Forensic Science International: Genetics 15 (2015) 76–83

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Identifying the most likely contributors to a Y-STR mixture using the discrete Laplace method

Mikkel Meyer Andersen [a,*], Poul Svante Eriksen [a,1], Helle Smidt Mogensen [b,2], Niels Morling [b,3]

[a] Department of Mathematical Sciences, Aalborg University, Denmark
[b] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

ARTICLE INFO

Keywords:
Forensic genetics
Discrete Laplace method
Likelihood ratio
Mixture separation
Mixture deconvolution

ABSTRACT

In some crime cases, the male part of the DNA in a stain can only be analysed using Y chromosomal markers, e.g. Y-STRs. This may be the case in e.g. rape cases, where the male components can only be detected as Y-STR profiles, because the fraction of male DNA is much smaller than that of female DNA, which can mask the male results when autosomal STRs are investigated. Sometimes, mixtures of Y-STRs are observed, e.g. in rape cases with multiple offenders. In such cases, Y-STR mixture analysis is required, e.g. by mixture deconvolution, to deduce the most likely DNA profiles from the contributors.

We demonstrate how the discrete Laplace method can be used to separate a two person Y-STR mixture, where the Y-STR profiles of the true contributors are not present in the reference dataset, which is often the case for Y-STR profiles in real case work. We also briefly discuss how to calculate the weight of the evidence using the likelihood ratio principle when a suspect's Y-STR profile fits into a two person mixture. We used three datasets with between 7 and 21 Y-STR loci: Denmark ($n = 181$), Somalia ($n = 201$) and Germany ($n = 3443$). The Danish dataset with 21 loci was truncated to 15 and 10 loci to examine the effect of the number of loci. For each of these datasets, an out of sample simulation study was performed: A total of 550 mixtures were composed by randomly sampling two haplotypes, h_1 and h_2, from the dataset.

We then used the discrete Laplace method on the remaining data (excluding h_1 and h_2) to rank the contributor pairs by the product of the contributors' estimated haplotype frequencies. Successful separation of mixtures (defined by the observation that the true contributor pair was among the 10 most likely contributor pairs) was found in 42–52% of the cases for 21 loci, 69–75% for 15 loci and 92–99% for 10 loci or less depending on the dataset and how the discrete Laplace model was chosen. Y-STR mixtures with many loci are difficult to separate, but even haplotypes with 21 Y-STR loci can be separated.

1. Introduction

In some crime cases, the male part of the DNA in a stain can only be analysed using Y chromosomal markers such as short tandem repeats (Y-STRs). This may be the case in serious crime cases such as rape cases, where the male component(s) can only be detected as Y-STR profiles, because the fraction of male DNA is much smaller than that of female DNA, which can mask the male results when autosomal STRs are investigated. Sometimes, mixtures of Y-STRs are observed, e.g. in rape cases with multiple offenders. In such cases, Y-STR mixture analysis is required, e.g. by mixture separation to deduce the most likely DNA profiles from the contributors.

Until now, only few mathematical models have been offered regarding Y-STR mixture separation and calculation of the weight of the evidence by means of likelihood ratio principles. This is mainly because both applications require a method for estimating haplotype frequencies that includes estimation of the frequencies of unobserved haplotypes. Using the discrete Laplace method [1], it is possible to model the probability distribution of Y-STR haplotypes. With the discrete Laplace method, it is furthermore possible to perform cluster analysis [2].

[*] Corresponding author at: Fredrik Bajers Vej 7G, DK-9220 Aalborg East, Denmark. Tel.: +45 99408800.

E-mail addresses: mikl@math.aau.dk (M.M. Andersen), svante@math.aau.dk (P.S. Eriksen), helle.smidt@sund.ku.dk (H.S. Mogensen), niels.morling@sund.ku.dk (N. Morling).

[1] Fredrik Bajers Vej 7G, DK-9220 Aalborg East, Denmark. Tel.: +45 99408800.
[2] Frederik V's Vej 11, DK-2100 Copenhagen East, Denmark. Tel.: +45 35326212.
[3] Frederik V's Vej 11, DK-2100 Copenhagen East, Denmark. Tel.: +45 35326115.

http://dx.doi.org/10.1016/j.fsigen.2014.09.011
1872-4973/© 2014 Elsevier Ireland Ltd. All rights reserved.

Here, we focus on separating two person Y-STR mixtures (mixture deconvolution), where the Y-STR profiles of the true contributors are not present in the Y-STR reference dataset, which is often the case in real casework. We also briefly discuss how to calculate the weight of the evidence using the likelihood ratio principles of Y-STR results of two person mixtures when a suspect's Y-STR profile fits into a two person mixture.

2. Method

We used three datasets with varying numbers of Y-STR loci. The results of DYS385a/b were excluded due to ambiguity problems [3]. We also excluded profiles with intermediate alleles and null alleles as these rather rare phenomena are not yet modelled well by any mathematical model. All haplotypes consisted of loci with only one allele. The datasets used were Danes [4,5] ($n = 181$, 21 loci using the PPY23 kit excluding DYS385a/b), Somalis [6] ($n = 201$, 10 loci using the PowerPlex Y kit excluding DYS385a/b) and a compilation of 14 German populations [7] ($n = 3443$, 7 loci excluding DYS385a/b, the used kits were stated). The Danish dataset was truncated to 15 loci (corresponding to Yfiler (Thermo Fisher)) and 10 loci (corresponding to PowerPlex Y (Promega)) in order to observe the effect of varying numbers of loci. Hence, a total of five datasets were used. The loci (in order) were: DYS19, DYS389I, DYS389B (DYS389II-DYS389I), DYS390, DYS391, DYS392, DYS393, DYS438, DYS439, DYS437, DYS448, DYS456, DYS458, DYS635, YGATAH4, DYS576, DYS481, DYS549, DYS533, DYS570 and DYS643. Thus, the 15 locus Danish dataset consisted of alleles of the first 15 loci (corresponding to Yfiler), the 10 locus Danish and Somali dataset of alleles of the first 10 loci (corresponding to PowerPlex Y) and the 7 locus German dataset consisted of alleles of the first 7 loci (corresponding to the minimal haplotype [8] without DYS385a/b). We assumed that no error, drop-out, drop-in, etc., had happened.

For each of the datasets, an out-of-sample simulation study was performed: A total of 550 Y-STR mixture profiles were composed by randomly sampling two haplotypes, h_1 and h_2, from the dataset and combining these. Hence, a Y-STR trace mixture profile may look like that shown in Table 1. We then excluded one observation of h_1 and one observation of h_2 from the dataset to simulate the situation where the Y-STR profiles of the true contributors are not in the reference dataset and estimated the parameters of a discrete Laplace model [1]. We then found all 2^{m-1} possible contributor pairs for the Y-STR trace mixture, where m is the number of loci with two alleles. These contributor pairs were then ranked decreasingly according to the product of the contributors' estimated haplotype frequencies. Refer to Table 2 for details of the datasets and the simulation study.

We fitted a discrete Laplace model for 1 to 50 clusters for all mixtures except those from the German dataset, where we fitted 1 to 100 clusters. We used three different criteria to determine the optimal number of clusters in the discrete Laplace method: Akaike Information Criterium (AIC) [9], corrected AIC (AICc) [10] that penalise for the number of parameters and Bayesian Information Criteria (BIC) [11]. Let L be the likelihood function of the model, n the number of observations and k the number of parameters in the model. The optimal number of clusters in the discrete Laplace method using AIC is obtained by minimising $2k - 2\ln(L)$, where $\ln(L)$ is the natural logarithm of L. With AICc, the optimal model is obtained by minimising $AIC + 2k(k+1)/(n-k-1)$. Note that AICc converges to AIC as n gets large. With BIC, the optimal model is obtained by minimising $k\ln(n) - 2\ln(L)$.

Let $h_1 \cup h_2$ denote at trace mixture formed by mixing the alleles of h_1 and h_2 at each locus. For a trace mixture, T, let $\mathcal{H}_T = \{\{h_1, h_2\} : h_1 \cup h_2 = T\}$ denote the set of all possible contributor pairs. Let $\hat{P}(h)$ be the estimated haplotype frequency for haplotype h. We can then sort the elements of $\mathcal{H}_T$ in decreasing order by the product of the estimated haplotype frequencies, $\hat{P}(h_1)\hat{P}(h_2)$, to obtain $\mathcal{C}_T = (c_1, c_2, \ldots, c_{2^{|\mathcal{H}_T|}})$, where $c_i = \{h_{i,1}, h_{i,2}\}$. Denoted by $\hat{p}_i = \hat{P}(h_{i,1})\hat{P}(h_{i,2})$, is the estimated probability of element c_i. Then, $\hat{p}_1 \geq \hat{p}_2 \geq \ldots \geq \hat{p}_{|\mathcal{H}_T|}$ and c_1 is the contributor pair with the highest likelihood of being the source of the mixture under the assumption of independence between haplotypes.

In the simulation study, we can then find the rank of the true contributor pair by inspection of $\mathcal{C}_T$.

2.1. Software example

The Y-STR mixture separation was performed using the `disclapmix` library [12] in the statistical software package `R` [13]. A self-contained example is given in Appendix A (using data included in the `disclapmix` library).

2.2. Likelihood ratio

Let T be a Y-STR trace mixture and h_1 a haplotype. Denote by $h_2 = T \ominus h_1$ the haplotype of the unknown if assuming that h_1 is a contributor to the Y-STR trace mixture. Note, that h_2 is well defined for mixtures with at least one locus with two alleles. For each locus, if the trace mixture has two alleles, then h_2 has the allele that h_1 does not have, and if the trace mixture has only one allele, then h_1 and h_2 share this allele. Thus, for a trace mixture, T, and a haplotype h_1, there is one and only one h_2 such that $h_1 \cup h_2 = T$.

Let

$$LR(T, h) = \frac{P(T \ominus h)}{2\sum_{(h_1,h_2)\in\mathcal{H}_T} P(h_1)P(h_2)} \quad (1)$$

be the likelihood ratio of h when the evidence is the observed trace mixture, T. The number, 2, in the denominator is due to the fact that $\mathcal{H}_T$ disregards the order of the contributors.

Then, if h_S is the haplotype of a suspect, then $LR(T, h_S)$ is the LR under the hypotheses:

H_p : The suspect together with one additional contributor left the Y chromosome DNA in the crime stain.

Table 1
Example of trace from the Danish dataset (with 15 loci).

	19	389I	389II′	390	391	392	393	438	439	437	448	456	458	635	Y GATA H4
Trace	15	13	16, 17	24	10, 11	11, 13	13	11, 12	10, 11	14	19, 21	16	15	23, 24	12, 13
True h_1	15	13	16	24	11	13	13	12	11	14	19	16	15	23	12
True h_2	15	13	17	24	10	11	13	11	10	14	21	16	15	24	13

DYS prefixes for loci were omitted. Y-STR alleles in the trace are written in numeric order. The product of the frequencies of the two contributing haplotypes was the highest among all the possible contributor pairs; hence, the true contributor pair was ranked number 1. There were 8 loci with 2 alleles, meaning that there were $2^{8-1} = 128$ possible haplotype combinations when ignoring contributor order. Fitting a discrete Laplace model using the Danish dataset (15 loci) without the two haplotypes, h_1 and h_2, constituting the trace, 4 clusters were found optimal using the BIC criterium. With this model, $\hat{p}_{h_1} = 5.31 \times 10^{-7}$ (the estimated haplotype frequency of h_1) and $\hat{p}_{h_2} = 2.58 \times 10^{-7}$.