Current challenges in statistical DNA evidence evaluation

Cereda, G.

Introduction

One of the main aims of forensic statistics is to evaluate to what degree some piece of evidence supports one or the other of the hypotheses of interest in a judicial setting. The largely accepted method to perform this evaluation is the calculation of the *likelihood ratio*, a statistic that expresses the relative plausibility of the observations under the hypotheses. For instance, a typical piece of evidence may be a trace, found at the crime scene, containing DNA material from a single donor and whose profile corresponds to a known suspect's DNA profile. A couple of mutually exclusive hypotheses is typically defined, of the kind of 'the crime stain came from the suspect' (h_p) and 'the crime stain came from an unknown donor' (h_d). The likelihood ratio is the ratio of the probabilities of observing the matching profiles under these two hypotheses.

In case of DNA mixtures (traces containing DNA from several contributors) common genotyping techniques do not allow to distinguish the DNA profile of each contributor. This complicates the statistical evaluation of this kind of evidence, since different combinations of DNA profiles are compatible with belonging to the (known and unknown) contributors to the stain. Moreover, using standard techniques, if the quantitative share of DNA of one of the contributors is less than 10% of the total DNA quantity, his DNA profile is generally 'masked' by the DNA profile of the other contributor(s). As a consequence, it is very difficult to detect this minor DNA with classical methods of genotyping. Unbalanced mixtures of this kind are quite common, for example in cases of sexual assaults when the victim's DNA is largely predominant. This means that there is a paramount need for reliable solutions.

The advent of a new technology, the DIP-STR (short for Deletion Insertion Polymorphisms - Short Tandem Repeats) marker system, constitutes an answer to the problem represented by the extremely unbalanced mixtures.

The initial aim of this thesis was to develop a Bayesian statistical model to evaluate DIP-STR results in the light of competing hypotheses of interest: this represents an essential element for rendering the potential of this new typing technique useful for practitioners. Furthermore, in this initial project we compared, from a statistical and forensic perspective, the usefulness and usability of the DIP-STR markers with that of traditional marker systems, such as classical STR and Y-STR markers.

While in progress, we were confronted with several delicate methodological issues regarding forensic statistics: first, we noticed that the Bayesian methods used in the literature can be seen as ad hoc approximation to the full Bayesian solution. Then, we were confronted with the so-called 'rare type match problem', the situation in which there is a match between the characteristics of some recovered material and those of the control material, but these characteristics have not been observed yet in previously collected samples (i.e., they do not occur in any existing database of interest for the case). The statistical evaluation of such a scientific finding depends on the rarity of the characteristic of interest (such as a DNA profile) in the population of reference. Indeed, the rarer it is, the higher is the likelihood ratio. The uncertainty over this rarity is usually dealt with using the observed relative frequency of the profile in some available database, but in case of no occurrence existing solutions are not satisfactory. The rare type match problem is particularly significant when Y-STR (or mitochondrial) DNA profiles are used, and in presence of new genotyping techniques (such as DIP-STR markers), for which the available database size is still limited.

We decided to start working with Y-STR data to study both new and existing solutions for the evaluation of rare type matches: classical Bayesian methods (beta-binomial and Dirichlet-multinomial) were revisited and compared to a Bayesian nonparametric approach tailored explicitly for the rare type match problem. Two frequentist solutions are also analysed: the discrete Laplace method and a new solution based on the Good-Turing estimator.

While studying frequentist solutions, we realised that different methods are based on different reductions of data, and that this is seldom discussed in forensic literature. Moreover, frequentist solutions involve different levels of uncertainty which have to be considered and discussed. Working on both frequentist and Bayesian frameworks allowed us to better understand the difference between the two approaches, and between the full Bayesian approaches and the plug-in approximations found in the literature. A lemma is also developed to help obtaining the full Bayesian likelihood ratio.

As a closing loop for this project, one of the Bayesian methods developed as a solution for the rare type match problem is applied to DIP-STR data. Also, the evaluative model built for DIP-STR data, in the initial stages of the research, has been improved and extended to incorporate parameter uncertainty in a consistent Bayesian way.

Scope and Propositions

This thesis covers different problems concerning the evaluation of DNA evidence. It is mainly divided into two parts: the first regards the DIP-STR genotyping techniques. It addresses the imperative need of developing a model to assign the likelihood ratio for DIP-STR results, and compares, from a statistical and forensic perspective, the advantages of these novel set of markers compared to traditional marker systems, such as STR and Y-STR.

The second part deals with several more general statistical aspects involved in the evaluation of DNA evidence. It aims at defining the differences between full Bayesian methods and ad hoc plug-in approximations, and at solving the rare type match problem for Y-STR data. The issues of the different reductions of data and of the levels of uncertainty involved in frequentist solutions are also discussed.

These two parts are connected in the final project, by developing a Bayesian solution for the rare type match problem for DIP-STR marker system. Moreover, the initial model for

DIP-STR data is improved in the light of the statistical discussion of the second part: any ad hoc solution is avoided to obtain a full Bayesian approach.

Novelty

Extremely unbalanced mixtures are still a challenging area of DNA analysis. The DIP-STR marker system is a recently developed technique of genotyping, which has only been discussed from a biological point of view: when this PhD project started, no Bayesian statistical solutions for the likelihood ratio assessment of DIP-STR evidence was available, making this research innovative and extremely useful. Moreover, this research will take advantage of the use of graphical probability models, in particular OOBNs, because of their very intuitive and flexible structures. They provide a powerful language for constructing knowledge-based models for reasoning under uncertainty, and significantly simplify the calculation of the statistics of interest. This represents an improvement if compared to the formulaic approach, classically used in literature.

The rare type match problem is still an open challenge of forensic statistics: it is so important a problem that it has been called ‘the fundamental problem of forensic mathematics’ (Brenner, 2010). This research produced, compared, and discussed new methodologies (both Bayesian and frequentist) to address this problem. Some methods have been developed specifically with this purpose, others have been adapted from existing ones. The use of Bayesian nonparametric methods represents a novelty in forensic science. Lastly, a solution for the rare type match problem encountered with the DIP-STR marker system is provided and discussed.

An additional contribution of this thesis to the theoretical statistical framework of forensic science is the discussion about the distinction between Bayesian and frequentist approaches to likelihood ratio assessment. This is important inasmuch the ad hoc plug-in approximations can be seen as hybrid solutions between the two. Moreover, a new Lemma is introduced and proved. It is of very broad application, since it can be used in all those situations (very common in forensic science) in which one wants to obtain the Bayesian likelihood ratio for data that depends on parameters, when prosecution and defence agree on the distribution of part of the data, but disagree on the distribution of the rest of the data.

Outline

The core structure of this research is represented by a series of papers written during the five years of this doctoral research. The papers are:

- Cereda, G., Biedermann, A., Hall, D., and Taroni, F. (2014) “Object-oriented Bayesian networks for evaluating DIP-STR profiling results from unbalanced DNA mixtures” *Forensic Science International: Genetics,* Vol. 8, pag. 159-169.
- Cereda, G., Biedermann, A., Hall, D., and Taroni, F. (2014) “An investigation of the potential of DIP-STR markers for DNA mixture analyses” *Forensic Science International: Genetics,* Vol. 11, pag. 229-240.

- Cereda, G. (2016) “Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach)” *Scandinavian Journal of Statistics,* In Press.
- Cereda, G. (2016) “Bayesian approach to LR for the rare match problem” *Statistica Neerlandica*, In Press.
- Cereda, G. “Nonparametric Bayesian approach to LR assessment in case of rare haplotype match” (submitted to *Annals of Applied Statistics*).
- Cereda, G., Gill, R. D., and Taroni, F. “A solution for the rare type match problem when using the DIP-STR marker system” (submitted to *Forensic Science International: Genetics*).

The dissertation is structured in three parts.

Part I
Chapter 1 contains the forensic and statistical knowledge essential to appreciate the results of this research, while Chapter 2 summarizes the results which constitute the content of each paper, and provides the logical thread that links the diverse studies.

Part II
Chapters 3 to 8 are each in the form of a separate research paper, in the order in which they are written.

Part III
Chapter 9 discusses the contribution and implications of the results in a larger statistical and forensic context, and describes further research directions and open questions.

Notation

Throughout Chapter 1 and 2, random variables and their values are denoted, respectively, with uppercase and lowercase characters: x is a specific realisation of X. Random vectors and their values are denoted, respectively, by uppercase and lowercase bold characters: $\mathbf{p}$ is a realisation of the random vector $\mathbf{P}$. Bayesian probability is denoted with $\Pr(\cdot)$, while density of a continuous random variable X is denoted by $p(x)$ or $f(x)$. For a discrete random variable Y, both the continuous notation $p(y)$ and the discrete one $\Pr(Y = y)$ will be used. Frequentist probability will be denoted as $\mathscr{P}r$. Occasionally we deviate from this rule. For instance, when dealing with graphical models, the notation changes: `teletype` is used for classes, **bold** for nodes and *italic* for states. Chapters 3, 4, 5, 6, 7, 8 stand aside and have a distinct convention, mostly based on the journal requirement for each paper.

Part I

Background and summary