Forensic Science International: Genetics 6 (2012) 762–774

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

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

Exploratory data analysis for the interpretation of low template DNA mixtures

H. Haned [a,*], K. Slooten [a,b], P. Gill [c,d]

[a] Netherlands Forensic Institute, Department of Human Biological traces, The Hague, The Netherlands
[b] VU University Amsterdam, Amsterdam, The Netherlands
[c] Norwegian institute of Public Health, Oslo, Norway
[d] University of Oslo, Norway

ARTICLE INFO

Keywords:
Low template
Mixtures
Drop-out
Drop-in
Likelihood ratios

ABSTRACT

The interpretation of DNA mixtures has proven to be a complex problem in forensic genetics. In particular, low template DNA samples, where alleles can be missing (allele drop-out), or where alleles unrelated to the crime-sample are amplified (allele drop-in), cannot be analysed with classical approaches such as random man not excluded or random match probability. Drop-out, drop-in, stutters and other PCR-related stochastic effects, create uncertainty about the composition of the crime-sample, making it difficult to attach a weight of evidence when (a) reference sample(s) is (are) compared to the crime-sample. In this paper, we use a probabilistic model to calculate likelihood ratios when there is uncertainty about the composition of the crime-sample. This model is essentially exploratory in the sense that it allows the exploration of LRs when two key-parameters, drop-out and drop-in are varied within their plausible ranges of variation. We build on the work of Curran et al. [8], and improve their probabilistic model to allow more flexibility in the way the model parameters are applied. Two new main modifications are brought to their model: (i) different drop-out probabilities can be applied to different contributors, and (ii) different parameters can be used under the prosecution and the defence hypotheses. We illustrate how the LRs can be explored when the drop-out and drop-in parameters are varied, and suggest the use of Monte Carlo simulations to derive plausible ranges for the probability of drop-out. Although the model is suited for both high and low template samples, we illustrate the advantages of the exploratory approach through two DNA mixtures (involving two and at least three individuals) with low template components.

1. Introduction

The interpretation of DNA profiles obtained from low template DNA (LTDNA) samples has proven to be a particularly difficult problem [1,2]. LTDNA samples often comprise DNA from multiple contributors, in different quantities and in limited amounts, which cause PCR-related stochastic effects, such as drop-out (alleles in the sample that fail to PCR-amplify) and drop-in (alleles unassociated with crime-samples that are PCR-amplified) [3,4].

When a reference sample, e.g. from a suspect, is compared to a crime-sample profile, stochastic effects typically create discordances at several loci, making it impossible to use classical methods, such as random man not excluded or the random match probabilities, to report the weight of the DNA evidence. Several models have been proposed in the literature to overcome these issues, but none is in general use or are easily available (free software). They are all anchored in a likelihood ratio (LR) framework, and are traditionally classified in two categories based on the type of information they take into account: (i) continuous models, model the peak heights as continuous variables, and thus account for both the qualitative and quantitative data provided by the electropherograms (epgs) [5–7], and (ii) qualitative models that only use the list of alleles observed in a DNA profile [8–11]. Continuous models consider peak heights to be continuous random variables, and in principle, make the 'best use' of available data. However, when PCR-related stochastic effects such as drop-out and drop-in affect the sample profile (i.e. typical low-template DNA profiles), these models are less efficient because the variability of the signal is exacerbated and the uncertainty in the peak heights is difficult to assess [12]. Comparative studies have not yet been undertaken. Consequently, it is not clear yet how these models behave when applied to low template DNA (LTDNA) in practice, and there is little published on the matter of their robustness when used with these type of samples [7]. Because the utility of peak height information decreases as the amount of template decreases [13], the qualitative and continuous models must eventually converge.

It is possible to evaluate complex mixtures and account for the main stochastic effects related to LTDNA samples, namely,

* Corresponding author.
E-mail addresses: h.haned@nfi.minvenj.nl, hi.haned@gmail.com (H. Haned).

1872-4973/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2012.08.008

drop-out and drop-in, without explicitly modelling the peak heights as continuous variables. This is achieved by adopting a probabilistic model that evaluates likelihood ratios, conditioned on the probability of allelic drop-out and drop-in. Such a model has been described by Curran et al. [8] and Gill et al. [9]. The model enables the computation of LRs for DNA samples with several replicates, which may show drop-out and drop-in alleles, and with multiple contributors. Although this model falls within the qualitative category, it is more accurate to describe it as semi-continuous, since information derived from the epgs is included in the LR to account for uncertainty in the data [8]. In this paper, we improve this model by implementing three major modifications: (i) the probability of drop-out is split per contributor, (ii) the drop-out parameter can vary under the prosecution and the defence hypothesis and (iii) allele masking due to shared alleles between contributors is accounted for. The results of the modified model that we will refer to as the 'SplitDrop' model, are compared to the original 'basic' Curran model, as well as to a newly available software, LikeLTD [14], which also relies on the method described in [8]. The basic model and LikeLTD are essentially the same, but instead of exploring a range of values for these probabilities, likeLTD searches for single drop-out and drop-in estimates which maximise the likelihoods under the defence and the prosecution hypotheses. We illustrate how the SplitDrop model can be applied in practice to typical cases of DNA mixtures reported by the Netherlands Forensic Institute, and the Norwegian Institute of Public Health, and show how it can be employed as an exploratory approach to evaluate the strength of DNA evidence.

2. Theoretical considerations

2.1. The classical likelihood ratio

The classical likelihood ratio (LR) approach consists of a comparison of the likelihood of obtaining the observed DNA profiles given alternative competing hypotheses. The probability of observing the evidence E given hypothesis H, can be computed using probabilistic reasoning. The LR is usually written as:

$$\mathrm{LR} = \frac{Pr(E|H_p)}{Pr(E|H_d)} \quad (1)$$

Fig. 1 shows an example of three epgs of a crime-sample at a single locus. We want to evaluate the following hypotheses, assuming that it has exactly one contributor:

- H_p: the suspect contributed to the sample,
- H_d: an unknown person, unrelated to the suspect, contributed to the sample.

First consider the case where there is sufficient DNA in the sample for the alleles to faithfully reflect the genotype of the donor of the sample. If the observed profile matches that of the person of interest (the suspect in this case), then under H_p, the probability of observing the crime-sample profile is one, since the suspect is assumed to be the contributor. Under H_d, we assume that an unknown person is the contributor of the sample. This person, under the assumption of a single donor trace, needs to match the reference profile. In our example case A, the only 'unknown genotype' that can explain the profile is a heterozygote 9, 10. The probability of observing the evidence, conditioned on an unknown person contributing to the sample is the probability of observing the genotype in the target population. If the target population consists of the general population, unrelated to the offender, with

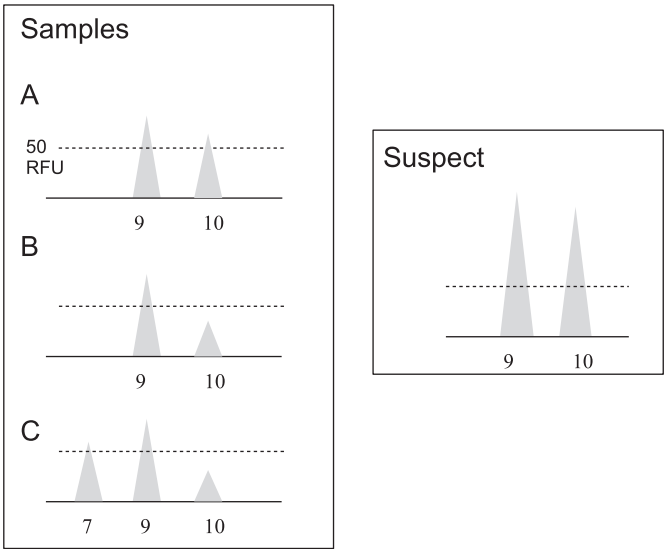

Fig. 1. Single source, single-locus examples. When the suspect is assumed to be the contributor to the samples: case A: no drop-out, no drop-in; case B: one drop-out, no drop-in; case C: one drop-out, one drop-in.

allele frequencies p_9 and p_{10} for alleles 9 and 10, then the LR in Eq. (1) is simply:

$$\mathrm{LR} = \frac{1}{2\,p_9\,p_{10}} \quad (2)$$

Let us assume now that that it is no longer certain that the observed alleles in the sample faithfully reflect the trace donor's genotype, a situation that arises in a low-template crime-sample profile. For example, certain alleles may have failed to PCR-amplify, or there could also be alleles unrelated to the contributor(s) that appear in the sample epg (allele drop-in). In the classical LR approach (unjustly ignoring the uncertainties), the probability $Pr(E|H_p)$ can be zero. This happens if the crime-sample profile cannot be explained by the suspect profile, and one way to deal with this situation is to ignore the problematic locus, and to compute a statistic for loci that do not show drop-out, drop-in or other stochastic effects. However, this approach is biased as it effectively considers evidence to be 'neutral' (LR = 1) and obviously may be very anti-conservative [15]. Models are needed however that are able to fully evaluate any hypothesis.

The Curran et al. [8] model enables unrestricted computation of likelihood ratios when PCR-related stochastic effects such as drop-out and drop-in are possible. In the following section, we illustrate how unrestricted computation of likelihood ratios is enabled when the probability of the evidence is conditioned on the probabilities of allelic drop-out and drop-in.

2.2. Likelihood ratio allowing for drop-out and drop-in

Curran et al. [8] proposed a probabilistic model that enables the evaluation of low template DNA samples. The model is based on simple principles of probabilistic theory, and only makes use of qualitative data.

Suppose n replicates, $R_1, \ldots, R_n$, have been analysed. We want to compute the LR for two competing hypotheses, H_p and H_d, which state the alternative contributors to the crime-sample. To achieve this, we need first to compute $P(R_i|H)$, where H is a hypothesis stating the number of contributors, the genotype of some of these contributors (possibly none), the probabilities of observing each