Ian W. Evett,[1] *D.Sc.; Peter D. Gill,*[2] *Ph.D.; and James A. Lambert,*[2] *M.Sc.*

Taking Account of Peak Areas when Interpreting Mixed DNA Profiles

REFERENCE: Evett IW, Gill PD, Lambert JA. Taking account of peak areas when interpreting mixed DNA profiles. J Forensic Sci 1998;43(1):62–69.

ABSTRACT: This paper establishes a logical framework for taking account of peak areas when interpreting mixed DNA STR profiles. The principles apply wherever such data are available but they are illustrated here by means of data which have been collected from made up mixtures of known concentrations analyzed at short tandem repeat loci. The data have led to some modeling assumptions which are used for numerical examples. In actual casework the proportions of the various components will not be known and there is a discussion of whether they should be allowed for by integrating over a prior distribution. This is a conceptual paper, rather than a prescription for casework, and the scope for further work is outlined.

KEYWORDS: forensic science, interpretation, DNA, profiling, STR, PCR, mixtures, probability, statistics, Bayesian

A method for interpreting mixed RFLP profiles, taking no account of band intensity, was described by Evett, Buffery, Willott and Stoney (1), endorsed by the National Research Council of the USA (2) and recently extensively generalized by Weir et al. (3). The analysis requires the assignment of probabilities of all of the combinations of genotypes which could give rise to the observed mixture given each of the explanations for its occurrence. If reliable peak area data are available from the mixture then the various combinations can be weighted logically using the conceptual framework which we describe here.

Forensic scientists are accustomed to taking account of intensity informally when interpreting mixtures. Often, it is possible reliably to separate the major and minor components of simple mixtures by visual examination, particularly if the different contributions are in the proportions of less than 1:5. As the proportions approach equivalence, or if mixtures comprise contributions from 3 or more individuals, then interpretation becomes increasingly problematical. However, peak areas of alleles can easily be estimated by densitometric methods. All electrophoretic DNA profiles are amenable to this kind of analysis. Analysis of short tandem repeat (STR) loci is facilitated by the ABD 373A and 377 Genescan software which automatically records peak height and peak areas of STR alleles into a spreadsheet format.

The next section describes the materials and methods used for DNA analysis in this study. Then we present a theoretical justification for a general expression for the likelihood ratio (LR) for any number of peaks, loci and possible contributors, taking account of peak areas. Readers who so prefer may omit this and proceed to the following section which discusses the particular case where there are two contributors to the mixture and one suspect. The underlying principles are explained for the cases where the mixture has 4 and 3 peaks respectively at one given locus. Next we present some simple modeling assumptions based on the analysis of mixtures of known proportions. Illustrative calculations for multiple locus calculations are then given and the merits and demerits of using variables which are mixture independent are discussed with numerical examples. Finally, we indicate how the path of future development may evolve.

Materials and Methods

Mixtures were prepared following the methods and protocols described by Sparkes et al. (4). Six individuals provided DNA which was blended in four different pairwise combinations and profiled at 6 loci (plus a sex test, omitted from this study). Known proportions by concentration (1:1, 1:2, 1:5, 1:10, 10:1, 5:1, 2:1) were prepared for each pairwise combination. For each mixture thus prepared, 1 ng and 5 ng aliquots were separately amplified by PCR using multiplex conditions and loci described by Sparkes et al. (4) and Kimpton et al. (5) and electrophoresed on ABD 377 automated sequencers. Determination of allele sizes (bp), peak height and peak area, was carried out by GENESCAN™ and GENOTYPER™ software. There were, in the 24 single locus combinations, 8 four peaked profiles, 13 three peaked, 2 two peaked and one single peaked.

Details of Loci

The primers used to amplify a seven locus multiplex system are described elsewhere by Oldroyd et al. (6) and include HUMTH01, D21S11, D18S51, D8S502, HUMVWFA31/A, HUMFIBRA/FGA and an amelogenin sex test. The frequency distributions of loci for 3 different races are described by Evett et al. (7). The identification and nomenclature of STR alleles is described in detail by Gill et al. (8).

Theory

Assume that a crime stain has been profiled and that the non-scientific evidence indicates that *n* offenders contributed DNA to the stain. We consider the case where there is a single person who is suspected of having been one of the contributors. Then we consider two alternatives:

[1]Forensic Science Service, Metropolitan Laboratory, 109 Lambeth Rd, London, SE1 7LP, UK.
[2]Forensic Science Service, Priory House, Gooch St North, Birmingham, B5 6QQ, UK.

Received 18 Oct. 1996; and in revised form 31 March, 30 June 1997; accepted 30 June 1997.

C: the suspect and $(n - 1)$ unknown people contributed to the mixture

$\overline{C}$: n unknown people contributed to the mixture

For the time being we consider one locus. Let $\mathbf{A} = (A_1, A_2, \ldots A_r)$ denote a set of r peaks, each representing an allele. Each peak may comprise a contribution from more than one person. Let $\mathbf{w} = (w_1, w_2, \ldots w_r)$ denote a corresponding set of peak areas (for convenience, they will be normalized to sum to 1, but this is not essential) and let $\mathbf{f} = (f_1, f_2, \ldots f_r)$ denote the corresponding set of allele frequencies. Let G_1 denote the suspect's genotype, then the likelihood ratio (LR) is:

$$\frac{P(E|C)}{P(E|\overline{C})} = \frac{P(\mathbf{A}, \mathbf{w}, G_1|C)}{P(\mathbf{A}, \mathbf{w}, G_1|\overline{C})} = \frac{P(\mathbf{A}, \mathbf{w}|C, G_1)}{P(\mathbf{A}, \mathbf{w}|\overline{C}, G_1)} \cdot \frac{P(G_1|C)}{P(G_1|\overline{C})} \quad (1)$$

If we assume that G_1 is independent of whether or not C is true then the second ratio is one. The first ratio can be simplified by the assumption that $\mathbf{A}$ is independent of G_1 given C, to give:

$$LR = \frac{P(\mathbf{A}, \mathbf{w}|C, G_1)}{P(\mathbf{A}, \mathbf{w}|\overline{C})} \quad (2)$$

It is necessary to keep in mind that both of these probabilities are conditioned also on the background information, or *circumstances*, I. Let $\mathbf{G}_i$ denote a set of n genotypes $(G_1, G_2, \ldots G_n)$, where each G_i is a pair of alleles (A_k, A_l). Then it is necessary to sum over all of the possible G_i that could give rise to the mixture:

$$LR = \frac{\sum_i P(\mathbf{A}, \mathbf{w}|C, \mathbf{G}_i, G_1)P(\mathbf{G}_i|C, G_1)}{\sum_i P(\mathbf{A}, \mathbf{w}|\overline{C}, \mathbf{G}_i)P(\mathbf{G}_i|\overline{C})} \quad (3)$$

Let $\mathbf{m}$ denote a set of n mixing proportions $(m_1, m_2, \ldots m_n)$ where the sum of the m_i is unity. In casework, these proportions will, in general, be unknown. Then the notation thus far is summarized in Table 1, with examples in the third column for the particular case where $r = 4$ and $n = 2$. The LR is evaluated by integrating over all values of $\mathbf{m}$:

TABLE 1—*Illustrative notation for a four peak mixture where there were two offenders. Bold letters denote matrices.*

Letter	Description	Example for a four peak mixture
r	Number of allelic peaks in profile	4
$\mathbf{A}$	A set of r bands	(A_1, A_2, A_3, A_4)
$\mathbf{f}$	A set of allele frequencies	(f_1, f_2, f_3, f_4)
$\mathbf{w}$	A set of peak areas	(w_1, w_2, w_3, w_4)
n	Number of genotypes contributing to profile	2
$\mathbf{G}$	A matrix of n genotypes	(G_1, G_2) e.g: $G_1 = (A_1, A_2)$ $G_2 = (A_3, A_4)$
$\mathbf{m}$	A matrix of n mixing proportions: i.e., the proportions in which the genotypes are present. These are unknown in casework.	(m_1, m_2) n.b.: $m_1 + m_2 = 1$

$$LR = \frac{\int \sum_i P(\mathbf{G}_i|C, G_1)P(\mathbf{A}, \mathbf{w}|C, \mathbf{G}_i, G_1, \mathbf{m})p(\mathbf{m}|C, \mathbf{G}_i, G_1)d\mathbf{m}}{\int \sum_i P(\mathbf{G}_i|\overline{C})P(\mathbf{A}, \mathbf{w}|\overline{C}, \mathbf{G}_i, \mathbf{m})p(\mathbf{m}|\overline{C}, \mathrm{G}_i)d\mathbf{m}} \quad (4)$$

Where $p(\mathbf{m}|C, \mathbf{G}_i, G_1)$ and $p(\mathbf{m}|\overline{C}, \mathbf{G}_i)$ are prior probability density functions. We first assume that these are the same for numerator and denominator and independent of the genotypes, i.e.,:

Assumption 1:

$$(a)\ p(\mathbf{m}|C, \mathbf{G}_i, G_1) = p(\mathbf{m})$$

$$(b)\ p(\mathbf{m}|\overline{C}, \mathbf{G}_i) = p(\mathbf{m})$$

Next we note that:

$$P(\mathbf{A}, \mathbf{w}|C, \mathbf{G}_i, G_1, \mathbf{m}) = p(\mathbf{w}|\mathbf{A}, C, \mathbf{G}_i, G_1, \mathbf{m})P(\mathbf{A}|C, \mathbf{G}_i, G_1, \mathbf{m})$$

$$P(\mathbf{A}, \mathbf{w}|\overline{C}, \mathbf{G}_i, \mathbf{m}) = p(\mathbf{w}|\mathbf{A}, \overline{C}, \mathbf{G}_i, \mathbf{m})P(\mathbf{A}|\overline{C}, \mathbf{G}_i, \mathbf{m})$$

For any given genotype configuration $\mathbf{G}_i$, the set of allelic peaks will either be $\mathbf{A}$ or not. We are interested only in those configurations which will result in $\mathbf{A}$ and so, for the terms in the summations:

Assumption 2:

$$(a)\ P(\mathbf{A}|C, \mathbf{G}_i, G_1, \mathbf{m}) = 1$$

for all terms i in the numerator summation

$$(b)\ P(\mathbf{A}|\overline{C}, \mathbf{G}_i, \mathbf{m}) = 1$$

for all terms i in the denominator summation.

We also make the reasonable assumption that the peak areas depend only on the genotypes present and the mixing proportion:

Assumption 3:

$$(a)\ p(\mathbf{w}|\mathbf{A}, C, \mathbf{G}_i, G_1, \mathbf{m}) = p(\mathbf{w}|\mathbf{G}_i, G_1, \mathbf{m})$$

$$(b)\ p(\mathbf{w}|\mathbf{A}, \overline{C}, \mathbf{G}_i, \mathbf{m}) = p(\mathbf{w}|\mathbf{G}_i, \mathbf{m})$$

Incorporating assumptions 1 to 3, equation (4) becomes:

$$LR = \frac{\int \sum_i P(\mathbf{G}_i|C, G_1)p(\mathbf{w}|\mathbf{G}_i, G_1, \mathbf{m})p(\mathbf{m})d\mathbf{m}}{\int \sum_i P(\mathbf{G}_i|\overline{C})p(\mathbf{w}|\mathbf{G}_i, \mathbf{m})p(\mathbf{m})d\mathbf{m}} \quad (5)$$

To generalize to s loci we assume that the prior distribution for $\mathbf{m}$ is the same for all loci. Then write (5) for the q 'th locus as:

$$LR_q = \frac{\int \pi_C(\mathbf{w}_q|\mathbf{m})p(\mathbf{m})d\mathbf{m}}{\int \pi_{\overline{C}}(\mathbf{w}_q|\mathbf{m})p(\mathbf{m})d\mathrm{m}}$$

TABLE 2—*List of genotypes for analysis of a four peak mixture.*

j	G_j
1	A_1, A_2
2	A_3, A_4
3	A_1, A_3
4	A_2, A_4
5	A_1, A_4
6	A_2, A_3

If we now assume that genotypes are independent between loci, then the overall *LR* is:

$$LR = \frac{\int \prod_q \pi_C(\mathbf{w}_q|\mathbf{m})p(\mathbf{m})d\mathbf{m}}{\int \prod_q \pi_{\overline{C}}(\mathbf{w}_q|\mathbf{m})p(\mathbf{m})d\mathbf{m}} \quad (6)$$

From here we consider the case where $n = 2$. In such cases r can be 1, 2, 3 or 4. For the present, we consider only the cases of 3 and 4 peaks.

Illustration of the Principles

Consider first the case where there are four peaks at one locus. We denote the four alleles implied by the peaks as $\mathbf{A} = (A_1, A_2, A_3, A_4)$ and, without loss of generality, we assume that they are ordered in increasing allele number. Let the suspect's genotype, $G_1 = (A_1, A_2)$, for this explanation: the method for cases where $G_1 = (A_1, A_3)$ etc. follows in an obvious manner. This notation is summarized in the last column of Table 1.

The numerator now simplifies to one term, because there is only one genotype configuration which could give rise to the profile. We denote this configuration by $\mathbf{G}_1 = (G_1, G_2)$, where $G_2 = (A_3, A_4)$. Then $P(\mathbf{G}_1|C, G_1) = 2f_3f_4$ and the numerator of the *LR*, from (5) is, with a little more simplification:

$$2f_3\, f_4 \int p(\mathbf{w}|\mathbf{G}_1, \mathbf{m})p(\mathbf{m})d\mathbf{m}$$

For the denominator, let G_j, $j = 1,2\cdots6$, denote six genotypes as itemized in Table 2. Then the first three $\mathbf{G}_i$ are as listed as ordered pairs in Table 3: the next three $\mathbf{G}_i$ are the same as these with the ordering reversed. It is necessary to distinguish between the orderings because it will later be taken that the proportion m applies to the first member of a pair and $(1 - m)$ to the second. For each of these six, $P(\mathbf{G}_i|\overline{C}) = 4f_1f_2f_3f_4$ and the denominator can be written as, in this case, $4f_1f_2f_3f_4 \sum_{i=1..6} \int p(\mathbf{w}|\mathbf{G}_i, \mathbf{m})p(\mathbf{m})d\mathbf{m}$, and it follows that the *LR* is:

$$LR = \frac{\int p(\mathbf{w}|\mathbf{G}_1, \mathbf{m})p(\mathbf{m})d\mathbf{m}}{2f_1\, f_2 \int \sum_{i=1..6} \mathrm{p}(\mathbf{w}|\mathbf{G}_i, \mathbf{m})p(\mathbf{m})d\mathbf{m}} \quad (7)$$

TABLE 3—*Ordered pairs of genotypes for a four peak mixture.*

i	$\mathbf{G}_i$
1	G_1, G_2
2	G_3, G_4
3	G_5, G_6

If we set the numerator and each of the six terms in the denominator to unity then this becomes the expression which would be derived using the method in Evett et al. (2) where no account is taken of intensity.

The analysis for three peaks in the crime profile is a little more complicated and there are two different cases to consider for the numerator, depending on whether the suspect is homozygous for heterozygous. Here we consider the latter case. Let $\mathbf{A} = (A_1, A_2, A_3)$ and let $G_1 = (A_1, A_2)$ as before. The full list of G_j is shown in Table 4 and the list of the first six ordered $\mathbf{G}_i$ is shown in Table 5. Now there are three terms in the numerator and six in the denominator. These terms need different kinds of treatment, as summarized in Table 6. Whatever the combination of genotypes, there will be one peak which consists of two contributions of the same allele and there are two ways in which this can happen: either one of the two contributors is homozygous, in which both contributions come from the same person (lines 1, 5, 6); or two heterozygous contributors share a common allele (lines 2, 3, 4). We return to the analysis of this case in the subsequent sections.

Derivation of Modeling Assumptions

We now consider how the probability densities $p(\mathbf{w}|\mathbf{G}_i, \mathbf{m})$ in equation (5) may be evaluated. In particular, we base modeling

TABLE 4—*List of genotypes for analysis of a three peak mixture.*

j	$\mathbf{G}_j$
1	A_1, A_2
2	A_3, A_3
3	A_1, A_3
4	A_2, A_3
5	A_2, A_2
6	A_1, A_1

TABLE 5—*Ordered pairs of genotypes for a three peak mixture.*

i	$\mathbf{G}_i$
1	G_1, G_2
2	G_1, G_3
3	G_1, G_4
4	G_3, G_4
5	G_3, G_5
6	G_4, G_6

TABLE 6—*Summary of genotype combinations for the three peak case.*

i	$\mathbf{G}_i$	$P(\mathbf{G}_i\|C)$	$P(\mathbf{G}_i\|\overline{C})$	Shared peaks
1	G_1, G_2	f_3^2	$2f_1f_2f_3^2$	Peak 3—one person
2	G_1, G_3	$2f_1f_3$	$4f_1^2f_2\ f_3$	Peak 1—two persons
3	G_1, G_4	$2f_2f_3$	$4f_1f_2^2\ f_3$	Peak 2—two persons
4	G_3, G_4	0	$4f_1f_2f_3^2$	Peak 3—two persons
5	G_3, G_5	0	$2f_1f_2^2\ f_3$	Peak 2—one person
6	G_4, G_6	0	$2f_1^2f_2f_3$	Peak 1—one person

assumptions on the data from the two person mixtures prepared as described above. We continue to consider evaluation at a single locus and later extend it to the multilocus case by assuming conditional independence given **m** as in equation (7). We also continue to consider mixtures with three and four peaks. Peak areas in two peak mixtures tend not to be informative.

It is necessary to employ summary functions of **w** which make best use of the intensities in discriminating between the various alternatives. An exhaustive evaluation would require extensive data but we maintain that there are good grounds to believe that in every situation there are simple and fairly obvious summary functions. These can be classified into two groups—mixture dependent and mixture independent. We continue with the convention that peaks are ordered in increasing size and we also assume that the w_i have been normalized to sum to 1. Then we suggest that informative summary functions for the three and four banded cases are as shown in Table 7. Note the following.

1. We adopt the convention of always putting the area of the lighter weight allele (or pair of alleles) as the numerator of a ratio.
2. The notation extends in an obvious way to other genotype configurations.
3. In the simplest modeling, the expected values of the functions on lines *a*, *b* and *d* would be unity and that for the function on line *f* would be 0.5.
4. The expected value for the mixture dependent ratios on lines *c*, *e* and *g* would be the (unknown) mixture proportion.

We now explain how the data collected from the made up mixtures suggest a number of modeling assumptions for the various functions.

Functions a, b *and* d—These are all mixture independent functions as each is the ratio of peak areas for the two peaks of one genotype. There are indications from the data that these functions will have distributions which vary from locus to locus. An obvious example is VWA, where the lighter weight peak tends to be the more intense, so the mean of the distribution is positive. This is illustrated in Fig. 1. The mean and SD are 1.2 and 0.2 respectively, excluding an extreme outlier at 4.95. Summaries for five of the loci are shown in Table 8. It is also likely that the distribution is dependent on the difference in allele sizes. Investigation of these effects can only be undertaken by further experimentation. For the time being, the data from all loci have been combined to give estimates for an underlying Normal distribution adopted for modeling purposes. The combined observed distribution is shown in Fig. 2. The overall mean and SD are 1.10 and 0.22 respectively. There is a suggestion of skewness so the Normal assumption should be no more than provisional. The observation that the mean exceeds one is in agreement with the tendency for lighter fragments to be amplified more efficiently than heavier ones.

TABLE 7—*Summary statistics based on peak areas for three and four peak profiles.*

	Peak Configuration	Genotype Configuration	Summary Function	Mixture Dependent?
a	$A_1A_2A_3A_4$	$(A_1A_2)(A_3A_4)$	w_1/w_2	No
b			w_3/w_4	No
c			$(w_1 + w_2)/(w_3 + w_4)$	Yes
d	$A_1A_2A_3$	$(A_1A_2)(A_3A_3)$	w_1/w_2	No
e			$(w_1 + w_2)/w_3$	Yes
f	$A_1A_2A_3$	$(A_1A_2)(A_2A_3)$	w_2	No
g			w_1/w_3	Yes

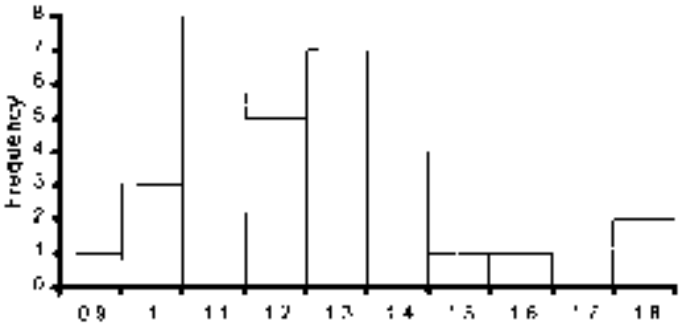

FIG. 1—*Distribution of the ratio of the peak areas of the lighter and heavier alleles for VWA genotypes (functions* a, b *and* d*).*

TABLE 8—*Estimated means and standard deviations for functions* a, b, *and* d.

	Mean	SD
D18	1.10	0.24
D21	1.06	0.09
D8	1.06	0.30
FGA	1.02	0.10
VWA	1.21	0.21

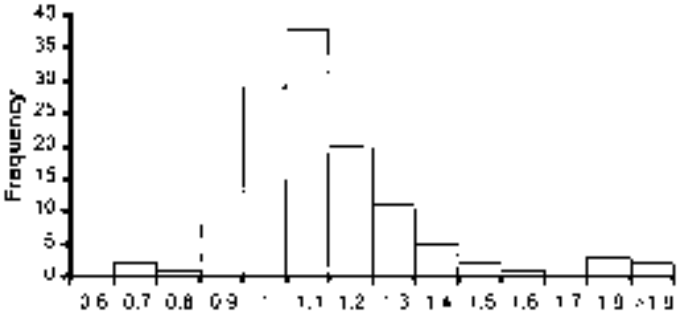

FIG. 2—*Distribution of the ratio of the peak areas of the lighter and heavier alleles for all genotypes (functions* a, b *and* d*).*

Functions c *and* g—These are mixture dependent functions. For a two person mixture, the mixture matrix **m** becomes simply $\{m, (1 - m)\}$ where m is the proportion contributed by the first genotype in a given combination. In casework m will not, in general, be known but they can be studied by the data from the known mixtures described earlier. As *c* and *g* are ratios then it makes sense to study their variation with regard to the mixture proportion $m/(1 - m)$: Table 9 shows summary statistics for *c* and *g* combined. Each row shows min/max etc., for the ratio of combined peak areas for the respective value of the mixture ratio. Note that

TABLE 9—*Summary statistics for functions* c *and* g.

$m/(1 - m)$	Min	Max	Mean	SD	cv
0.1	0.044	0.363	0.15	0.078	0.519
0.2	0.086	1.224	0.472	0.362	0.767
0.5	0.2	1.087	0.533	0.273	0.512
1	0.451	2.416	1.157	0.568	0.49
2	0.944	5.996	2.773	1.618	0.584
5	1.81	8.25	4.696	2.16	0.46
10	3.951	26.23	11.29	5.618	0.497

TABLE 10—*Variation in function* e *with mixture ratio,* q.

$m/(1-m)$	Min of e	Max of e	Mean of e	SD	cv
0.1	0.10	0.36	0.21	0.11	0.51
0.2	0.24	0.74	0.49	0.21	0.42
0.5	0.69	1.11	0.90	0.30	0.34
1	1.33	1.97	1.67	0.32	0.19
2	2.73	3.78	3.25	0.43	0.13
5	7.72	11.42	9.57	2.62	0.27
10	10.25	23.51	14.78	6.04	0.41

the observed ranges tend to be large—and this is in spite of the fact that a few extreme outliers were omitted from the summary. In part, this variation is attributable to the practical difficulty of making DNA mixtures to precise proportions, nevertheless, it suggests that the predictive power of this ratio would not appear to be particularly good. For simple exploratory modeling, we have taken the functions to be Normal with mean equal to the mixture ratio and a coefficient of variation (cv) of 0.5. Improved modeling might be yielded by transformations of the data though we suspect that the improvement would be marginal.

Function e—It would seem natural to expect function *e* to have the same distribution as *c* and *g* but this proved not to be the case. Note that for function *e* there are three configurations, depending whether the homozygote peak is the lightest, middle or heavy allele. These have not been distinguished between for the time being and for the sake of summarizing the data the convention adopted has been to divide the sum of the two heterozygous peaks by the homozygous peak. The data are summarized in Table 10. Note that the mean of the function is consistently larger than the mixture ratio $m/(1-m)$, an indication that when two copies of the same allele are present from one person, they are amplified less efficiently that two different alleles. We are not aware of this observation having been made before and we have no simple explanation for it. For simple modeling we propose a Normal distribution with mean of 1.5 times the mixture ratio and cv of 0.5. Once again, transformations of the data may prove superior but, we suspect, not greatly so.

Function f—In theory, the distribution should be mixture independent with mean 0.5. However, it will not be quite as simple as this in practice because it depends to some extent on whether the combined peak is the smallest (A), middle (B) or heaviest allele (C) as Table 11 shows. However, once again for the sake of simplicity, we model all three with a Normal distribution with mean 0.5 and a generous SD of 0.06; we recognize that this is sacrificing, to some extent, informativeness for simplicity.

The modeling assumptions are summarized in Table 12. In the calculations that are described in the next section the observed values of functions are transformed to N(0, 1) variables as shown in the last column. γ denotes the mixture ratio $m/(1-m)$.

TABLE 11—*Function* f.

Shared Peak	Mean	SD
A	0.53	0.03
B	0.50	0.02
C	0.48	0.02

Illustrative Calculation

We have carried out many experimental calculations on the mixture data but, for illustration we restrict ourselves to one particular combination. This is a set of 6-locus profile data for 5 ng of a 10:1 mixture of the DNA of two of the known individuals. We take the mixture data as representing a crime profile and take one of the contributors to the mixture as a hypothetical suspect. The data are summarized in Table 13. The second column shows that there were four peaks for D21 and VWA; three peaks for D18, D8 and FGA; and two peaks for THO1. The allele designations are in the second column, the genotype of the "suspect" is shown in the third column, followed by the peak areas in the fourth column. Normalized peak areas are in column 5 and Caucasian allele frequencies are in the last column. Visual inspection of, for example the areas for D21 and VWA shows that they lend support to the presence of the suspect's genotype in the mixture.

The principles of the calculation are illustrated by the summary for D21 in Table 14. The possible genotype configurations given two contributors are shown in the second column with the corresponding genotype combination probabilities shown in the third column: that for the numerator is $2f_af_b$. As this is a four peaked case, each of the genotype combinations for the denominator has the same probability of $4f_af_bf_cf_d$ (in the three banded case, the probability will vary from line to line—AB, BC has a different probability from that of AA, BC, for example). The next two columns are the mixture independent functions *a* and *b*: on the first line, for example, *a* is the ratio of the peaks C and D and *b* is the ratio of peaks A and B. Next comes function *c*, which is mixture dependent. The three probability densities are calculated in the following three columns as N(0, 1) ordinates using the transforms of the data in the last column of Table 12. Note that the combination CD, AB in the denominator is associated with the highest probability density as shown in the last column and is thus the best supported (note that the order CD, AB is only important in the sense that the proportion *m* is applied to the first genotype: for practical purposes it is indistinguishable from AB, CD). Evaluation of the probability density for function *c* requires a value for the mixture proportion but in casework this will not, in general, be known, and so it is necessary to integrate over the range of possible values. The example shown has been calculated for $m = 0.9$ and the *LR* conditioned on this value of *m* is the ratio of the numerator and denominator.

Similar calculations are carried out locus by locus and the overall numerator and denominator arrived at by multiplying the individual single locus values together. Table 15 shows how the log(base 10) of the overall *LR*, for the data in Table 13 varies when different values of *m* are used. Not surprisingly, the *LR* has a maximum at $m = 0.9$, bearing in mind that the mixture was made up to $m = 0.91$.

Table 16 shows the variation of the log of the *LR* with the input value of *m* for each of the different mixtures that were made up for this particular pair of individuals. It will be seen that, as expected, the *LR* tends to peak when the input value approximates to the true value.

In casework, when *m* is not known, it can be seem from equation 6 that it is necessary to integrate the numerator and denominator over a prior distribution for *m*. If we accept that prior ignorance can be represented by a uniform probability density function for *m* then, a mean *LR* can be calculated and the log of the mean for each of the mixtures is shown in the last row of Table 16. Note that the *LR* is of the order 10^7 whereas, if we were not allowing for intensity then based on frequencies alone it would have been

TABLE 12—*Modeling assumptions for peak area functions.*

Function	Form of Function (x)	Explanation	Mixture Dependent?	Transformation
a, b, d	w_i/w_j	Ratio of two peaks for the same heterozygote	No	$(x - \mu_1)/\sigma_1$
c	$(w_i + w_j)/(w_k + w_l)$	Ratio of the sums of the peaks from two heterozygotes	Yes	$(x - \gamma)/0.5q$
e	$(w_i + w_j)/w_k$	Ratio of the sum of two peaks from a heterozygote to the peak from a homozygote	Yes	$(x - 1.5\gamma)/0.75\gamma$
f	w_2	Area of a peak shared by two heterozygotes	No	$(x - \mu_2)/\sigma_2$
g	w_i/w_j	Ratio of peaks from two different heterozygotes	Yes	$(x - \gamma)/0.5q$

TABLE 13—*Summary of the details of a 10:1 mixture of the DNA of two people (5 ng).*

Locus	Crime Profile	Alleles	Suspect's Genotype	Peak Areas	Normalized Areas	Allele Frequencies
TH01	A	8	A	17441	0.438	0.108
	B	9.3	B	22368	0.562	0.304
D21	A	59		1226	0.060	0.031
	B	65		1434	0.070	0.258
	C	67	C	8816	0.433	0.069
	D	70	D	8894	0.437	0.09
D18	A	13	AA	38985	0.909	0.125
	B	16		1914	0.045	0.137
	C	17		1991	0.046	0.115
D8	A	10	A	6416	0.515	0.094
	B	11		383	0.031	0.066
	C	14	C	5659	0.454	0.209
VWA	A	16	A	4669	0.444	0.216
	B	17		931	0.089	0.27
	C	18	C	4724	0.449	0.219
	D	19		188	0.018	0.093
FGA	A	21	A	16099	0.582	0.187
	B	22	B	10538	0.381	0.165
	C	23		1014	0.037	0.139

TABLE 15—*Variation in log_{10} LR with input value of m for the case where the mixture has been made up to a value of $m = 0.91$.*

m	γ	$\log_{10} LR$
0.1	0.11	−22.96
0.2	0.25	−21.99
0.3	0.43	−21.30
0.4	0.67	−20.84
0.5	1.00	−20.50
0.6	1.50	−20.25
0.7	2.33	−13.33
0.8	4.00	−0.31
0.9	9.00	7.70

5800. Thus the information in the peak intensities lends considerable extra support to the (correct) hypothesis that the suspect is a contributor to the mixture.

There are two further issues which can best be discussed by considering data from a single locus. First, it is also the case that the peak intensities can lead to a dramatic *reduction* to the *LR* and thus provide an additional tool for discrimination. Second, the mixture dependent functions cause a problem in that the mixture

TABLE 14—*Summary of calculation of likelihood ratio for D21.*

			Functions of Peak Areas			Probability Densities for Functions of Peak Areas			
	$\mathbf{G}_i$	$P(\mathbf{G}_i\|\cdot)$	a	b	c	a	b	c	$P(\mathbf{G}_i\|\cdot)p(\mathbf{w}\|\cdot)$
Numerator	CD AB	0.015996	0.991	0.855	6.658	0.3530	0.2145	0.3484	0.000422
Denominator	AB CD	0.000199	0.855	0.991	0.150	0.2145	0.3530	0.0577	8.68E-07
	CD AB	0.000199	0.991	0.855	6.658	0.3530	0.2145	0.3484	5.24E-06
	AC BD	0.000199	0.139	0.161	0.972	2.87E-05	4.44E-05	0.0813	2.06E-14
	BD AC	0.000199	0.161	0.139	1.028	4.44E-05	2.87E-05	0.0831	2.1E-14
	AD BC	0.000199	0.138	0.163	0.987	2.8E-05	4.56E-05	0.0817	2.08E-14
	BC AD	0.000199	0.163	0.138	1.013	4.56E-05	2.8E-05	0.0826	2.1E-14
								Denominator:	6.11E-06

TABLE 16—*Variation of log_{10} (*LR*) with input value of* m *for all of the different made up mixtures for the selected combination of two individuals.*

Value of *m* Used to Calculate *LR*	Value of *m* to Which Mixtures Were Made Up						
	0.09	0.17	0.33	0.50	0.67	0.83	0.91
	log(Base 10) of *LR*						
0.1	7.60	7.63	−4.30	−8.61	−8.17	−19.87	−22.96
0.2	6.22	7.81	5.08	−2.88	−17.95	−23.82	−21.99
0.3	0.27	6.78	7.45	5.69	−14.86	−23.06	−21.30
0.4	0.25	3.01	7.46	6.58	0.72	−22.31	−20.84
0.5	0.20	0.24	5.85	5.98	5.76	−17.62	−20.50
0.6	0.11	0.09	0.39	5.43	7.18	−5.30	−20.25
0.7	−0.05	−0.44	−1.70	4.91	7.65	5.68	−13.33
0.8	−0.94	−2.39	−1.51	1.17	7.68	8.16	−0.31
0.9	−3.69	−3.87	−0.53	−0.36	0.74	8.09	7.70
log_{10} (mean *LR*)	7.31	7.13	7.33	6.02	6.93	8.04	7.67

is not known and so it is necessary to average over a prior distribution: if the mixture dependent functions are not affecting discrimination, then it may be possible to dispense with them. We consider these issues together, first at locus D21, then at locus D18.

Locus D21

It will be seen from Table 13 that the peak area data at this locus clearly support the genotype combination AB, CD. Now, if the peak areas for alleles B and C are interchanged it will be seen that the data now tend to favor combination AC, BD. To explore this quantitatively, the peak areas for B and C were transformed in the following way:

$$B' = Bx + C(1 - x)$$

$$C' = Cx + B(1 - x)$$

where B, C denote the original peak areas and B', C' the transformed areas. Thus, $x = 0$ interchanges the peak areas and $x = 1$ has them unchanged. The *LR* was calculated for values of x ascending in intervals of 0.1 from zero to one, using, for each x, two methods: the mixture dependent method included the function c, following the approach described in the previous section, and averaging numerator and denominator over a uniform prior distribution for m and taking the ratio; the mixture independent method was based on functions a and b only so there was no need to average over a distribution for m. The variation in *LR* from both calculations is shown in Fig. 3. The horizontal unbroken line shows the *LR* calculated without taking account of intensity; the dotted line is the mixture dependent calculation; and the dashed line the mixture *in*dependent calculation. At $x = 0$ the *LR* is of the order 10^{-6}, which would be sufficient to reduce the overall six locus *LR* below one, confirming that peak areas can provide a powerful additional tool for discrimination. It is notable that both peak area based *LR*'s pass the fixed *LR* at similar values of x. In this particular example, the mixture independent calculation gives higher *LR*'s, but this is not always the case.

D18

Reference to Table 13 shows that the peak areas support the genotype combination AA, BC. Consider the situation, when the peak areas are:

A	38985
B	1914
C	38985

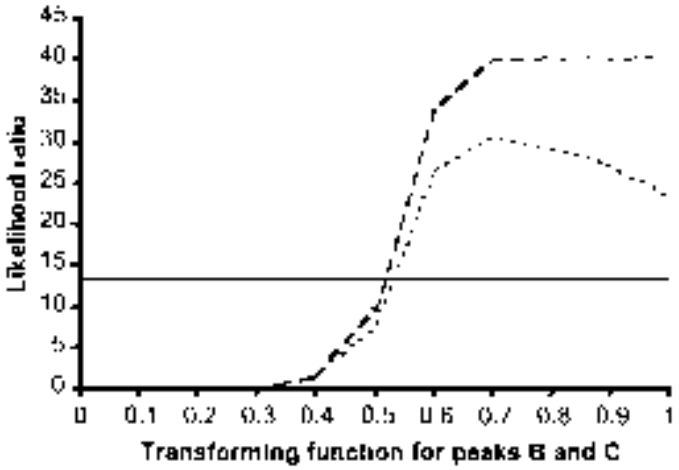

FIG. 3—*Demonstration of the effect on the* LR *of variation in peak area at locus D21. The unbroken line is that calculated ignoring intensity; the dotted line is that taking account of mixture dependent and mixture independent functions; and the dashed line is that which incorporates only the mixture dependent functions.*

The best supported combinations now are of the kind AC, BC and AB, AC and the *LR* for the presence of the suspect's DNA in the mixture is of the order 10^{-4}. To investigate variation of the *LR* with the area of peak C, it was transformed as:

$$C' = A(1 - x) + Cx$$

where the notation has the same sense as before. Figure 4 shows how the mixture dependent and mixture independent *LR*'s varied with x: there is very little to choose between them.

The results from these two loci suggest a way forward based solely on the mixture independent functions which we have called a, b, d and f. Clearly, there is a need to study more complex mixtures, but our preliminary assessments suggest that mixture independent functions will always be computable, however complex the mixture.

Further Development

We recognize that in the present analysis we have considered only one special case: where there are known to be two contributors to a mixture, there is one suspect and one unknown contributor. There will be many different kinds of casework situation, each with its pair of hypotheses to be tested against each other, and

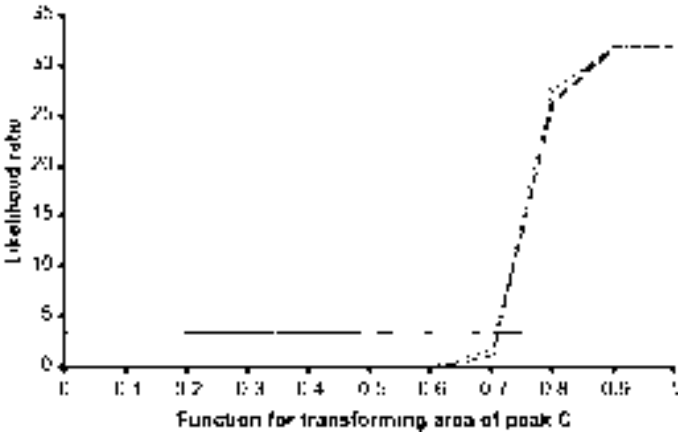

FIG. 4—*Demonstration of the effect on the* LR *of variation in peak area at locus D18. The unbroken line is that calculated ignoring intensity; the dotted line is that taking account of mixture dependent and mixture independent functions; and the dashed line is that which incorporates only the mixture dependent functions.*

each with its own analysis to be carried out. We consider that we have established the principles of a method which takes account of peak area and that those principles can be applied to any particular situation. The complexity of the actual analysis may best be resolved by computer programs of sufficient flexibility and we see the development of such facilities as an important area for the future. There is also a need to combine the requirement for such a system with the capability of dealing impartially with other factors, in particular, peak artefacts, such as pull-up and stutter, and we believe that expert systems technology represents the most profitable path. The first stages in this kind of approach have been described by Gill et al. (8). There is also a need to collect more data. The modeling assumptions that we have adopted are based on a relatively limited set of data. Further data will enable us to refine and extend the modeling: one improvement, for example will be take account of the consideration that peak area ratios should depend on the separation of the peaks. It is also worth noting that the modeling will be protocol dependent and we expect that other workers in the field will want to collect peak area data from mixtures, though a copy of the data used here can be obtained from the authors. We hope that the framework that we have presented here will encourage other workers to carry out similar studies.

References

1. Evett IW, Buffery C, Willott G, Stoney DA. A guide to interpreting single locus profiles of DNA mixtures in forensic cases. J Forensic Sci Soc 1991;31:41–7.
2. National Research Council. The evaluation of DNA evidence. National Academy Press 1996. Washington DC.
3. Weir BS, Triggs CM, Starling L, Stowell LI, Walsh KAJ, Buckleton J. Interpreting DNA mixtures. J Forensic Sci. In press.
4. Sparkes R, Kimpton C, Watson S, Oldroyd N, Clayton T, Barnett L, et al. The validation of a 7-locus multiplex STR test for use in forensic casework (I): Mixtures, ageing, degradation and species studies. Int J Leg Med 1996;109:186–94.
5. Kimpton CP, Oldroyd NJ, Watson SK, Frazier RRE, Johnson PE, Millican ES, et al. Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification. Electrophoresis 1996;17:1283–93.
6. Oldroyd NJ, Urquhart AJ, Kimpton CP, Millican ES, Watson SK, Downes T, et al. A highly discriminating octoplex short tandem repeat polymerase chain reaction system suitable of human individual identification. Electrophoresis 1995;16:334–7.
7. Evett IW, Gill PD, Lambert JA, Oldroyd N, Frazier R, Watson S, et al. Statistical analysis of data for three British ethnic groups from a new STR multiplex. Int J Leg Med. 1997;110:5–9.
8. Gill P, Urquhart A, Millican E, Oldroyd N, Sparkes R, Kimpton C. A new method of STR interpretation using inferential logic—development of a criminal intelligence database. Int J Leg Med 1996; 109:14–22.

Additional information and reprint requests:
Ian W. Evett
Forensic Science Service
Metropolitan Laboratory
109 Lambeth Rd., London
SE1 7LP, UK