Prentice Hall, 2004.

Current address

Marina Andrade, Professor Auxiliar
Iscte – Lisbon University Institute
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
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