Forensic Science International: Genetics 16 (2015) 71–76

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Inclusion probability with dropout: An operational formula

E. Milot [a,b,*], J. Courteau [c], F. Crispino [a,b], F. Mailly [d]

[a] Département de chimie, biochimie et physique, Université du Québec à Trois-Rivières, 3351 boul. des Forges, CP 500, Trois-Rivières, QC, Canada G9A 5H7
[b] Centre international de criminologie comparée, Université du Québec à Trois-Rivières, 3351 boul. des Forges, CP 500, Trois-Rivières, QC, Canada G9A 5H7
[c] Groupe de recherche PRIMUS, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, 3001, 12e Avenue N., Sherbrooke, QC, Canada J1H 5N4
[d] Laboratoire de sciences judiciaires et de médecine légale, Ministère de la sécurité publique du Québec, 1701, rue Parthenais, Montréal, QC, Canada H2K 3S7

ARTICLE INFO

Article history:
Received 7 February 2014
Received in revised form 6 November 2014
Accepted 26 November 2014

Keywords:
Inclusion probability
Dropout
Number of contributors
Forensic mixture
Low-template DNA
RMNE

ABSTRACT

In forensic genetics, a mixture of two or more contributors to a DNA profile is often interpreted using the inclusion probabilities theory. In this paper, we present a general formula for estimating the probability of inclusion (*PI*, also known as the RMNE probability) from a subset of visible alleles when dropouts are possible. This one-locus formula can easily be extended to multiple loci using the cumulative probability of inclusion. We show that an exact formulation requires fixing the number of contributors, hence to slightly modify the classic interpretation of the *PI*. We discuss the implications of our results for the enduring debate over the use of *PI* vs likelihood ratio approaches within the context of low template amplifications.

1. Introduction

The analysis of mixtures from low template (LT) DNA profiling is opening a new era in forensic genetics by providing an opportunity to extract more information than ever relevant to judiciary casework from crime scene traces. At the same time, it raises major probabilistic challenges in the evaluation of the evidentiary weight of genetic profiles by necessitating the assessment of alleles potentially present but below the analytical threshold, the so-called "dropouts".

Two major schools of mixture interpretation have been cohabiting for years in the forensic science community [1]: inclusion probability (*PI*) theory, also known as the "random man not excluded" (RMNE) approach, and likelihood ratios (*LR*). In brief, the former provides a measure of how inclusive a mixture is by estimating the proportion of a relevant population expected to have genotypes such that these individuals cannot be excluded as possible contributors to a mixture DNA profile. The second uses the same mixed profile to evaluate two or more competing hypotheses about the source of a trace. A debate has run over the merits of each approach [2–4] but consensus is growing over the superiority of LR to deal with mixture data in general and dropouts in particular, and for the evaluation of evidentiary weight in the court [5,6].

Nevertheless, the *PI* approach can serve for investigative purposes (i.e. forensic intelligence) such as to evaluate the power of discrimination or "quality" of a mixture prior to the collection of a suspect profile [2,7], especially in situations where no known profiles can be assumed to be present. Thus, it may help the investigator focus his efforts on the most useful evidence. The PI can also serve to decide whether a mixture should be searched against a crime scene or convicted offender databank [2].

For evidentiary purposes, the Scientific Working Group on DNA Analysis Methods (SWGDAM) guidelines recommend the removal of loci that exhibit peaks below the stochastic level for cumulative probability of inclusion (*CPI*) calculations prior to comparison with a suspect's profile [8], unless higher RFU alleles can be interpreted as a distinct group, in which case a lab could calculate a restricted *CPI* using only these alleles (SWGDAM 2010, Sections 4.6.3 and 5.3.5) [8]. However, it has been reported that a widespread practice was to exclude from the *CPI* calculations loci for which a suspect shows discordant alleles. This may not always be conservative and may produce evidence prejudicial to suspects [9]. While it is true that many advanced statistical tools for *LR* calculations are becoming available for such complex circumstances, a proper understanding and efficient use of this approach implies much more than implementing new software within a laboratory. The effort and time required to properly train analysts should not be

* Corresponding author at: Département de chimie, biochimie et physique, Université du Québec à Trois-Rivières, 3351 boul. des Forges, CP 500, Trois-Rivières, QC, Canada G9A 5H7. Tel.: +1 819 376 5011x4397.
E-mail addresses: emmanuel.milot@uqtr.ca (E. Milot), josiane.courteau@usherbrooke.ca (J. Courteau), frank.crispino@uqtr.ca (F. Crispino), france.mailly@msp.gouv.qc.ca (F. Mailly).

http://dx.doi.org/10.1016/j.fsigen.2014.11.023
1872-4973/

underestimated. Moreover, there is still some resistance to the *LR* framework within the judicial system [1]. Thus, many labs still use the *PI* both for evidential and investigative purposes and until the use of *LR* becomes more widespread, they are left with few options to deal with dropouts.

It may seem problematic to develop a *PI* that accounts for dropouts because any allele from the population genetic pool could have been present in the mixture before dropping out, hence, in theory no "random man" should be excluded at all. Van Nieuwerburgh et al. [7] proposed a *PI* formulation that allows for dropout occurrence. The method requires the user to specify the number x of dropouts assumed to have occurred on a given mixture and then considers as not excluded any genotype in the population matching that restriction. Thus, it makes no assumption about dropout rates, which has the advantage of avoiding their difficult estimation. However, Van Nieuwerburgh et al.'s *PI* is unduly conservative when $x > 0$ (Tables 1 & 2). This is because the inclusion of a given genotype with x discordant allele(s) is done only on the basis of the frequency of the latter in the population rather than on the joint probability that discordant alleles occurred in the pool of contributor genotypes in first place and then dropped out.

The above discussion underscores another major issue with current *PI* calculations in that they imply a post-hoc interpretation of mixtures to assess whether, how many or at what loci dropouts did actually occur, an error-prone process that may be hard to justify in court. Therefore, one "paints the target around the arrow" [10]. Ideally, a *PI* accounting for dropouts should come as near as possible to the value that would be obtained if it were calculated with the standard formulation (see Section 2) when all alleles are visible (no dropouts). Moreover, it would not rely on post-hoc evaluation of dropouts that may have occurred. An exact formulation necessitates some knowledge about both dropout rates and the probability that alleles were present in the trace (or equivalently, in the pool of contributor genotypes) before dropping out. No such formulation has yet been proposed, likely because developers of statistical tools incorporating dropout probabilities mostly adopt the *LR* approach. Here we develop such a formulation and show that it requires fixing the number of contributors to a trace. Therefore, the interpretation of the *PI* slightly changes and we discuss the implications of this. Nevertheless, this formulation constitutes a more rigorous solution to deal with dropouts than alternative *PI* methods proposed thus far.

2. Methods

2.1. Some notations and properties of the probability theory

Let A_1, A_2, …, A_N the N distinct alleles represented in the population at a specific locus and let p_1, p_2, …, p_N the corresponding alleles frequencies or probabilities. Let C the set of N_C visible distinct alleles from a mixture of *NCo* contributors. Without loss of generality, we can assume that $C = \{A_1, \ldots, A_{N_C}\}$. Let G_c be the set of all distinct alleles of the *NCo* contributors.

Table 1
Frequency of each alleles for three loci used in Van Nieuwerburgh et al.'s [7].

Allele	Locus 1	Locus 2	Locus 3
A_1	0.18	0.15	0.12
A_2	0.19*	0.15	0.12
A_3	0.20*	0.16*	0.13*
A_4	0.21*	0.17*	0.14*
A_5	0.22	0.18*	0.15*
A_6	–	0.19	0.16
A_7	–	–	0.18

The asterisk (*) means that the allele is observed in the evidence profile.

Let $Pr(A)$ designates the probability of an event A. Probability theory states that $Pr(A \cap B) = Pr(A|B)Pr(B)$ and, more generally, $Pr(A) = \sum_n Pr(A \cap B_n) = \sum_n Pr(A|B_n)Pr(B_n)$ where $\{B_n\}$ is a partition of the sample space. This property refers to the *law of the total probability*.

2.2. General expression for one locus

The exclusion probability (*PE*) is the probability that a random man (or woman) would be excluded at a focal locus. Then $PE = 1 - PI$, where *PI* is the inclusion probability or the RMNE probability – the random man not excluded. A general expression is

$$PI = \sum_{i=1}^{N(N+1)/2} Pr(g_i)\gamma_i \tag{1}$$

where $Pr(g_i)$ is the probability that an individual chosen at random from the population is of genotype g_i at the locus, γ_i is the probability that the distinct alleles of g_i (denoted g_i') are included in G_c and $N(N+1)/2$ is the total number of distinct genotypes made from N distinct alleles.

Under Hardy–Weinberg (HW) equilibrium, $Pr(g_i) = \phi p_{i,1} p_{i,2}$, where $p_{i,1}$ and $p_{i,2}$ are the frequencies of the first and second alleles ($A_{i,1}$, $A_{i,2}$) of g_i, respectively, and $\phi = 2$ if the individual is heterozygous (i.e. $A_{i,1} \neq A_{i,2}$) and $\phi = 1$ otherwise. Since $C = G_c$ in the absence of dropout

$$\gamma_i = \begin{cases} 1 & \text{if } g_i' \subset C \\ 0 & \text{otherwise} \end{cases}$$

and Eq. (1) reduces to the classical formula

$$PI = \sum_{i=1}^{N_C(N_C+1)/2} \phi\, p_{i,1}\, p_{i,2} = \left(\sum_{j=1}^{N_C} p_j\right)^2 \tag{2}$$

2.3. Modelling dropouts

In the presence of dropouts, we need to include not only all genotypes compatible with C, but also all genotypes not in C due to the occurrence of dropouts on the mixture. These genotypes are those g_i that satisfy:

$$g_i' \subset G_c; g_i' \not\subset C$$

Note that the first condition does not imply that the genotype g_i is one of the mixture's contributors but implies instead that its alleles are compatible with the mixture. The term $Pr(g_i)$ in Eq. (1) remains the same, but the problem is to derive a mathematical expression for γ_i.

Lets recall that γ_i is the probability that the distinct alleles of g_i (g_i') are included in G_c. When dropouts are possible, all we know from G_c is the visible subset C ($C \subset G_c$). Let $D_k = G_c \setminus C$ be the set of dropout alleles in G_c distinct from C (then $C \cap D_k = \varnothing$ and $C \cup D_k = G_c$). This is equivalent to saying that D_k represents one of the K possible events (or outcomes) of dropout alleles forming the invisible part of the mixture. Then, conditioning on the invisible alleles (alleles in dropout) and by summing over all possible sets of invisible alleles (law of the total probability (see Section 2.1)), we have

$$\begin{aligned} \gamma_i = Pr(g_i' \subset G_c) &= \sum_{k=0}^{K} Pr(g_i' \subset G_c | D_k) Pr(D_k) \\ &= \sum_{k=0}^{K} Pr(g_i' \subset C \cup D_k | D_k) Pr(D_k) \end{aligned} \tag{3}$$