International Congress Series 1288 (2006) 477–483

www.ics-elsevier.com

Forensic interpretation of haploid DNA mixtures

M. Krawczak

Institut für Medizinische Informatik und Statistik Christian, Albrechts, Universität Kiel, Germany

Abstract. The mathematical concept previously introduced for the forensic interpretation of DNA mixtures using non-associated genetic markers has been adapted to the assessment of haplotypes. Such calculus is required, for example, when mitochondrial or Y-chromosomal markers are used in forensics. Two approaches exist to its practical computational implementation, involving either the inclusion–exclusion principle of probability theory or a recursion in the number of unknown contributors invoked. The former approach scales better in cases of practically relevant complexity and for sufficiently diverse markers. Simulation based upon the Y-chromosomal Haplotype Reference Database suggests that the exclusion chance of a non-contributor varies between 95% in the case of two contributors to the trace, and 70% for five contributors. These estimates are however likely to be conservative since only haplotypes known to occur in YHRD have been involved in the simulations.

Keywords: DNA mixture; Likelihood ratio; Trace matching; Y chromosome; Mitochondrion

1. Introduction

In forensics, the interpretation of a DNA mixture naturally entails the assessment of the probability that a given number of known and unknown individuals, respectively, have contributed to the DNA pool in a sample of biological material taken from a crime scene or victim [1]. With currently available molecular genetic technology, it will not generally be possible to infer the number of contributors to the DNA pool from the laboratory results alone. Rather, the information available for inference making will comprise sets of alleles, known to be present in the mixture for a particular genetic marker. Without loss of generality, we may assume that all known contributors "match" the trace, i.e. that their alleles are included in the trace for all markers. For this situation, Weir et al. [1] have provided a general framework for quantifying the evidence in favour, or against, particular hypotheses about the creation of the DNA mixture in the trace.

E-mail address: krawczak@medinfo.uni-kiel.de.

0531-5131/
doi:10.1016/j.ics.2005.10.041

2. The analytical framework

For a given marker, let V denote the set of alleles in the trace and let W denote the set of alleles that are included in the trace but which are not found in any of the known contributors. The key quantity of interest is the probability of making precisely this observation, namely a trace that contains alleles V and for which a subset W is not explained by known contributors. This probability, denoted by

$$P_n(V;W), \tag{1}$$

obviously depends upon the number n of unknown contributors. Assuming that the population of interest is in Hardy-Weinberg equilibrium, $P_n(V;W)$ can be calculated from known allele frequencies using the inclusion–exclusion principle from probability theory [1], viz

$$P_n(V;W) = f(V)^x - \sum_{T \subset W} (-1)^{\mathrm{card}(T)-1} \cdot [f(V) - f(T)]^x. \tag{2}$$

Here, $f(X)$ denotes the total frequency of alleles in set X, summation is over all non-empty subsets T of W, and x equals $2n$ for diploid and n for haploid markers.

3. Trace probabilities for haploid markers

Most of the genetic markers currently used in forensic practice have been specifically designed and selected to be statistically independent, i.e. they should show neither linkage disequilibrium nor any other type of allelic association in the population under consideration. The probability of observing a trace comprising alleles from multiple loci can be obtained simply by multiplying the marker-specific probabilities obtained from formula (2). Although this is possible for virtually all autosomal loci of practical interest, it is obviously not valid for markers on the Y chromosome or the mitochondrion. The major purpose of this article is to address the specific mathematical and inferential problems arising in the forensic interpretation of haploid DNA mixtures. For mitochondrial DNA markers, these problems may no longer arise since currently available molecular techniques including denaturing high-performance liquid chromatography (DHPLC) allow the physical separation and subsequent characterisation of the individual components of a mitochondrial DNA mixture provided they are sufficiently different [2]. In the following, I will therefore concentrate on the forensic use of Y-chromosomal variation since the length and dispersion of useful variants of the Y chromosome still render the physical separation of haplotypes impracticable. Nevertheless, Y-chromosomal markers are particularly useful in forensics because they are sensitive to male, and insensitive to female, contributions. This implies that, for crimes like prostitute murder or gang rape, the genetic identification of potential perpetrators from the DNA pool of a trace would not be obscured by female genetic material, which is usually present in much larger quantities. Furthermore, the use of Y-chromosomal markers allows evidence about the contribution of a male suspect to be obtained from the analysis of close male relatives. This circumstance, however, may also represent a disadvantage of Y-chromosomal markers; in that close male relatives of a contributor can not be excluded from being contributors themselves. One of the biggest disadvantages, however, is the formal

equivalence of the Y chromosome to a single albeit highly diverse locus. This limitation implies that the population frequency of individual allele combinations, or haplotypes, is difficult to estimate accurately from reasonably sized databases. Furthermore, the calculation of match probabilities, for which knowledge of the population frequencies is a prerequisite, is not as straightforward as in the case of single autosomal loci.

The approach by Weir et al. [1] of using the exclusion–inclusion principle for the calculation of trace probabilities has recently been extended to haploid, statistically non-independent markers [3]. This extension was based upon the notion that, generally, for m loci, and V_i and W_i defined as before at the ith locus, the probability of observing the complete DNA pool can be written as

$$P_n\left(\prod_{i=1}^{m} V_i; \prod_{i=1}^{m} W_i\right). \tag{3}$$

As mentioned above, the Cartesian product formation over allele sets in formula (3) is equivalent to product formation over probabilities in the case of independent loci. For non-independent loci, the inclusion–exclusion principle has to be applied to the whole Cartesian set products themselves. Thus, for $V_i=\{1, \ldots, s_i\}$ and $W_i=\{1, \ldots, k_i\}$ with $k_i \leq s_i$, we have

$$P_n\left(\prod_{i=1}^{m} V_i; \prod_{i=1}^{m} W_i\right) = \sum_{T\subseteq\{(i,j):i\leq m, j\leq k_i\}} (-1)^{\mathrm{card}(T)} \cdot P_n\left(\prod_{i=1}^{m} V_i \setminus \{j:(i,j)\in T\}\right). \tag{4}$$

Here, $P_n(X)$ denotes the probability that n individuals have all their alleles in set product X. If the potential contributors are unrelated, $P_n(X)$ equals $P_1(X)^n$. The base of the latter term is simply the frequency of all haplotypes included in X. Instead of using the inclusion–exclusion principle, the probability given in formula (3) can also be calculated by recursion over n [3]. It can be shown, however, that this is computationally less efficient in most practically relevant situations, namely when the number of unknown contributors invoked is larger than twice the average number of mandatory alleles in the trace [3].

At some stage, both the exclusion–inclusion principle and the recursion approach will involve nested sums of the form

$$\sum_{j_1\in X_1} \cdots \sum_{j_m\in X_m} f(j_1..j_m) \tag{5}$$

where the $f(j_1, \ldots, j_m)$ are haplotype frequencies. The efficiency of these summations can be increased in two ways. First, calculation can be conditioned upon the presence of the respective pair-wise allele combinations in the population of interest, i.e. j_k is only taken into consideration if sub-haplotype $j_{k-1}j_k$ exists. Second, it may sometimes be more efficient to take all haplotypes in a forensic database into account sequentially, and to assess whether each haplotype would be compatible with the product of the X_i or not, rather than running through all haplotypes in the set product during summation.

4. Graphical models

An alternative approach to computing probabilities for complex genetic problems, also suggested for the forensic interpretation of stain mixtures [4], is the use of graphical models [5]. The most important type of graphical models are so-called "Bayesian networks", which are directed, acyclic graphs with node set K. The nodes $k \in K$ represent random variables $Y=(Y_k)_{k\in K}$ that have a joint probability distribution function of the form

$$f(y) = \prod_{k\in K} f\left(y_k | y_{pa(k)}\right). \tag{6}$$

i.e. the nodes k are statistically independent given the "parents", pa(k), of each node. A Bayesian network suitable for the analysis of trace mixtures is given in Fig. 1. Mortera et al. [4] noted, however, that this representation of the DNA mixture problem by a graphical model is computationally very inefficient. If the nodes representing haplotypes of unknown contributors, Y_{U1} to Y_{U3}, are assumed to be independent, then computation of the probability in (3) using the Bayesian network from Fig. 1 is even less efficient than recursion over the number of unknown contributors n. However, graphical models are more flexible in practical terms because they allow a variety of hypotheses to be included in one model. This is reflected in Fig. 1 by the possible introduction of an extra latent node for quantity n. This node would take values according to a particular prior distribution if such a prior distribution is indeed known (for a detailed discussion, see below). Furthermore, the node representing the suspect haplotype Y_S can be turned from an observable into a latent variable, and linked to another latent node Y_{RS} representing the Y-chromosomal haplotype of a close male relative. In any case, if such flexibility is not needed, following the approach of Weir et al. [1] would be computationally superior to the use of graphical models. This is particularly true since an alternative and potentially more efficient graphical model suggested by Mortera et al. [4] cannot easily be transferred to the Y-chromosomal situation as it would require a number of latent nodes that is proportional to the number of known haplotypes. Even for the small number of Y-chromosomal

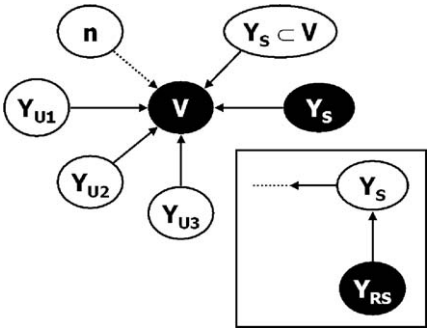

Fig. 1. Graphical model of a Y-chromosomal DNA mixture. Latent random variables are depicted by open circles, observable variables are blackened. The dotted arrow connects an optional latent node for the unobserved number, n, of unknown contributors to the network. In the inset, an extension is depicted whereby the suspect node Y_S becomes latent and is linked to the observed haplotype, Y_{RS}, of a male relative.

microsatellite markers currently used in forensics [6], this leads to a prohibitively high complexity of the resulting model.

5. Unknown number of unknown contributors

So far we have assumed that the number of unknown contributors to a trace is known exactly. In practice, however, this will not generally be the case. One way around this problem would be to replace $P_n(V;W)$ by the aggregate $f(V)$ that takes only the match between the suspect and trace into account. Inference making in forensics has generally been accepted to be based upon the likelihood ratio $L(E)$ of the prosecutor and defence hypothesis, given the evidence E. In the case of a mixture trace, these hypotheses usually entail different numbers of known and unknown contributors, so that

$$L(E) = \frac{P_{n_p}(V; W_p)}{P_{n_d}(V; W_d)} \tag{7}$$

If only the match itself is considered, the likelihood ratio would simplify to

$$L(E) = \frac{1}{f(V)} \tag{8}$$

which does not depend on n anymore. However, a lot of information, specifically about the genotype of the suspect, is lost, which implies that the mere evaluation of a match according to formula (8) would be extremely inefficient and therefore not practically useful. If prior probabilities for n were known, then the nominator and denominator in formula (7) would turn into weighted sums of trace probabilities. However, unless the number of unknown contributors is known from eyewitnesses, it is difficult to imagine how reliable priors could be provided by anybody involved in a particular case. In order to solve this problem, Brenner et al. [7] and Lauritzen and Mortera [8] proposed the calculation of an upper limit for the denominator in formula (7), thereby obtaining a lower limit for the respective likelihood ratio. However, these authors assumed that the prosecution hypothesis would invoke a fixed number of unknown contributors whereas the defence would consider varying values of n. It is not however clear why this should generally be true. It would also appear plausible that the defence hypothesis, in place of the suspect, merely invokes one more unknown contributor than the prosecution, i.e.

$$L(E) = \frac{P_n(V; W \setminus S)}{P_{n+1}(V; W)} \tag{9}$$

where S is the set of suspect alleles. If n is allowed to vary in formula (9), a useful lower limit for $L(E)$ may no longer exist since the likelihood ratio will often be a decreasing function of n that converges to zero, i.e. the more unknown contributors are invoked the less incriminating is a match. Therefore, it is not generally possible to minimise $L(E)$ as an act of courtesy to the suspect. Instead, a coherent solution can be provided if the philosophy of the Neyman-Pearson theorem [9], which is reflected in

the widely accepted use of the likelihood ratio as a decision making criterion in "simple" forensic cases, is also followed for composite hypotheses. In this case,

$$L(E) = \frac{\max_i P_i(V; W_p)}{\max_j P_j(V; W_d)} \tag{10}$$

should be used for decision making. In other words, if the prosecution and the defence put forward different but non-overlapping sets of hypotheses, then the most logically coherent way in which a decision should be made, on the basis of some evidence E, would allow both sides to maximise the probability of E over their specific sets of hypotheses. There is no reason why the two mutually exclusive hypotheses chosen in formula (10) should be in any way be linked (as, for example, by the same number of unknown contributors).

6. Practical use of Y-chromosomal markers

In terms of their practical use, one of the major problems of haploid markers is that the highly diverse spectrum of allele combinations, likely to be present in a population as a whole, is difficult to capture by reasonably sized databases. This implies that many rare haplotypes are probably first seen when they occur in the context of an actual crime case. If a suspect matches the Y-chromosomal profile of a trace for which there is no reason to believe that it represents a mixture of different male contributions, Dawid and Mortera [10] rightly maintained that the probability of a match under the defence hypothesis is approximately equal to the inverse of the database size if the suspect/trace haplotype is not found in the database. With DNA mixtures, however, the situation may be more complicated. For example, all those involved in a case may have agreed upon a certain number of unknown contributors being correct, but the number of haplotypes present in a database that need to be invoked in order to produce the trace in question turns out to be higher than this. This may either lead to a situation where some evidence that is independent of the genetic finding needs to be revisited, and its impact upon the case revised, or the DNA expert is asked for a guess of the match probability assuming the existence of hitherto not observed haplotypes. The haplotype frequencies necessary for the latter calculations can potentially be derived by extrapolation, exploiting the molecular relationship between the haplotypes present in a database [11], but this method has not yet been validated systematically nor has it been adopted to the simultaneous extrapolation of more than one haplotype frequency.

In any case, simulations based upon the Y Haplotype Reference Database [6] have shown that the median likelihood ratio for a suspect and an additional unknown trace donor is about 40:1 if the suspect DNA is in the trace, and 10:1 if the suspect DNA is not in the trace [3]. The exclusion chance of a non-donor is approximately 95%. These simulations have taken the seven microsatellites comprising the so-called "core haplotype" set into account [6]. The median likelihood ratio drops to 2:1 for a non-donor, and to 5:1 for donors, if five or four unknowns have contributed to the trace, respectively. In this case, the exclusion chance of a non-donor is approximately 70%. These estimates may appear sobering at first glance, but they are likely to be very conservative. This is because many low frequency haplotypes occurring in the German population had not been

included in the simulations, and database counts tend to overestimate haplotype frequencies [3].

7. Conclusions

In summary, the coherent interpretation of haploid DNA mixtures in a forensic context can follow the same principles as derived earlier for statistically independent, autosomal systems. Specific problems may arise, however, from the formal equivalence of haploid genomes to single highly diverse loci. In order to maintain the potential of haploid DNA systems for solving otherwise intractable forensic cases, national and international collaborations are thus upon to establish appropriate databases fit to this task.

References

[1] B.S. Weir, et al., Interpreting DNA mixtures, J. Forensic Sci. 42 (2) (1997) 213–222.
[2] P.B. Danielson, et al., Separating human DNA mixtures using denaturing high-performance liquid chromatography, Expert Rev. Mol. Diagn. 5 (1) (2005) 53–63.
[3] A. Wolf, et al., Forensic interpretation of Y-chromosomal DNA mixtures, Forensic Sci. Int. 152 (2–3) (2005) 209–213.
[4] J. Mortera, A.P. Dawid, S.L. Lauritzen, Probabilistic expert systems for DNA mixture profiling, Theor. Popul. Biol. 63 (3) (2003) 191–205.
[5] S.L. Lauritzen, N.A. Sheehan, Graphical models for genetic analyses, Stat. Sci. 18 (4) (2003) 489–514.
[6] L. Roewer, et al., Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes, Forensic Sci. Int. 118 (2–3) (2001) 106–113.
[7] C.H. Brenner, R. Fimmers, M.P. Baur, Likelihood ratios for mixed stains when the number of donors cannot be agreed, Int. J. Leg. Med. 109 (4) (1996) 218–219.
[8] S.L. Lauritzen, J. Mortera, Bounding the number of contributors to mixed DNA stains, Forensic Sci. Int. 130 (2–3) (2002) 125–126.
[9] F. Taroni, et al., Evaluation and presentation of forensic DNA evidence in European laboratories, Sci. Justice 42 (1) (2002) 21–28.
[10] A.P. Dawid, J. Mortera, Coherent analysis of forensic identification evidence, J. R. Stat. Soc., B 58 (2) (1996) 425–443.
[11] L. Roewer, et al., A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males, Forensic Sci. Int. 114 (1) (2000) 31–43.