Comparing forensic hypotheses from PCR results in cases involving mixtures of body fluids

Mike Steel[1] and **Michael Taylor**[2]
[1]*Biomathematics Research Centre*
University of Canterbury,
Christchurch, New Zealand
[2]*Evidence Assessment Unit,*
Environmental Science and Research Ltd.
27 Creyke Road, Ilam, Christchurch, New Zealand

No. 133 November, 1995

SUMMARY

Likelihood ratios provide a convenient and widely-used measure for assessing the relative support for forensic hypotheses given certain evidence. We extend earlier work to provide techniques for computing these ratios when a crime sample provides a profile of (multiple locus) genetic markers of mixed origin. Generic formulae are provided, illustrated with an example, and some extensions are discussed briefly.

Key Words: Polymerase chain reaction, DNA profiling, likelihood ratio, multiple locus profiles, Bayesian, evidence, forensics.

1. Introduction

The Polymerase Chain Reaction (PCR) has become established as a method for routine DNA typing in many forensic laboratories. The advantages of the method are well known and include additional sensitivity. Because of this, cases which previously had insufficient DNA for analysis are now yielding results from PCR tests.

This extra sensitivity leads to increased chances of detecting mixtures of body fluids. This is particularly true for sexual assault cases. For example, current methods of preferential extraction cannot guarantee complete separation of the male and female DNA. With the less sensitive RFLP method this has seldom been a problem, but the additional sensitivity of the PCR method accentuates this limitation.

Furthermore, PCR methods give only a limited amount of concentration information which is often insufficient to unambiguously associate alleles. Consequently, the interpretation of mixtures can be complex. In most sexual assault cases it must be assumed that observed alleles may be paired in a variety of combinations and could come from male or female sources. In fact some laboratories avoid interpreting cases that clearly involve a mixture. This represents a significant limitation of the overall method.

Traditional methods for computing coincident frequencies have limitations for the interpretation of results from mixtures. However Bayesian based methods for assessing evidential significance are ideally suited to such cases (Evett et al. 1991). Bayesian arguments provide a powerful tool for assessing support for hypotheses in the light of new evidence. For example, suppose one is comparing two hypotheses H_1 and H_2, given evidence E. Provided these two hypotheses are sufficiently well defined one can often compute the probability of the evidence E arising under each of these two hypotheses. Let $\mathbb{P}[E|H_1]$ and $\mathbb{P}[E|H_2]$ denote these two probabilities, and let LR denote their ratio; that is:

$$LR = \frac{\mathbb{P}[E|H_1]}{\mathbb{P}[E|H_2]}$$

LR is called a *likelihood ratio*. Given prior probabilities $\mathbb{P}[H_1]$ and $\mathbb{P}[H_2]$ for H_1 and H_2 (that is, probability assessments arrived at before the evidence E is taken into account), the number LR provides a simple way of determining how the **ratio** of these two prior probabilities changes once E is considered.

More precisely, by Bayes' Theorem,

$$\frac{\mathbb{P}[H_1|E]}{\mathbb{P}[H_2|E]} = LR \times \frac{\mathbb{P}[H_1]}{\mathbb{P}[H_2]}$$

where $\mathbb{P}[H_i|E]$ is the (posterior) probability of H_i given E (for $i = 1, 2$).

Thus, LR is the factor that the prior probability ratios (of H_1 to H_2) are multiplied by to obtain the posterior probability ratios (of H_1 to H_2) when the evidence E is taken into account. For an informal account of Bayesian methods in science, see Howson, C. and Urbach, (1991). For applications in forensic science, see Evett et al. (1991), Aitken (1995), and Robertson and Vignaux (1995).

The calculation of likelihood ratios in forensic cases involving genetic mixtures from different individuals can be quite complex, and it is easy to miss certain combinations of genotypes that would produce the observed evidence E. Here we provide a systematic approach to this problem, and give exact formulae for computing the likelihood ratios (Section 2). An independence assumption on the loci reduces the calculation to considering each locus separately. In Section 3 we provide a formula to count the possible combinations of genotypes that could produce E under each hypothesis - such a formula can be a useful check that one has not omitted considering cases (for example, by a programming error in implementing the formulae in Section 2). In Section 4 we give an example of the use of the formulae.

2. Generic formulae

For a given locus, a *genotype* g, consists of a pair of alleles, and as one does not distinguish between the order of these alleles, it is convenient to write g as the unordered pair ab. If $a = b$, then g is *homozygous*, otherwise g is *heterozygous*. If a locus has $r > 1$ alleles, exactly $r(r+1)/2$ distinct genotypes are possible. For example, a locus with three alleles a, b, c, has six possible genotypes aa, bb, cc, ab, ac and bc.

For genotype g, let f_g denote the frequency of g in the population P from which it is drawn. For example, if the population P is in Hardy-Weinberg equilibrium we have:

$$f_g = \begin{cases} f(a)^2, & \text{if } g = aa. \\ 2f(a)f(b) & \text{if } g = ab, a \neq b. \end{cases}$$

where $f(a), f(b)$ are the allelic frequencies of a, b in P, respectively.

In this paper we do not require P to be in Hardy-Weinberg equilibrium since we will always work with estimates of genotype frequencies explicitly (these always determine the allelic frequencies but not conversely). For computational similicity we do invoke one mild assumption, namely, that P is at least moderately large (at least several hundred).

Suppose we are using $K \geq 1$ genetic loci as the basis of our forensic analysis: - let X^j be the collection of alleles at locus j that are present in a crime sample, and let X denote the collection $[X^1, ..., X^K]$. For example, if $K = 2$, and locus 1 has alleles $\alpha, \beta, \gamma, \delta$ and locus 2 has alleles x, y, z, then X might be $[\{\alpha, \beta\}, \{x, y\}]$.

We wish to compute the likelihood ratio for the following two hypotheses:

H_1: X arose from contributions from a collection of p individuals $\mathrm{I} = \{I_1, ..., I_p\}$ of known genotype at the K loci, and $r_1 \geq 0$ other unknown individuals from population P.

H_2: X arose from contributions from a collection of p' individuals $\mathrm{I}' = \{I'_1, ..., I'_{p'}\}$ of known genotype at the K loci and $r_2 \geq 0$ other unknown individuals from population P.

Note that I and I′ will generally include individuals in common (the victim, and perhaps a consenting partner in the case of a sexual assault investigation), and that r_1 and r_2 are specified in the hypothesis (often one of them is zero, and only rarely is either of them more than 2).

It is possible to use the results presented here (or straightforward modifications of them) to analyse situations where:

(1) r_1 and/or r_2 is not known exactly;

(2) not all the individuals of I and/or I′ may have contributed to the crime sample;

(3) the genotypes of some individuals in I (and/or I′) are not known exactly;

(4) X is not known exactly;

(5) the unknown individuals can come from different populations;

(we describe this further after Proposition 1), so for simplicity we will assume that H_1 and H_2 are precisely as described above.

Notice that H_1 and H_2 are of the same generic form, which simplifies the analysis, as we can compute $\mathbb{P}[X|H_1]$ and $\mathbb{P}[X|H_2]$ by a common formula. Thus, it simplifies notation to consider the generic form of hypothesis 1 and 2 as the following (single or multiple locus) hypothesis H:

H: X arose from contributions from individuals I of known genotype, and $r \geq 0$ other unknown individuals from population P.

To carry out calculations, we suppose that we have accurate estimates of the genotype frequencies of the K loci in P, and that there is approximate **independence** between the loci in P. That is, the proportion of individuals in P of genotype combination $(g^1, ..., g^K)$, where g^i is the genotype at locus i, is approximately equal to the product $f_{g^1} \times f_{g^2} \ldots \times f_{g^K}$.

Proposition 1: Assuming independence between loci, $\mathbb{P}[X|H] = \prod_{i=1}^{K} \mathbb{P}[X^i|H]$.

Proof: Let S_i be the set of those ordered r-tuples $(g_1, g_2, ..., g_r)$ for which: (i) each component is a possible genotypes at locus i, and (ii) the alleles present amongst these r genotypes, together with those alleles present in the genotypes, at locus i, of the individuals from I are precisely the alleles present in X^i. Let $S = S_1 \times S_2 \times \ldots \times S_K$. For $p = (p_1, ..., p_K) \in S$, let $P[p; j]$ denote the proportion of individuals in P that have, for $i = 1, ..., K$, the j-th component of p_i as their genotype at locus i. Then,

$$\begin{aligned}
\mathbb{P}[X|H] &= \sum_{p \in S} \prod_{j=1}^{r} P[p; j] \\
&= \sum_{p \in S} \prod_{i=1}^{K} \prod_{j=1}^{r} f_{(p_i)_j} \\
&= \prod_{i=1}^{K} \Big(\sum_{q \in S_i} \prod_{j=1}^{r} f_{q_j} \Big) \\
&= \prod_{i=1}^{K} \mathbb{P}[X^i|H] \, .
\end{aligned}$$

which completes the proof.

Thus, for hypotheses stated as above, the independence assumption reduces a multiple locus calculation to several single locus calculations. For some of the extensions mentioned above (to hypotheses H^* not of type H) it is helpful to invoke the identity:

$$\mathbb{P}[X|H^*] = \sum_i \mathbb{P}[X|H^* \wedge A_i]\mathbb{P}[A_i|H^*]$$

where $A_1, \ldots$ are partitioning events chosen so that, for each i, the conjoint event $H^* \wedge A_i$ is a hypothesis of type H (and for which Proposition 1 can then be applied). For example, in the extension (1) discussed above, A_i would be the event that the number of unknown individuals from P that contributed to X is i. For example, if one believed it was 80% (resp. 20%) likely that one (resp. two) unknown individual contributed under hypothesis H^*, then (invoking Proposition 1):

$$\mathbb{P}[X|H^*] = 0.8 \prod_{i=1}^{K} \mathbb{P}[X^i|H^* \wedge A_1] + 0.2 \prod_{i=1}^{K} \mathbb{P}[X^i|H^* \wedge A_2] \, .$$

If one wished to adopt a more conservative approach, and not impose any prior probabilities for $\mathbb{P}[A_i|H]$ then one can still obtain an upper (resp. lower) bound on the likelihood ratio LR by selecting the event(s) from the partitioning events A_1,... that maximize (resp. minimize) LR. Whatever approach is adopted it is worth noting that the factorization in Proposition 1 does not apply if H is replaced by H^*.

In view of Proposition 1, calculations are reduced to the single-locus case; and the computation of $\mathbb{P}[X|H]$ is given by Proposition 2 (below). To describe this result we first introduce the following definitions.

Definitions (1) For a set A of alleles, let $Gen(A)$ denote the collection of all those genotypes whose alleles both lie in A. For example, suppose $A = \{a, b, c\}$, then $Gen(A) = \{aa, bb, cc, ab, ac, bc\}$.

(2) For a set G of genotypes, let $All(G)$ denote the collection of alleles which lie in at least one genotype in G. For example, if $G = \{ab, cc, cd\}$ then $All(G) = \{a, b, c, d\}$.

Note that $All(Gen(A)) = A$; $G \subseteq Gen(All(G))$; and $|Gen(A)| = |A|(|A| + 1)/2$.

(3) For two sets of alleles X and Y, with $Y \subseteq X$, order $Gen(X)$ as $g_1, \ldots, g_G$, and let $N_r(X, Y)$ denote the set of ordered G-tuples of non-negative integers, $(n_1, \ldots, n_G)$ which sum up to r, and which have the property that $All(\{g_i : n_i > 0\})$ contains $X - Y$.

Formally,

$$N_r(X,Y) = \{(n_1,\ldots,n_G) : n_i \geq 0 \text{ for all } i, \sum_j n_j = r, \text{and } X - Y \subseteq All(\{g_i : n_i > 0\})\}$$

For example, suppose $Y = \{a, b\}$ and $X = \{a, b, c\}$. Order $Gen(X)$ as follows: aa, bb, cc, ab, ac, bc. Then, for $r = 2$, we have:

$$\begin{aligned} N_2(X,Y) \;=\; & \{(0,0,2,0,0,0),(0,0,0,0,2,0),(0,0,0,0,0,2),(1,0,1,0,0,0),\\ & (1,0,0,0,1,0),(1,0,0,0,0,1),(0,1,1,0,0,0),(0,1,0,0,1,0),\\ & (0,1,0,0,0,1),(0,0,1,1,0,0),(0,0,1,0,1,0),(0,0,1,0,0,1),\\ & (0,0,0,1,1,0),(0,0,0,1,0,1),(0,0,0,0,0,1)\}. \end{aligned}$$

These 15 values are the possible ways that two genotypes chosen from $X = \{a, b, c\}$, together with the alleles $\{a, b\}$ make up a total allele set of $X = \{a, b, c\}$ (in Proposition 3 we give an exact formula for the number of such cases for all r, X, Y).

Proposition 2: For a single locus ($K = 1$), let Y be the collection of alleles that occur at this locus in at least one individual of I, and X the alleles on a crime sample. Then,

(a) $\mathbb{P}[X|H] = 0$ if either

(i) Y contains an allele not present in X, or

(ii) $2r < |X - Y|$.

(b) $\mathbb{P}[X|H] = 1$ if $r = 0$, and $X = Y$.

(c) In all other cases,

$$\mathbb{P}[X|H] \;=\; \sum_{(n_1\ldots n_G)\in N_r(X,Y)} \frac{r!}{n_1!n_2!\ldots n_G!} \prod_{j=1}^{G} f_{g_j}^{n_j}$$

where $n! = n \times (n-1) \times \ldots \times 1$ (and $0! = 1$), and $f_{g_j}^0 = 1$.

Proof: The condition $2r \geq |X - Y|$ in (a) is necessary for $\mathbb{P}[X|H] \neq 0$ since each individual from P can contribute at most two new alleles towards extending Y into X. The other statements in (a) and (b) are trivial. The formula in (c) arises since $\mathbb{P}[X|H]$ is just a sum of multinomial probabilities, assuming the r individuals are chosen randomly from the (large) population P.

Special Cases: (**1**) $r = 1$. In this case, $\mathbb{P}[X|H] = 0$ unless Y is contained in X, and $|X - Y| = 0, 1$ or 2. In these three cases we have:

$$\mathbb{P}[X|H] = \begin{cases} \sum_{g \in Gen(Y)} f_g, & \text{if } X = Y \\ f_{aa} + \sum_{y \in Y} f_{ay}, & \text{if } X - Y = \{a\} \\ f_{ab}, & \text{if } X - Y = \{a, b\} \end{cases}$$

(**2**) $r = 2$. In this case, $\mathbb{P}[X|H] = 0$ unless Y is contained in X, and $|X-Y| = 0, 1, 2, 3, 4$. In these five cases we have:

(i) If $X = Y$,

$$\mathbb{P}[X|H] = \sum_{g \in Gen(Y)} f_g^2 + \sum_{g \neq g' \in Gen(Y)} 2 f_g f_{g'}$$

(ii) if $X - Y = \{a\}$,

$$\mathbb{P}[X|H] = f_{aa}^2 + 2 f_{aa} \sum_{y \in Y} f_{ay} + \sum_{y \in Y} f_{ay}^2 + \sum_{y \neq y' \in Y} 2 f_{ay} f_{ay'} + 2 f_{aa} \sum_{g \in Gen(Y)} f_g + \sum_{y \in Y, g \in Gen(Y)} 2 f_{ay} f_g$$

(iii) if $X - Y = \{a, b\}$,

$$\begin{aligned} \mathbb{P}[X|H] = \; & f_{ab}^2 + 2(f_{aa} f_{bb} + f_{aa} f_{ab} + f_{bb} f_{ab}) + \\ & \sum_{y \in Y} 2[f_{ab}(f_{ay} + f_{by}) + f_{ay} f_{by}] + 2 f_{ab} \sum_{g \in Gen(Y)} f_g \\ & + \sum_{y \neq y' \in Y} 2 f_{ay} f_{by'} + 2 f_{aa} \sum_{y \in Y} f_{by} + 2 f_{bb} \sum_{y \in Y} f_{ay} \end{aligned}$$

(iv) if $X - Y = \{a, b, c\}$,

$$\begin{aligned} \mathbb{P}[X|H] = \; & 2(f_{ab} f_{ac} + f_{ac} f_{bc} + f_{ab} f_{bc} + f_{aa} f_{bc} + f_{bb} f_{ac} + f_{cc} f_{ab}) \\ & + \sum_{y \in Y} 2[f_{ab} f_{cy} + f_{ac} f_{by} + f_{bc} f_{ay}] \end{aligned}$$

(v) if $X - Y = \{a, b, c, d\}$,

$$\mathbb{P}[X|H] = 2(f_{ab} f_{cd} + f_{ac} f_{bd} + f_{ad} f_{bc}).$$

3. Enumerating all cases

It is useful to have a formula to count all of the cases whereby the allele set X can arise under hypothesis H. This amounts to determining the size of $N_r(X,Y)$ and is provided as follows.

Proposition 3: For $Y \subseteq X$,

(1) $$|N_r(X,Y)| = \sum_{s=0}^{x-y} \binom{x-y}{s} (-1)^s \binom{\binom{x-s+1}{2}+r-1}{r}$$

where $x = |X|$, $y = |Y|$, and

$$\binom{m}{k} = \frac{m!}{k!(m-k)!}.$$

(2) In case $x - y > 2r$, $|N_r(X,Y)| = 0$,

and for $x - y = 2r$, $|N_r(X,Y)| = \frac{(2r)!}{2^r r!}$.

Proof: For $S \subseteq X$ let $M_r(X,S)$ denote the number of $|Gen(X)|-$tuples $(n_1, \ldots, n_G)$ where $G = |Gen(X)| = \binom{|X|+1}{2}$, $n_i \geq 0$ for all i, $\sum_{j=1}^{G} n_j = r$, and such that if genotype g_i contains an allele from S then $n_i = 0$. From the definition we have:

$$M_r(X,S) = M_r(X-S,\phi),$$

and

$$M_r(X-S,\phi) = \binom{\binom{x-s+1}{2}+r-1}{r}, \quad \text{where } s = |S|,$$

since the left hand side of this second equation is just the number of ways of selecting r objects from $|Gen(X-S)| = \binom{x-s+1}{2}$ objects, with repetition.

Now, by the principle of inclusion and exclusion (see Anderson, 1974) we have:

$$N_r(X,Y) = \sum_{s\geq 0}(-1)^s \sum_{\substack{S\subseteq X-Y \\ |S|=s}} M_r(X,S).$$

Applying the above two identities to this equation gives:

$$\begin{aligned} N_r(X,Y) &= \sum_{s\geq 0}(-1)^s \sum_{\substack{S\subseteq X-Y \\ |S|=s}} M_r(X-S,\phi) \\ &= \sum_{s\geq 0}(-1)^s \binom{x-y}{s} \binom{\binom{x-s+1}{2}+r-1}{r}, \end{aligned}$$

since there are $\binom{x-y}{s}$ subsets S of $X-Y$ of size s. This establishes part (1) of Proposition 3. For part (2), the result for $x-y>2r$ is clear. In case $x-y=2r$ then $N_r(X,Y)$ is precisely the number of matchings on $X-Y$, and by a classical result (see Anderson 1974) this is $\frac{(2r)!}{2^r r!}$. This completes the proof.

As an example, consider the case $X=\{a,b,c\}, Y=\{a,b\}$ and $r=2$ considered earlier. Applying Proposition 3, we have $|N_2(X,Y)| = \binom{7}{2} - \binom{4}{2} = 15$, as before.

4. Example

The following example, based on a real case, involved the HLA DQA.1 locus which has six alleles $(1.1, 1.2, 1.3, 2, 3, 4)$. The case involved a rape by two offenders, in which alleles $1.1, 2, 4$ were detected in a sample from the victim's panties. It is possible that allele 1.2 was also present, as, in this case, it could be masked by the presence of alleles 1.1 and 4. The victim's genotype was $(1.2, 4)$, and the genotypes of two suspects S_1 and S_2 were: $S_1 = (1.1, 4)$; $S_2 = (1.1, 2)$. We wish to compare four hypotheses:

H_1 : S_1 and one other individual (unknown) contributed to sample.
H_2 : S_2 and one other individual (unknown) contributed to the sample.
H_{12} : S_1 and S_2 contributed to the sample.
H_0 : Two individuals (unknown) contributed to the sample.

Since the victim has allele type 1.2, and since we will assume that the victim may have contributed alleles to the sample on the panties we take $X=\{1.1, 1.2, 2, 4\}$. Our calculations are performed on the basis of genotype frequencies for the New Zealand population (Stringer et al. 1995).

For H_1 we have $Y=\{1.1, 1.2, 4\}$, and $r=1$, so that,

$$\mathbb{P}[X|H_1] = f_{1.1,2} + f_{1.2,2} + f_{2,2} + f_{4,2} = 0.204$$

For H_2, $Y=\{1.1, 1.2, 2, 4\}$, and $r=1$, so that (Proposition 2, special case 1):

$$\mathbb{P}[X|H_2] = f_{1.1,1.1} + f_{1.2,1.2} + f_{2.2} + f_{4,4} + f_{1.1,1.2} + f_{1.1,2} + f_{1.1,4} + f_{1.2,2} + f_{1.2,4} + f_{2,4} = 0.606$$

For H_{12} we have $Y=\{1.1, 1.2, 2, 4\} = X, r=0$, so that $\mathbb{P}[X|H_{12}] = 1$.

For H_0 we have $Y=\{1.2, 4\}$, and $r=2$ so that, by Proposition 3, there are $\binom{11}{2} - 2\binom{7}{2} + \binom{4}{2} = 19$ terms in the sum for $\mathbb{P}[X|H_0]$. Applying special case (2ii) of Proposition

2 we get:

$$\mathbb{P}[X|H_0] = 0.082$$

The likelihood ratio LR for each pair of hypotheses is displayed in the following array (where the entry for row H_α and column H_β is $\frac{\mathbb{P}[X|H_\beta]}{\mathbb{P}[X|H_\alpha]}$):

	H_1	H_2	H_{12}	H_0
H_1	1.0	3.0	4.9	0.4
H_2	0.3	1.0	1.7	0.1
H_{12}	0.2	0.6	1.0	0.1
H_0	2.5	7.4	12.2	1.0

A programme (in C) for performing the calculations like those above, and extensions to multiple locus analysis is available, upon request from the authors. Note that, since the genotype frequencies are generally estimated from a sample from the population, it will be important to consider the sampling error introduced into these estimates, particularly if the sample is small. This in turn can affect the estimates of joint genotype frequencies for multiple loci and thereby the likelihood ratios. The assumption of independence of loci should also be addressed (see Sudbury et al. 1993). Thus, two important future extension to our results would be to explicitly incorporate (i) sampling effects, and (ii) linkage between the genotype frequencies of different loci.

References

[1] Aitken, C.G.G. (1995). *Statistics and the evaluation of evidence for forensic scientists.* Brisbane: John Wiley and Sons.

[2] Anderson, I. (1974). *A first course in combinatorial mathematics.* Oxford: Clarendon Press.

[3] Evett, I. W., Buffery, C., Willott, G. and Stoney, D. (1991). A guide to interpreting single locus profiles of DNA mixtures in forensic cases. *Journal of the Forensic Science Society* **31(1)**: 41-47.

[4] Howson, C. and Urbach, P. (1991). Bayesian reasoning in science. *Nature* **350**:371-374.

[5] Robertson, B. and Vignaux, G.A. (1995). Interpreting evidence: Evaluating forensic science in the courtroom. Chichester: John Wiley and Sons.

[6] Stringer, P., Triggs, C. M., Baldwin, L. C., Melia, L. M. and Savill, M. G. (1995). Distribution of HLA DQA.1 alleles in New Zealand Caucasian, Maori and Pacific Islander populations. *International Journal of Legal Medicine* **108(1)**:2-7.

[7] Sudbury, A. W., Marinopoulos, J. and Gunn, P. (1993). Assessing the evidential value of DNA profiles matching without using the assumption of independent loci. *Journal of the Forensic Science Society* **33(2)**: 73-82.