Forensic Science International: Genetics 26 (2017) 40–47

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Research paper

Identifying common donors in DNA mixtures, with applications to database searches

K. Slooten [a,b]

[a] Netherlands Forensic Institute, P.O. Box 24044, 2490 AA The Hague, The Netherlands
[b] VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 11 March 2016
Received in revised form 13 August 2016
Accepted 1 October 2016
Available online 12 October 2016

Keywords:
DNA mixtures
Dropout
Deconvolution
DNA database search
Likelihood ratio distributions

ABSTRACT

Several methods exist to compute the likelihood ratio $LR(M, g)$ evaluating the possible contribution of a person of interest with genotype g to a mixed trace M. In this paper we generalize this LR to a likelihood ratio $LR(M_1, M_2)$ involving two possibly mixed traces M_1 and M_2, where the question is whether there is a donor in common to both traces. In case one of the traces is in fact a single genotype, then this likelihood ratio reduces to the usual $LR(M, g)$. We explain how our method conceptually is a logical consequence of the fact that LR calculations of the form $LR(M, g)$ can be equivalently regarded as a probabilistic deconvolution of the mixture.

Based on simulated data, and using a semi-continuous mixture evaluation model, we derive ROC curves of our method applied to various types of mixtures. From these data we conclude that searches for a common donor are often feasible in the sense that a very small false positive rate can be combined with a high probability to detect a common donor if there is one. We also show how database searches comparing all traces to each other can be carried out efficiently, as illustrated by the application of the method to the mixed traces in the Dutch DNA database.

1. Introduction

Over the last years, much attention has been paid to the development of software for the calculation of likelihood ratios for comparisons of a mixture and a known person, to evaluate the weight of evidence in favour of that person's contribution to the mixture. There are several such programs now available, broadly speaking divided into two classes: the models that evaluate as evidence the recorded alleles of the mixture (called binary or semi-continuous models, depending on their ability to handle dropout) and those that also take the peak heights into account, called the continuous models. An overview of available software and their features is provided in [1]. What these programs have in common with each other is that they enable the computation of likelihood ratios (LR's) of the form $LR(M, g) = P(M \mid H_1)/P(M \mid H_2)$ where H_1 and H_2 are hypotheses stating the contributors of the mixture M; the difference between H_1 and H_2 is usually only that H_1 states the contribution of an individual with genotype g and H_2 has replaced that person by an unknown individual. In this article we are going to compute LR's denoted $LR(M_1, M_2)$ which compare two mixed traces M_1, M_2 with each other, to evaluate whether or not they have a donor in common. This is a generalization of the LR's of the form $LR(M, g)$, since if the mixture M_2 is in fact a single source trace from a person whose genotype g can be deduced with certainty, then $LR(M_1, M_2) = LR(M_1, g)$. Likelihood ratios of the form $LR(M_1, M_2)$ can obviously be useful in order to establish a connection between the two cases in which the mixtures M_1 and M_2 have been found. The structure of this paper is as follows. First, we explain how the computation of LR's of the form $LR(M_1, M_2)$ can be done efficiently by recalling that the computation of LR's $LR(M, g)$ in fact is equivalent to a probabilistic deconvolution of the mixture M, giving us a probability distribution on the genotypes of the donors of the mixture. It then suffices to do this for both M_1 and M_2 and match the derived donors with each other, taking the probability distributions into account. Having established the necessary theory, we then apply this method to various simulated mixtures, in order to determine whether or not the obtained LR's are in practice sufficiently well discriminating between the situations with and without a common donor. Finally, we report on the application of this method to the mixtures in the Dutch DNA database. The comparison of mixtures with possible dropout to a database of individuals in order to find the donors of the mixture has been described in various places, e.g. [2–4]. The method presented here is a generalization towards the comparison of any pair of DNA profiles, i.e., mixed traces or reference profiles,

E-mail address: k.slooten@nfi.minvenj.nl.

http://dx.doi.org/10.1016/j.fsigen.2016.10.003
1872-4973/

including the possibility to take an arbitrary number of replicate analyses into account.

We have worked throughout with the semi-continuous model described in [5], where all donors of the mixture are assigned their own probability of dropout. This allows to also investigate whether a major donor to one mixture, is possibly the same person as a minor donor of another mixture. We note however, that although we used a particular choice of mixture evaluation model here, the approach described in this article can in principle be applied to other mixture evaluation models, in particular also for the continuous models. However, the computations can become overly demanding if the likelihood ratio does not factorize over the considered loci, as is the case when it is obtained by integration over parameters affecting the mixture likelihoods. Finally, a reason to work with the semi-continuous model is that we apply the method to mixtures stored in a database. In this database there is no peak height information.

2. Methods

2.1. Semi-continuous model

We start by recalling the characteristics of the semi-continuous model that we use in this article. This description is a summary of the one given in [6] and we refer the reader to that paper for further details. Suppose that a mixture has n contributors. Let their dropout rates be $d_1, \ldots, d_n$ with $0 \le d_i \le 1$, and let $c \ge 0$. We define the probability that allele a is detected in mixture M as

$$P_{\vec{d},c}(a \in \mathcal{M}|\vec{g}) = 1 - e^{-cp_a} \prod_{i=1}^{n} d_i^{n_{i,a}}, \tag{2.1}$$

where $n_{i,a} \in \{0, 1, 2\}$ is the number of alleles a present in g_i, the genotype of contributor i (by definition, $0^0 = 1$). Note that, when $c = 0$, we see from this formula that an allele is recorded unless it drops out for all the contributors that have that allele. In [5,6] we have used the approximation $e^{-cp_a} \approx 1 - cp_a$ for $c \ll 1$. The formula used here corresponds to a Poisson distribution with parameter c for the number of alleles that drop in. The parameter c is therefore equal to the expected number of alleles dropping in per locus. Given the number of alleles that drop in, the alleles that drop in are then obtained as a multinomial sample using the allele frequencies. In particular they are allowed to be identical to each other or to an already present allele coming from a contributor. To compute the probability that the observed mixture $\mathcal{M}$ is equal to the set of alleles M, one simply uses (2.1) to obtain

$$P_{\vec{d},c}(\mathcal{M} = M|\vec{g}) = \prod_{x \in M} P_{\vec{d},c}(x \in \mathcal{M}|\vec{g}) \prod_{x \notin M} P_{\vec{d},c}(x \notin \mathcal{M}|\vec{g}). \tag{2.2}$$

The probability to observe $\mathcal{M} = M$ when some of the donors have unknown genotypes is obtained by summing (2.2) over the set of possible genotypes for these donors, weighted by their prior probability to be the donor's genotypes. Suppose that the population frequency of genotype g is denoted $p(g)$, by which we mean the probability that a person chosen at random from the population has genotype g. In standard mixture calculations, without relatedness in the hypotheses, and without applying the θ-correction (cf. [7]), one sets

$$P_{\vec{d},c}(D_i = g) = p(g) \tag{2.3}$$

as the a priori distribution of the genotype of D_i of donor i. Note that this is of course independent of $\vec{d}$ and c: regardless of the mixture or of how we evaluate it, we assume prior to having any mixture data that each donor is a random person from the population.

In this article we will not apply a θ-correction. In principle, the methodology described below can be generalized to incorporate this correction, but from a computational point of view this is unattractive, as we will point out below. Another justification lies in the fact that comparing mixtures to test whether they have a donor in common amounts to the retrieval of information for investigative purposes, and not to the calculation of the weight of evidence against a specific suspect.

2.2. Deconvolution

Suppose that we have a mixture M, with donors $D_1, \ldots, D_n$. We view the D_i as random variables on the set of possible genotypes. Suppose that there is a person of interest (PoI) S with genotype g. Let $D_1, \ldots, D_k$ (with $k \ge 0$) be the undisputed contributors to the mixture, all of whose genotypes $D_i = g_i$ are known. A standard approach is to define hypotheses H_1 and H_2, where H_1 states that $D_1 = g_1, \ldots, D_k = g_k, D_{k+1} = g$, the other $n - k - 1$ contributors being unknown, and H_2 states that $D_1 = g_1, \ldots, D_k = g_k$, and the other $n - k$ contributors are unknown. The dropout probabilities $\vec{d}$ and the parameter c are the same for both hypotheses. We now have sufficient information to compute the likelihood of the mixture data under these hypotheses and the quotient of these is equal to (summarizing by I the information $\{D_1 = g_1, \ldots, D_k = g_k\}$ on the undisputed contributors)

$$LR_{\vec{d},c}(M, g) = \frac{P_{\vec{d},c}(M|D_{k+1} = S, S = g, I)}{P_{\vec{d},c}(M|D_{k+1} \neq S, S = g, I)} = \frac{P_{\vec{d},c}(M|D_{k+1} = g, I)}{P_{\vec{d},c}(M|I)}. \tag{2.4}$$

The second equality follows since we assume that the genotypes of unrelated individuals are independent, and hence is violated when $\theta > 0$. As we have shown in [8], calculation of these likelihood ratios is exactly the same process as probabilistic inference of the donor's genotypes since

$$\begin{aligned} LR_{\vec{d},c}(M, g) &= \frac{P_{\vec{d},c}(M|D_{k+1} = g, I)}{P_{\vec{d},c}(M|I)} = \frac{P_{\vec{d},c}(D_{k+1} = g|M, I)}{P_{\vec{d},c}(D_{k+1} = g|I)} \\ &= \frac{P_{\vec{d},c}(D_{k+1} = g|M, I)}{p(g)}. \end{aligned} \tag{2.5}$$

Thus $LR_{\vec{d},c}(M, g)$ has two interpretations: first, it tells us how many more times the mixture data are likely to be found if assume that donor $k + 1$ has genotype g, compared to the a priori probability having no information on that donor; and second, it tells us how many more times it becomes likely that donor $k + 1$ has genotype g, given the mixture data, compared to the a priori probability without mixture data.

In particular, we see from (2.5) that the genotype probabilities of the searched donor are obtained from the LR and the population frequencies:

$$P_{\vec{d},c}(D_{k+1} = g|M, I) = p(g) LR_{\vec{d},c}(M, g). \tag{2.6}$$

A corollary is that we must have, as is dictated by common sense as well, that $LR_{\vec{d},c}(M, g) \le 1/p(g)$. Note also that, when we consider the hypotheses H_1 and H_2 regarding contribution of the suspect, the LR acts as factor between prior and posterior odds on these hypotheses, whereas in (2.6) it acts as factor between prior and posterior probabilities.

2.2.1. Computational considerations

As explained in [6], $LR_{\vec{d},c}(M, g)$, when considered a function of the allele frequencies, only depends on the frequencies of the observed alleles in the mixture M for this semi-continuous model. In other words, for every allele in genotype g that is unobserved in the mixture, the LR would have been the same if that allele would have been another unobserved allele. We therefore define $U = \{a \mid a \notin M\}$ as the set of alleles that were not observed in the mixture. In a computer implementation of the model the set U can