Forensic Science International: Genetics 9 (2014) 134–141

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Database extraction strategies for low-template evidence

Øyvind Bleka [a,*], Guro Dørum [b], Hinda Haned [c], Peter Gill [a]

[a] Department of Forensic Genetics, Norwegian Institute of Public Health, Oslo, Norway
[b] Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway
[c] Department of Human Biological Traces, Netherlands Forensic Institute, The Hague, The Netherlands

ARTICLE INFO

Article history:
Received 11 June 2013
Received in revised form 14 November 2013
Accepted 28 November 2013

Keywords:
Database extraction
Match score
STR mixtures

ABSTRACT

Often in forensic cases, the profile of at least one of the contributors to a DNA evidence sample is unknown and a database search is needed to discover possible perpetrators. In this article we consider two types of search strategies to extract suspects from a database using methods based on probability arguments. The performance of the proposed match scores is demonstrated by carrying out a study of each match score relative to the level of allele drop-out in the crime sample, simulating low-template DNA. The efficiency was measured by random man simulation and we compared the performance using the SGM Plus kit and the ESX 17 kit for the Norwegian population, demonstrating that the latter has greatly enhanced power to discover perpetrators of crime in large national DNA databases. The code for the database extraction strategies will be prepared for release in the R-package *forensim*.

1. Introduction

Mixtures are commonly encountered in casework, and they are often low-level which means that the profiles may be partial. This increases the complexity of carrying out database searches, since it is difficult to recognize the individual profiles based on deconvolution techniques in order to positively identify genotypes that can be extracted to interrogate national DNA databases. Often there are no suspect(s) available, or there is no match between the suspect(s) and the evidence. A way forward is to interrogate a national DNA database (also known as a "cold hit" search) using a match score as described by Chung and Fung [4] and Chung et al. [5]. The match score is a statistic which indicates the strength of evidence of the match between a profile in the database and the crime scene evidence. In this article we evaluate, via simulation, the efficiency of database extraction for two strategies, based on different match score statistics. A priori we expect that the likelihood ratio will be the best match score since it is known to be the most efficient statistic to distinguish between two hypotheses [1]. The strategies aim to extract the true donor, given that he is in the database, with a certain probability, for low-template data [3]. We simulated a virtual DNA database comprising five million random individuals (similar in size to the UK national DNA database), and carried out searches for individuals of interest by generating two-person mixtures where one of the contributors is the true donor T and the other is an unknown unrelated individual outside the database. The mixtures generated may be subject to allele drop-out in order to simulate low-template samples. If the database size is large and the interrogated evidence is a DNA mixture, it is expected that some random profiles in the database would provide high strength of evidence which may be greater than that provided by the true donor himself. In particular, we expect this to occur more often when alleles of the donor profile (G_T) have dropped out. In this article we investigated how the database extraction strategies were influenced by the number of markers in the kit by comparing the SGM Plus kit (with ten autosomal markers) with the Promega ESX 17 kit (with sixteen autosomal markers).

In the first part of the method section we compared four different statistical match scores by investigating the rank of the true donor from a database search. In the second part, the best match statistic was applied to the database search strategies and validated. The software used is available in the R-package *forensim* [7].

2. Method

In order to investigate the efficiencies of the different match scores, we first carried out a simulation procedure on a virtual database with five million individuals (based on Norwegian allele frequencies). In this procedure, we investigated the rankings of the true donors of 4000 virtual two-person DNA mixtures, when different match statistics were used and different numbers of alleles were dropped out in the true donor. We defined the best

* Corresponding author at: Department of Forensic Genetics, Norwegian Institute of Public Health, Rikshospitalet, Sognsvannsveien 20, 0372 Oslo, Norway. Tel.: +47 21077643.
E-mail address: Oyvind.Bleka@fhi.no (Ø. Bleka).

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.11.006

match score to be the one that provided the highest ranking for the true donor (i.e. the true donor is within the top ranked candidates required after the search). We evaluated two common match statistics: The Matching Allele Count (MAC) and the Likelihood Ratio (LR). We also introduced two additional statistical match scores; the *P*-value of Matching Allele Count (PMAC), and the Frequency Product (FP).

2.1. Comparing match statistics

Let one of the alleles in the genotype for one marker of an individual in database D be given as a with allele frequency p_a. Let E be the allele information of the evidence for one marker. The four match scores are defined as follows (see Appendix A.3).

1. Matching Allele Count (**MAC**): Simple count of number of matching alleles between the individual's profile and evidence E summed across all markers. An allele a is counted if a is seen in the evidence E.
2. *P*-value of Matching Allele Count (**PMAC**): The probability of observing a better set of matches (MAC per locus) than the one observed (*p*-value). This is based on the frequencies of the alleles of the matching profiles' genotype. The final match score is the negative signed product of the *p*-values across all markers.
3. Frequency Product (**FP**): For each matching allele a between the individual and the evidence E, we multiply the product $p_a(1-p_a)$. If the allele a does not match, the value one is assigned and the product remains unchanged. The final match score is the negative signed product over all markers.
4. The Likelihood Ratio value (**LR**): This match score is the likelihood ratio between the probability of evidence given the hypotheses H_p^j : 'The individual j in the database and one unknown individual contributed to the evidence' versus H_d : 'Two unknown individuals contributed to the evidence' [8]. For simplicity we name the likelihood of the observed evidence given the hypothesis H_d as l_d. A drop-out probability of 0.1, a drop-in probability of 0.05 and no inbreeding (θ-correction [2] equal to zero) were specified in the LR model. The likelihood values were calculated using the R-package *forensim* [7].

Although the values yielded by these four match scores are not directly comparable to each other, they have in common that a large value means a better match between a given candidate in the database and the evidence.

2.1.1. Example 1: illustration of the four match statistics

We illustrate rankings of individuals in a database based on different match scores with an example. The following two-person mixture evidence was given: D3S1358: (15,16), vWA: (14,15,17), D16S539: (11,12,13), D2S1338: (20,23,24), D8S1179: (11,12,13,14). A search for a possible contributor was carried out in the very small database given in Table 1. We assumed that individual 1 was the true contributor. Table 2 shows the results of the rankings when each match score was applied to every individual in the database. Rankings are shown in brackets. From this it was observed that the true contributor was top ranked with all match scores even though two alleles had dropped out.

Table 1
Shows the genotypes of the individuals in database D for five markers.

ID	D3S1358	vWA	D16S539	D2S1338	D8S1179
1	(14,15)	(14,17)	(12,13)	(24,18)	(12,14)
2	(15,14)	(15,19)	(13,11)	(18,20)	(13,13)
3	(14,18)	(15,18)	(13,11)	(23,17)	(12,13)
4	(15,15)	(15,18)	(11,8)	(18,23)	(11,14)
5	(14,18)	(15,18)	(11,12)	(17,24)	(11,13)

Table 2
Match scores for the individuals in database D based on the four match statistics, and the resulting rankings in brackets.

ID	MAC(A)	PMAC(B)	FP(C)	LR(D)
1	8 [1]	−7.29e−05 [1]	−1.81e−07 [1]	3.09e+00 [1]
2	7 [3]	−9.56e−04 [3]	−2.56e−06 [3]	2.41e−03 [5]
4	6 [5]	−2.30e−03 [4]	−4.66e−06 [5]	8.59e−03 [4]
5	7 [3]	−2.30e−04 [2]	−6.26e−07 [2]	9.83e−02 [2]
6	6 [5]	−2.35e−03 [5]	−4.46e−06 [4]	8.88e−03 [3]

2.1.2. Efficiency of the match scores

To find the best match score method, a much larger simulation comprising 4000 mixtures and a database size similar in size to the UK national DNA database was applied. For each match score, the distribution of the rank of the true donor was simulated using a database D with 5 million random genotypes and different number of allele drop-outs $n_{D.O.}$ in the true donor profile.

For each simulation $m = 1, \ldots, 4000$ carry out the following:

1. Choose individual m (index) in the database D as the true donor T with genotype G_T.
2. Drop out a fixed number $n_{D.O.}$ of randomly chosen alleles in the genotype G_T so it becomes $\tilde{G}_T$.
3. Draw independently one random individual U with genotype G_U without allele drop-out.
4. Merge the genotypes $\tilde{G}_T$ and G_U together to create the mixture evidence E.
5. For each of the four match scores the individuals in the database are sorted (based on the match scores) and the rank of the individual with index m (i.e. the true donor T) is stored.

By comparing the simulated distributions we found the best match score for further use in the probabilistic database extraction strategies.

2.2. Probabilistic database extraction strategies

We now introduce two different probabilistic strategies to extract a list of candidates from the database. By probabilistic we mean that the true donor is in the extracted list, given that he is in the database, with probability α (i.e. efficiency). The first strategy was based on "the conditional method" used for familial searching in [9], while the second strategy was inspired by the "the profile-centered method" in [9] and takes low-template data into consideration by conditioning on drop-out probabilities. While the first strategy is limited by using the LR match score, the second strategy may use any of the match scores described in the previous subsection. If the drop-out probabilities are not known for the given evidence, they must be estimated. We split the second strategy, "the low-template threshold method", into (a) and (b), where two different methods were used to estimate the drop-out probability (see Appendix A.1) to take into account the uncertainty range of the drop-out probability estimation.

Assume that a suspect database, D, of size N, and the evidence E are given. The match score was calculated for all N individuals in the database D and assigned as a decreasing sorted set of values $\{u_j\}_{j=1}^{N}$. We define D_n as the subset of D which includes the n individuals with greatest match scores.

2.2.1. The conditional method

For a given α we choose the smallest subset D_n with the criterion

$$\sum_{j \in D_n} u_j \geq \alpha \sum_{j \in D} u_j \qquad (1)$$