Journal of
Mathematics, Statistics and Allied Fields
Volume 2, Issue 2, 2008

Evidence evaluation in DNA mixture traces

Marina Andrade, Assistant Professor, ISCTE Business School, marina.andrade@iscte.pt
Manuel A. M. Ferreira, Full Professor, ISCTE Business School, manuel.ferreira@iscte.pt
José António Filipe, Assistant Professor, ISCTE Business School, jose.filipe@iscte.pt

Abstract

Nowadays the use of DNA profiles in forensic identification problems is a very common procedure. The observance of a mixture trace resulting of a crime that has been committed is a very interesting and challenging phenomenon for the laboratories and for the judicial systems. The hypotheses that can be considered and compared in cases of mixture traces are discussed in this work. In court two hypotheses, the prosecution hypothesis and the defence hypothesis, lead the dispute in the case presented. The use of Bayesian networks in the analysis of DNA mixture traces is a powerful tool in complex cases. That allows an easy evaluation of the likelihoods for the whole hypotheses suggested by each case.

1 Introduction

The development of molecular biology and the knowledge of DNA structure allowed its use as genetic information vehicular, and therefore DNA has been more commonly used for the clarification of judicial forensic identification problems. Every human being has 23 pairs of chromosomes in the nuclear of a cell. A locus, also named a gene for simplification, is an area on a chromosome and the DNA composition on that area is an allele. A DNA marker is a locus for which it is known the possible alleles that can be measured. Thus, it is possible to determine an individual DNA profile, i. e., for an individual to measure his/her genotype on several markers. An unordered pair of alleles, one inherited from the individual's father and the other from the mother, although it is not possible to distinguish which is which, composes the genotype. It is assumed independence of the individual's alleles both within and across markers, i.e., Hardy - Weinberg and linkage equilibrium assumptions. This work intends to discuss different approaches in the analysis of DNA mixture profiles. In a mixture with two contributors there are four hypotheses to compare. A more complex and real case may become very burdensome in the number of hypotheses and algebraic calculations to perform. In section 2 are presented the different hypotheses to test according to the possible mixture traces. Beyond the two hypotheses emergent in court, for each case, others may also be of interest. Therefore it is needed a form to evaluate those in an efficient way. Starting with an algebraic approach then it is pursued to the use of Bayesian networks, particularly important in complex cases of mixture profiles. In section 3 the discussion comprises possible analysis and the potential use of Bayesian networks in this context.

2 Hypotheses

Mixture traces are generally observed in criminal cases. When a DNA mixture profile is mentioned it means that were observed more than two bands for one or more loci in the set of known genetic markers used to analyze the trace found. Here it is discussed for illustration the set of hypotheses of a mixture with two contributors. After this the data of a more complex mixture is presented. For the purpose intended it is presented data only for two markers.

2.1 Algebraic approach

Before proceeding to a more complex case, it is briefly discussed the hypotheses to test in a case to which a mixture trace was found and connected with a certain crime. Assume that there were only two donors. Two known individuals are measured: a victim v and a suspect s. In such a case the competing hypotheses are:

(i) $s\&v$ (ii) $s\&u$ (iii) $v\&u$ (iv) $2u$

with u an unknown individual in the population. Where (i) means that the mixture is composed with genetic material of the victim and the suspect; (ii) the mixture composition is formed with DNA of the suspect and an unknown individual; (iii) the mixture is composed with DNA of the victim and an unknown individual; and (iv) the mixture is formed with DNA of two unknown individuals.

In court the hypotheses in dispute are the prosecution hypothesis, stating that the mixture is composed with genetic material from the victim and the suspect $\{v, s\}$ *versus* the defence hypothesis arguing that the mixture results of biological material from the victim and an unknown person $\{v, u\}$. Given the allele frequencies it is possible to determine algebraically the likelihood ratio values for the competing hypothesis, Weir *et al.* (1997). That can be more or less heavily depending on the mixture observed and the known individual genotype's involved. With the likelihood for each hypothesis one may want to compare their ratios: $(s\&v \text{ vs } v\&u)$ or $(s\&v \text{ vs } 2u)$ or $(s\&u \text{ vs } 2u)$.

Consider now an excerpt of a criminal case with two victims (v_1, v_2) and a suspect (s_2). The competing hypotheses are: the mixture has DNA from the victim's and the suspect *versus* the mixture has DNA from the victim's and an unknown person for the prosecution and the defence hypothesis, stating the suspect's guilty (H_P) and innocence (H_D), respectively. The set of data is:

Marker	Gv_1 (f)	Gv_2 (m)	Gs_2	E_c (mixture)
FES	A, C	C, C	B, B	A, B, C
FGA	B, E	B, C	A, C	A, B, C, E

Table 1: Mixture trace data

where G_x refers the genotype of individual x and E_c refers the crime scene evidence for the two known markers.

As in any criminal case the court has to answer the question $P(s_2 \text{ is guilty } |E)$? Therefore it seems natural just to want to compare that probability with the following $P(s_2 \text{ is not guilty } |E)$, with $E=(E_c, G_{v1}, G_{v2}, G_{s2})$. In order to compare the hypotheses one can determine the ratio of these two hypotheses as follows:

$$\frac{P(s_2 \text{ is guilty}|E)}{P(s_2 \text{ is not guilty}|E)} = \frac{P(E_c \mid G_{v1}, G_{v2}, G_{s2}, H_P)}{P(E_c \mid G_{v1}, G_{v2}, G_{s2}, H_D)} \times \frac{P(H_P)}{P(H_D)} \qquad (1)$$

Supported on the data the determination of the likelihood ratio (the first term of the right side of the equation 1) is easy. Thus, the probability of the evidence given the prosecution

hypothesis is one. The probability of the evidence conditional to data and the defence hypothesis can be obtained as the product of the last column of the table 2 below:

Marker	$P(E_c \mid G_{v1}, G_{v2}, G_{s2}, H_D)$
FES	$p^2_B + 2.\ p_A\ .\ p_B + 2\ .\ p_{\boldsymbol{B}}\ .\ p_C$
FGA	$p^2_A + 2.\ p_A\ .\ p_B + 2\ .\ p_{\boldsymbol{A}}\ .\ p_C + 2\ .\ p_{\boldsymbol{A}}\ .\ p_E$

Table 2: Probability of the evidence given the defence hypothesis

With these values one can compare the hypotheses. But, in such a case it is reasonable to be interested in a comparison of a more large set of hypotheses viewing the possible origin of the mixture - a source level proposition according to Cook *et al*. (1998). One of the complexities in the interpretation and in the evaluation of a mixture trace is to assign the number of total contributors. The various numbers of alleles present in the mixture suggest a minimum for that number but say nothing about a maximum to consider. For this Lauritzen and Mortera (2002) gave a useful low upper bound to the number of contributors to consider, and it can give some clues.

During the investigation period of a criminal case most of the time it is important to take into account other possible scenarios. In the mentioned case it was thought that three individuals were involved and the set of hypotheses to test may admit up to three unknown individuals to consider. Thus, consider up to six contributors in the mixture. The set of hypotheses to test will have a total of 32 states. A mixture with three contributors has eight hypotheses to test. In this case it must be considered those eight states for the known individuals plus those eight states combined with one, two and three unknowns (see table 3). Naturally to determine the expression and the value for each hypothesis becomes a difficult and slow task, making it difficult to proceed with the analysis. In court or during the investigation process of a real case it is extremely important to define, in reasonable time, the weight of each hypothesis in evaluation.

0) Ø	8) u	16) $2u$	24) $3u$
1) v_1	9) v_1, u	17) $v_1, 2u$	25) $v_1, 3u$
2) v_2	10) v_2, u	18) $v_2, 2u$	26) $v_2, 3u$
3) v_1, v_2	11) v_1, v_2, u	19) $v_1, v_2, 2u$	27) $v_1, v_2, 3u$
4) s_2	12) s_2, u	20) $s_2, 2u$	28) $s_2, 3u$
5) s_2, v_1	13) s_2, v_1, u	21) $s_2, v_1, 2u$	29) $s_2, v_1, 3u$
6) s_2, v_2	14) s_2, v_2, u	22) $s_2, v_2, 2u$	30) $s_2, v_2, 3u$
7) s_2, v_1, v_2	15) s_2, v_1, v_2, u	23) $s_2, v_1, v_2, 2u$	31) $s_2, v_1, v_2, 3u$

Table 3: Set of 32 states to test

2.2 Bayesian networks approach

The complex identification problems raised in the forensic area encourage the interest for the development of inference mechanisms allowing the search and attainment of answers for this kind of problems. The use of Bayesian networks in the analysis of DNA mixture profiles had its beginning with the works of Mortera (2003) and Mortera *et al*. (2003). Since then a more common discussion and the computational improvements achieved within the

software's grant a good support to the authorities, whether they are: the courts or the polices.

In section 2.1 the 32 hypotheses listed intend to cover different assumptions for different scenarios, considering the involvement of the three nominated individuals till the involvement of three unknown individuals. As it was highlighted the problems appear when it is needed to determine the expression and the correspondent value of each hypothesis. For this case Andrade and Ferreira (2007) have performed the analysis with object-oriented Bayesian networks (OOBN), considering there a total of five markers. After having built the networks, supported with a software program[1] , and inserted the evidence (data) the results were obtained in a simple and quick way. An example of the results is given in table 4, referring the 32 values of each hypothesis for marker FGA.

For some states the likelihood values are null. This happens when the hypothesis is not consistent with the minimum number of individuals' necessary to generate the mixture inserted, i.e., hypothesis not compatible with the data. For states 1) 2) 4) and 8) the likelihood is null because they comprise one only individual; so they do not refer a mixture. States 3) and 6) refer a mixture but its genotypes do not match the data and the correspondent likelihood is also zero. State 0) mentions the absence of any individual and its likelihood is obviously zero. Of course the likelihoods for state 0) and 1), 2) 4) and 8) are null whatever is the marker. But states 3) and 6) may be non-null for other markers. Some hypotheses present a larger value for the likelihood ratio than others that is discussed in the next section.

state	FGA	state	FGA	state	FGA	state	FGA
0)	0.0000	8)	0.0000	16)	0.0006	24)	0.0004
1)	0.0000	9)	0.0083	17)	0.0052	25)	0.0015
2)	0.0000	10)	0.0017	18)	0.0015	26)	0.0006
3)	0.0000	11)	0.0207	19)	0.0067	27)	0.0016
4)	0.0000	12)	0.0042	20)	0.0029	28)	0.0009
5)	0.3768	13)	0.0714	21)	0.0135	29)	0.0026
6)	0.0000	14)	0.0101	22)	0.0036	30)	0.0009
7)	0.3768	15)	0.0714	23)	0.0135	31)	0.0026

Table 4: Results for marker FGA and the given data

3 Comments

Mixture traces analysis and evaluation present obviously difficulties. For each case the analysis has to be performed according the particular situation. The algebraic approach becomes more complex when it is admitted one more person in a mixture. If there is interest in questioning a bigger number of contributors the complexities increase largely. In this paper a complex situation, not usually considered, is studied and an important resource to deal with these problems, the mentioned Bayesian networks, is used.

In a criminal case of forensic identification, before its evaluation in court, usually it is necessary to test and compare a certain number of hypotheses connected with the inherent

[1] www.hugin.com

conjectures. At an earlier stage of a case investigation process the policies have to define the reasonable scenarios and to determine the important ones that will be detailed studied, in order to determine the ones to be evaluated by the courts. And even in court beyond the main comparison some others may be of interest to consider. Whatever the circumstances are, to perform those comparisons as quick and efficiently as possible is an exigency of the whole parts involved in the judicial area. The results of table 4 give an illustration of what can be tested and that some conjectures should not be left to appreciate. For example, some of the hypotheses in which it is considered the presence of an unknown person are not to be depreciated.

Also worth to be mentioned is the modularity and flexibility of OOBN, which allow its possible use in cases with similar details and the extension to more complex cases. The different modules or instances can be reused to analyse various problems. In the new problem one can define the necessary new objects and combine them with the already defined ones, and deal with the singularity of each case. Even an actual case can be simulated using OOBN and compare the court ruling in it since the courts facilitate the information, which is often considered confidential. But in fact some of the problems studied using OOBN correspond to courts commands.

Acknowledgements

- The authors are members of UNIDE/ISCTE Research Centre and thank its support in this investigation.

- The authors wish to thank the editor and the referees their valuable comments that contributed very much to improve this paper.

References

Andrade, M., and Ferreira, M. A. M., 2007. Analysis of a DNA mixture sample using object-oriented Bayesian networks*. In proceedings of the 6th International Conference APLIMAT 2007*, Feb 6-9, Bratislava, Slovak Republic.

Cook, R., Evett, I. W., Jackson, G., Jones, P. J. and Lambert, J. A. (1998). A hierarchy of propositions: Deciding which level to address in casework. *Science and Justice*, 38, 151-156.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., Spiegelhalter, D. J. (1999). *Probabilistic expert systems*. Springer, New York.

Lauritzen, S.L. and Mortera, J., 2002. Bounding the number of contributors to mixed stains. *Forensic Science International*, 130, 125-126.

Mortera, J., 2003. Analysis of DNA mixtures using probabilistic expert systems. In: Green, P.J., Hjort, N.L., Richardson, S. (Eds.), *Highly Structured Stochastic Systems*. Oxford University Press.

Mortera, J., Dawid, A. P., and Lauritzen and S. L., 2003. Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology*, 63, 191-205.

Weir, B. S., Triggs, C. M., Starling, L., Stowell, L. I., Walsh, K. A. J., and Buckleton, J. S. (1997). Interpreting DNA mixtures. *Journal of Forensic Sciences*, 42, 213-22.