Inclusion Probability is a Likelihood Ratio: Implications for DNA Mixtures

Mark W. Perlin,* PhD, MD, PhD, Cybergenetics, Pittsburgh, PA USA

ABSTRACT

There has been much discussion recently amongst forensic scientists about the relative merits of inclusion and likelihood ratio (LR) methods for interpreting DNA mixtures. Advocates for the probability of inclusion (PI; also termed CPI, RNME or CPE) approach contend that it is a simpler statistic that is easier to explain in court. LR enthusiasts rejoin that theirs is a more informative method that preserves more of the identification information present in the DNA data. The debate implicitly assumes that there is some essential difference between PI and LR, suggesting that each perspective should be understood and evaluated on its own merits.

In fact, there are many different LR statistics for DNA mixture interpretation. And PI happens to be just one of them. However, amongst all currently used LRs, the PI version does have a special distinction – it is the least informative.

Recognizing that PI is just another LR has important consequences for forensic science practice.
1. The current PI vs. LR controversy can be finally put to rest.
2. Inclusion efficacy can be measured in terms of how well it preserves the data's identification information. The logarithm of the LR is a standard information measure, and PI is a LR, so this assessment is easily accomplished.
3. The inclusion method can be supported in court based on its scientific status as a valid LR.
4. The PI statistic can be better understood through the inclusion likelihood function used in its LR construction.
5. The relevance of PI can be challenged on particular DNA evidence by examining the appropriateness of its (inclusion likelihood) modeling assumptions for that data.

In this poster, we show by construction that PI is a LR. We first describe the inclusion likelihood function, and see how it naturally explains binary allele data. We next use Bayes theorem to form the inclusion genotype, represented by its probability mass function (pmf). Using an easily understood form of the LR (genotype probability gain), we then insert the inclusion genotype pmf into this LR expression to obtain the standard PI statistic. Having thus derived the PI as a LR, we then discuss what this result means for DNA mixture interpretation.

The poster visually explains the underlying concepts to the forensic practitioner, and uses no mathematics besides basic probability. We focus primarily on the forensic science implications of PI actually being a LR. This proper scientific foundation for PI may invite a re-examination of some prevalent DNA mixture interpretation practices.

MIXTURE DATA

A DNA mixture occurs when there is more than one contributor to an evidence sample (Figure 1). Mixtures commonly arise in sexual assault, homicide and property crime biological evidence. To fully account for the quantitative STR data, a likelihood function must examine a combination of allele pairs that forms a *continuous* pattern (Figure 2) that can be compared with peak heights (1, 2).

A *binary* DNA assay would instead produce all-or-none data indicating the presence or absence of an allele. Thus, a simple qualitative approximation might be to ignore peak height data (Figure 3), and just check for allele pair inclusion (3, 4). This simplifying interpretation procedure is easier for a human analyst to perform and explain (5), although it can discard considerable identification information (typically a LR factor of a million) (6). We use this binary "probability of inclusion" (PI) method to illustrate how a LR match statistic is constructed from a genotype probability distribution.

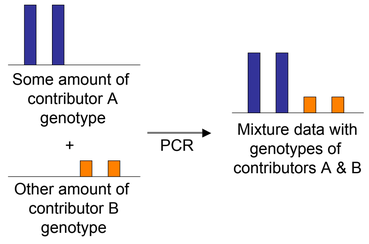


Figure 1. Forming DNA mixture data.

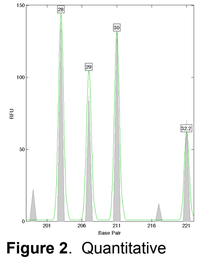


Figure 2. Quantitative likelihood function compares a continuous proposed genotype mixture pattern (gray triangles) with the observed quantitative peak height data (green curve).

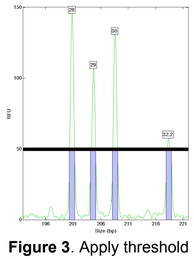
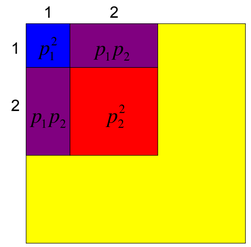


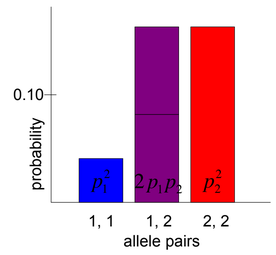
Figure 3. Apply threshold (black bar) to observed quantitative data (green curve) and ignore peak height information. This produces all-or-none allele events (blue bars).

PRIOR PROBABILITY

By Bayes theorem, we can construct a posterior pmf $q(x)=\Pr\{Q=x|d_Q\}$ for the questioned genotype Q by defining a prior probability and likelihood function (Table 1). Before we see any data, our *prior* belief $\pi_Q(x)=\Pr\{Q=x\}$ is that the probability of seeing an allele pair x is its population frequency $r(x)$, from the product rule (Table 1, column π_Q)

$$r(x)=\begin{cases} p_i^2 & x=i,i \\ 2p_ip_j & x=i,j\ (i\neq j) \end{cases}$$

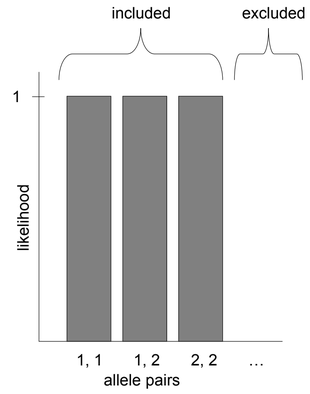




LIKELIHOOD FUNCTION

Examining the data requires a *likelihood* function $\lambda_Q(x)=\Pr\{d_Q|Q=x\}$. This likelihood describes the probability of observing the data d_Q when assuming a particular genotype value x. The inclusion method applies a preset threshold to mixture peak data to determine an inclusion set I containing K "allelic peaks". The qualitative inclusion likelihood function λ_Q is then the all-or-none rule (Table 1, column λ_Q)

$$\lambda_Q(x)=\Pr\{d_Q|Q=x\}=\begin{cases} 1 & \text{both alleles of } x \text{ are in inclusion set } I \\ 0 & \text{otherwise} \end{cases}$$



	allele pair x	prior $\pi_Q(x)$	likelihood $\lambda_Q(x)$	product $\lambda_Q(x)\cdot\pi_Q(x)$	posterior $q(x)$
included	1,1	p_1^2	1	p_1^2	$p_1^2/(p_1+p_2+\ldots+p_K)^2$
	1,2	$2p_1p_2$	1	$2p_1p_2$	$2p_1p_2/(p_1+p_2+\ldots+p_K)^2$
	…	$r(x)$	1	$r(x)$	$r(x)/(p_1+p_2+\ldots+p_K)^2$
	K,K	p_K^2	1	p_K^2	$p_K^2/(p_1+p_2+\ldots+p_K)^2$
excluded	$1,K+1$	$2p_1p_{K+1}$	0	0	0
	…	$r(x)$	0	0	0
	N,N	p_N^2	0	0	0
				$p_1^2+2p_1p_2+\ldots+p_K^2$ $=(p_1+p_2+\ldots+p_K)^2$	1

Table 1. Inferring the posterior genotype probability distribution, using an inclusion likelihood function.

POSTERIOR PROBABILITY

The *posterior* probability of allele pair of x after having seen the data is proportional to the product of likelihood and prior. The inclusion approach produces a posterior probability that is proportional to the prior for included allele pairs, having a data normalization constant δ_Q (Table 1, column $\lambda_Q\cdot\pi_Q$)

$$\begin{aligned}\delta_Q &= \sum_{x\in G}\lambda_Q(x)\cdot\pi_Q(x) \\ &= \sum_{x\in I}\pi_Q(x) \\ &= p_1^2+2p_1p_2+\ldots+p_K^2 \\ &= (p_1+p_2+\ldots+p_K)^2\end{aligned}$$

Dividing each likelihood-prior product by this data marginal constant δ_Q forms a posterior probability distribution (Table 1, column $q(x)$; Figure 4)

$$q(x)=\frac{\lambda_Q(x)\cdot\pi_Q(x)}{\delta_Q}=\frac{\lambda_Q(x)\cdot r(x)}{\delta_Q}=\begin{cases} r(x)/(p_1+p_2+\ldots+p_K)^2 & x \text{ included in } I \\ 0 & \text{otherwise} \end{cases}$$

that sums to one since $\sum_{x\in I}r(x)=(p_1+p_2+\ldots+p_K)^2$.

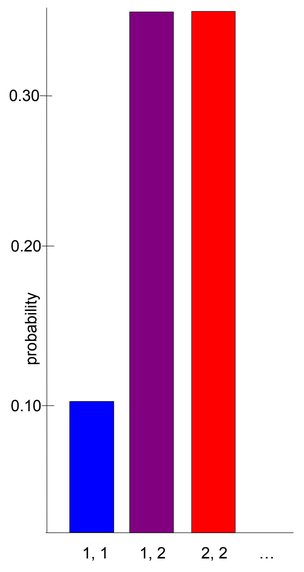


Figure 4. Posterior genotype probability distribution for inclusion is just the prior distribution, stretched up so that the included allele pair probabilities sum to one.

LIKELIHOOD RATIO

The LR information gain can be written as the ratio $q(x)/r(x)$ of posterior genotype probability to its prior at the unique allele pair x of a suspect reference, by a well-known LR posterior-to-prior form (7). For an included allele pair x, we can substitute for numerator posterior probability $q(x)$ the expression derived in Table 1, obtaining

$$LR=\frac{q(x)}{r(x)}=\frac{r(x)/(p_1+p_2+\ldots+p_K)^2}{r(x)}$$

We next cancel out the common prior population probability $r(x)$ from both numerator and denominator. The LR information gain (when using an inclusion likelihood function) is therefore the familiar PI match rarity statistic (8)

$$LR=\frac{1}{(p_1+p_2+\ldots+p_K)^2}$$

based on included allele frequencies.

REFERENCES

1. Gill P, Sparkes R, Pinchin R, Clayton TM, Whitaker JP, Buckleton J. Interpreting simple STR mixtures using allele peak area. Forensic Sci Int. 1998;91:41-53.
2. Perlin MW, Szabady B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. J Forensic Sci. 2001;46(6):1372-7.
3. Budowle B, Onorato AJ, Callaghan TF, Manna AD, Gross AM, Guerrieri RA, et al. Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. J Forensic Sci. 2009;54(4):810-21.
4. SWGDAM. Short Tandem Repeat (STR) interpretation guidelines (Scientific Working Group on DNA Analysis Methods). Forensic Sci Commun (FBI). 2000 July;2(3).
5. Buckleton J, Curran J. A discussion of the merits of random man not excluded and likelihood ratios. Forensic Sci Int Genet. 2008;2(4):343–8.
6. Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. Validating TrueAllele® DNA mixture interpretation. Journal of Forensic Sciences. 2011;56(November):in press.
7. Essen-Möller E. Die Biesweiskraft der Ähnlichkeit im Vater Schaftsnachweis; Theoretische Grundlagen. Mitteilungen der anthropologischen Gesellschaft in Wien. 1938;68(9-53).
8. Evett IW, Weir BS. Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists. Sunderland, MA: Sinauer Assoc, 1998.

Cybergenetics