Int J Legal Med (1998) 111 : 62–66

ORIGINAL ARTICLE

N. Fukshansky · W. Bär

Interpreting forensic DNA evidence on the basis of hypotheses testing

Received: 18 February 1997 / Received in revised form: 28 July 1997

Abstract Analysis of a mixed biological stain by means of highly polymorphic VNTR systems usually reveals a profile composed of multiple markers. If the victim and one or several suspects match the profile the evidential strength of the matches has to be very carefully analysed. It is shown that a treatment to evaluate mixed stains within the scope of a hypotheses testing approach yields a formula identical to one proposed recently by Weir und coauthors. A simple combinatorial proof of this formula is presented. Further, a special computational formula is derived for the cases where the formula of Weir turns out to be cumbersome.

Key words DNA · Forensic statistics · Mixed stains · Identification

Introduction

DNA analysis of biological stains has become one of the most powerful investigative tools in forensics and is now used worldwide. DNA profiling is especially powerful in detecting stain mixtures. However, it is still controversial how to present this kind of DNA evidence in court. The aim of any quantification in cases of non-exclusion is to statistically evaluate DNA profiles obtained from a stain to identify the assailants out of the group of tested persons – victims and suspects. Propositions for the statistical treatment of relativly simple cases based on the simple sorting out have been made by earlier workers [1–4]. A major progress marks the general formula proposed recently by Weir and coauthors [5] and proved then by Brenner (Appendix in Ref [5]) in a rather complicated way.

N. Fukshansky[1] · W. Bär (✉)
Institut für Rechtsmedizin Universität Zürich-Irchel,
Winterthurerstrasse 190/Bau 52, CH-8057 Zürich, Switzerland
Fax: +41 1 635 68 51

Present address:
[1] Birkenstrasse 4B, D-79256 Buchenbach, Germany

The entire problem can be formulated in terms of hypotheses testing. A DNA mixture from a crime scene – the stain – can contain contributions from one person or from a number of persons: victims, assailants. A comparison of the stain profile with the single person DNA profiles taken from a group of tested persons – victims, suspects – is performed with the aim to identify the assailants. However, such a comparison usually leaves room for more than one alternative.

This situation is typical to apply the theory of hypotheses testing to evaluate the probabilities of these alternatives. Each of mutually excluding hypotheses H_1, H_2, ..., H_N about the circumstances of the crime is a statement specifying members of a group of persons – among them all tested persons and, if necessary, non-tested persons (unknowns) – as either contributors or non-contributors to the crime sample. We call the content of each statement a scenario, Z_i.

According to the scheme of hypotheses testing the information from all DNA tests is accounted for in the calculation of the a-posteriori probability of H_i which is the conditional probability of H_i under the condition of all test results. Under test results or *"test"* we understand the entire set of the crime sample and personal tests. This probability (we designate it as $p(H_i/\text{test}) \equiv W_i$) is:

$$W_i = p(H_i) \cdot X_i / \sum_{j=1}^{N} p(H_j) \cdot X_j \quad (i = 1, \ldots, N), \tag{1}$$

where $X_i \equiv p(test/H_i)$ is nothing more than the probability of the scenario Z_i and $p(H_i)$ is the a-priori probability of the hypothesis H_i. Thus the quantitative assessment of the forensic DNA evidence comes to the calculation of X_i ($i = 1, ..., N$).

In this paper we show that the probability X_i can be calculated with the formula of Weir. Furthermore we give a simple combinatorical proof of this formula and derive another expression for X_i suitable for the cases where the formula of Weir appears very cumbersome.

We consider a number of examples and show that they all fit in the above scheme of hypotheses testing. As this

consideration reveals the common structure for the probabilities of the scenarios this structure is analysed in the next sections. Two subsequent sections contain a short combinatorial proof of the formula of Weir and the derivation of a convenient recursive formula for X_i. A numerical example is given in the last section.

The general structure of the identification problem

Before deriving the general structure of the identification problem on the basis of a number of examples we make a few comments and introduce necessary designations additional to those given in the introduction.

1. The analysis is performed for a polymorphic single locus DNA system. An extension to the case of several loci is trivial – the total probabilities X_i are products of X_i-s obtained for each locus provided that they are independent.

2. As will be seen below a scenario Z_i commonly consists of independent partial scenarios which we will designate as $Z_i(1)$, $Z_i(2)$, ..., $Z_i(m_i)$. Because of the independence of the partial scenarios

$$X_i = X_i(1) \cdot X_i(2) \cdots X_i(m_i).$$

The partial scenarios are of two kinds.

A scenario of the first kind represents one isolated tested person, *T*. The probability – designated $p(T)$ – is then simply the genotype frequency of *T*. A partial scenario of the second kind represents the entire group of the contributors to the stain, which commonly consists of both non-tested and tested persons. In the special cases where all contributors are tested persons, the corresponding partial scenario transforms to a set of partial scenarios of the first kind.

3. We will designate the alleles as A_1, A_2, ... and their frequencies as a_1, a_2, ...

The following assumptions underly the treatment:

- The entire analysis addresses codominant systems. This means that only homozygous persons show a single property in a specific DNA test.
- The non-tested persons (unknown) are not related to any participants (known or unknown) and belong to one ethnic group
- There are no exclusions i.e. all the participants are in concordance with the stain.

Now let us proceed to the examples.

I. *A mixed stain*. The stain showing three alleles A_1, A_2, A_3 is produced by two persons – the victim (*V*) and one non-identified assailant. The victim shows the alleles A_1, A_2. The suspect (*S*) shows the alleles A_2, A_3. Two hypotheses arise.

Hypothesis H_1: *S* is the stain producer

Hypothesis H_2: *S* is not the stain producer, an unknown person (*U*) is the stain producer.

The corresponding scenarios and their probabilities are:

$\mathbf{H_1}$	$Z_1(1)$:	V	(stain)	$X_1(1) = p(V)$		$= 2a_1a_2$
	$Z_1(2)$:	S	(stain)	$X_1(2) = p(S)$		$= 2a_2a_3$
				X_1	$= p(V) \cdot p(S)$	$= 4a_1a_2^2a_3$
$\mathbf{H_2}$	$Z_2(1)$:	S		$X_2(1) = p(S)$		$= 2a_2a_3$
	$Z_2(2)$:	$\{V, U\}$	(stain)	$X_2(2) = ?$		

To settle the probability $X_2(2)$ we need a short discussion. The events *U* and *V* are now not independent, because both of them are the contributers to the stain. The probability for the partial scenario $Z_2(2)$ is deduced from the formula for conditional probability:

$$X_2(2) = p(V) \cdot p(U/V),$$

where $p(U/V)$ is the probability for *U* to be a contributor to the stain under the condition that *V* is a contributor to the stain. Thus the second hypothesis can be represented as dealing with three independent persons: *S*,*V* and an fictive person *U/V*. This fictive person *may have* any alleles from the stain but it *must have* those of them which are not shown by *V* (in our example this is A_3). Hence the phenotype of *U/V* consists of three genotypes $A_3\,A_3$, $A_1\,A_3$, $A_2\,A_3$ and its probability is $a_3^2 + 2a_3(a_1 + a_2)$.

Thus we can paraphrase the hypothesis H_2 as:

$\mathbf{H_2}$	$Z_2(1)$: S	(stain)	$X_2(1) = p(S)$		$= 2a_2a_3$
	$Z_2(2)$: V	(stain)	$X_2(2) = p(V)$		$= 2a_1a_2$
	$Z_2(3)$: U/V	(stain)	$X_2(3) = p(U/V)$		$= a_3(a_3 + 2a_1 + 2a_2)$
			X_2	$= p(V)p(S)p(U/V)$	$= 4a_1a_2^2a_3^2(a_3+2a_1+2a_2)$

Using the same procedure one can easily construct the scenarios for the next somewhat more complicated example.

II. *The rape in group*. Let us assume that three persons have contributed to the stain: the victim and two non-identified assailants. Three persons have been tested: the victim (*V*) and two suspects (*S*1 and *S*2) and none of them is excluded. Four hypotheses settle the matter.

$\mathbf{H_1}$: *S*1 and *S*2 are both contributers to the stain.

$Z_1(1)$: V	(stain)	
$Z_1(2)$: $S1$	(stain)	$X_1 = p(V)p(S1)p(S2)$
$Z_1(3)$: $S2$	(stain)	

$\mathbf{H_2}$: The stain is produced by *V*, *S*1 and an unknown person (*U*); *S*2 does not contribute to the stain.

$Z_2(1)$: $S2$		
$Z_2(2)$: V	(stain)	$X_2 = p(V)p(S1)p(S2)p(U/V\&S1)$
$Z_2(3)$: $S1$	(stain)	
$Z_2(4)$: $U/V\&S1$	(stain)	

$\mathbf{H_3}$: The stain is produced by *V*, *S*2 and an unknown person (*U*); *S*1 does not contribute to the stain.

$Z_3(1)$: $S1$		
$Z_3(2)$: V	(stain)	$X_3 = p(V)p(S1)p(S2)p(U/V\&S2)$
$Z_3(3)$: $S2$	(stain)	
$Z_3(4)$: $U/V\&S2$	(stain)	

H_4: Both *S*1 and *S*2 are not the contributers; the stain is produced by *V* and two unknown persons (*U*1 and *U*2).

$Z_4(1)$: *S*1
$Z_4(2)$: *S*2 $\qquad X_4 = p(V)p(S1)p(S2)p(U1\&U2/V)$
$Z_4(3)$: *V* (stain)
$Z_4(4)$: (*U*1&*U*2)/*V* (stain)

Note that the product of the genotype frequences of all tested persons

$$\Pi = p(V)p(S1)p(S2)$$

is contained in each X_i, i.e.

$X_i = \Pi \cdot \tilde{X}_i$, where
$\tilde{X}_1 = 1$
$\tilde{X}_2 = p(U/V\&S1)$
$\tilde{X}_3 = p(U/V\&S2)$
$\tilde{X}_4 = p(U1\&U2/V)$

On the other hand a factor common to all X_i can be safely cancelled in Eq.(1). This means, that for the calculation of the a-posteriori probability of the hypotheses it is sufficient to know only the values of $\tilde{X}_i$.

The examples shown above facilitate the following general formulation of the identification problem for a mixed stain.

Let us assume that a hypothesis H_1 states that the stain is produced by *n* unknown persons U_1, U_2, ..., U_n and a number of tested persons T_1, T_2, ... Obviously, each of these unknown persons *can* show *only* alleles of the stain and the entire set of these persons in the aggregate *must* show *all* the alleles found in the stain which are not possessed by the tested contributors. We designate the number of these "must"-alleles as *k* and the alleles themselves - as A_1, A_2, ..., A_k. The probability of the event "*n* unknown persons contribute to the stain under condition that a fixed number of tested persons contribute to the stain" we designate as $p(n, k)$:

$$p(U_1\&U_2\& \ldots \&U_n/T_1\&T_2 \ldots) = p(n, k).$$

We see that the a-posteriori probability of a hypothesis H_i can be calculated as

$$W_i = p(H_i) \cdot \tilde{X}_i / \sum_{j=1}^{N} p(H_j) \cdot \tilde{X}_j, \text{ where}$$
$\tilde{X}_i = p(n, k)$ for $n > 0$
$\tilde{X}_i = 1$ for $n = 0$

Thus the entire identification study focuses on the calculation of the probabilities $p(n, k)$.

Solving the problem of hypotheses evaluation

Consider a stain showing m alleles A_1, A_2, ..., A_m with the frequencies in the population a_1, a_2, ..., a_m respectively. According to the current hypothesis the contributors to this stain are *n* unknown and a number of tested persons. The group of *n* unknown contributors *can have* only alleles from the stain and *must have k* alleles A_1, A_2, ..., A_k ($k \le m$, $k \le 2n$) which are not contained in the genotypes of the tested contributors. The probability of this event was designated $p(n, k)$.

To evaluate this probability we apply the following combinatorial reasoning. Let us view the genotypes of unknown contributors as ordered pairs A_iA_j (with the probability of appearance a_ia_j) so that A_iA_j is different from A_jA_i if $i \ne j$. Let us also consider a sequence of 2*n* places which should be occupied by the 2*n* alleles of *n* unknown persons. This sequence of 2*n* places should be filled by alleles belonging to the stain, but the first *k* alleles A_1, A_2, ..., A_k must be contained in the sequence at least once, i.e. *k* places are necessarily presented for them. The rest of this sequence containing 2*n*–*k* places can be occupied by any alleles from the stain. The number of times an allele A_i ($i = 1, \ldots, m$) is met in the rest sequence we designate n_i, ($0 \le n_i \le 2n - k$, $n_1 + n_2 + \ldots + n_m = 2n–k$).

The number of different variants to fill 2*n* places with *m* alleles under the condition that the first *k* alleles are met $n_i + 1$ times each, while the other *m*–*k* alleles are met n_i times each ($0 \le n_i \le 2n–k$) is given by the polynomial coefficient:

$$\frac{(2n)!}{(n_1 + 1)! \cdots (n_k + 1)! n_{k+1}! \cdots n_m!} \qquad (2)$$

and the probability of each of these distributions is:

$$a_1^{n_1+1} \cdots a_k^{n_k+1} a_{k+1}^{n_k+1} \cdots a_m^{n_m}.$$

The probability $p(n, k)$ to be found is the sum of the probabilities of all different distributions:

$$p(n, k) = \sum_{\substack{n_1+\cdots+n_m=2n-k \\ n_i \ge 0}} \frac{(2n)! a_1^{n_1+1} \cdots a_k^{n_k+1} a_{k+1}^{n_{k+1}} \cdots a_m^{n_m}}{(n_1 + 1)! \cdots (n_k + 1)! n_{k+1}! \cdots n_m!} \qquad (3)$$

The expression (3) in principle gives the general solution for $p(n, k)$. However, to compute the *m*-fold sum (3) is a very complicated problem, especially for large *m* and $2n - k$. Therefore we will derive in the next two sections a more convenient formula for this purpose.

A computational formula for $p(n, k)$ for smaller k values

Formula (3) can be transformed in a following way. Consider the polynomial formula

$$(x_1 + x_2 + \ldots + x_m)^n = \sum_{\substack{n_1+\cdots+n_m=n \\ n_i \ge 0}} \frac{n! x_1^{n_1} \cdots x_m^{n_m}}{n_1! \cdots n_m!} \qquad (4)$$

One can easily see that on the basis of (4) for k=0 Eq.(3) acquires the form

$$p(n, 0) = \sum_{\substack{n_1+\cdots+n_m=2n \\ n_i \ge 0}} \frac{(2n)! a_1^{n_1} \cdots a_m^{n_m}}{n_1! \cdots n_m!} = s^{2n}, \qquad (5)$$

where

$$s = a_1 + a_2 + \cdots + a_m \qquad (6)$$

is the sum of the allele frequencies for all the alleles detected in the stain.

The formula (5) can be seen as the probability of the distribution of $\{n_i\}$ when each of n_i (i = 1, ..., m) can also have a zero value. To obtain the corresponding probability for the case $k > 0$ one should subtract from the probability given by (5) the probability S of the following event: at least one of n_1, n_2, ..., n_k is zero. Let us derive the formula for the probability S.

We designate the probabilities of events $\{n_i = 0\}$, $\{n_i = 0, n_i = 0\}$,$\{n_i = 0, n_j = 0, n_l = 0\}$, ... as p_i; $p_{i,j}$; $p_{i,j,l}$; ... respectively and also introduce the sums

$$S_1 = \sum p_i;\ S_2 = \sum p_{i,j};\ S_3 = \sum p_{i,j,l};\ \ldots \quad i < j < l < \ldots < k$$

where S_r is the probability that exactly r items of n_1,..., n_k are equal to zero. Then the probability S that at least one of them is zero can be calculated according to the formula [6]:

$$S = S_1 - S_2 + S_3 - S_4 + ... + (-1)^{k-1} S_k$$

providing for the probability $p(n, k)$:

$$p(n, k) = p(n, 0) - S = s^{2n} - S_1 + S_2 - ... + (-1)^k S_k. \qquad (7)$$

The probabilities p_i; $p_{i,j}$; $p_{i,j,l.}$; ... can be easily computed with the formula (5) setting there $\{n_i = 0\}$; $\{n_i = 0, n_j = 0\}$; $\{n_i = 0, n_j = 0, n_i = 0\}$;... respectively:

$$p_i = \sum_{\substack{n_1+\cdots+n_m = 2n \\ n_i = 0}} \frac{(2n)!a_1^{n_1}\cdots a_{i-1}^{n_{i-1}} a_{i+1}^{n_{i+1}}\cdots n_m^{n_m}}{n_1!\cdots n_{i-1}! n_{i+1}!\cdots n_m!} = (s - a_i)^{2n} \qquad (8)$$

$$p_{i,j} = (s - a_i - a_j)^{2n};$$

$$p_{i,j,l} = (s - a_i - a_j - a_l)^{2n}; \ldots$$

Formulae (7, 8) are identical to the formula of Weir and are convenient for the calculations in the case of rather small k. For larger k the number of terms grows rapidly making the calculations cumbersome. An approach specifically designed for large k is therefore needed.

A computational formula for $p(n, k)$ for the case of large k values

The search for a formula for large k is substantiated by the following argument. Note that for $k = 2n$ the formula (7) contains $2n$ items. On the other hand in this case all the alleles of n unknown persons are different and allow for $(2n)!$ permutations which leads to a simple formula

$$p(n, 2n) = (2n)!a_1 a_2 \cdots a_{2n}$$

containing only one term.

When $k = 2n - 1$ then one place can be occupied either by an allele from the group a_{k+1}, ..., a_m with the probability $(2n)!a_1 ... a_k \cdot (a_{k+1} + ... + a_m)$ or, by an allele from the group a_1, ..., a_k with the probability $(2n)!a_1 ... a_k(a_1 + ... + a_k)/2!$. The factor 2! is necessary because the corresponding $A_i (i \le k)$ appears twice and their permutations are indistinguishable. Designating

$$a_{k+1} + ... + a_m = q \qquad (9)$$

we obtain for $k = 2n - 1$:

$$p(n, k) = (2n)!a_1 \cdots a_k(q + \sum_{i=1}^{k} a_i/2!) = (2n)!a_1 \cdots a_k(s + q)/2.$$

On the basis of analogous combinatorial reasoning the following general formula for the calculation of $p(n, k)$ in case of $k \le 2n$ can be obtained:

$$p(n, k = 2n - r) = (2n)!a_1 a_2 \cdots a_k \sum_{j=0}^{r} \frac{q^{r-j} Q(k, j)}{(r - j)!}, \qquad (10)$$

where

$$Q(k, j) = \sum_{\substack{n_1+\cdots+n_k = j \\ 0 \le n_i \le j}} \frac{a_1^{n_1}\cdots a_k^{n_k}}{(n_1 + 1)!\cdots(n_k + 1)!}.$$

Importantly, the term $Q(k, j)$ can be calculated in a recursive way. Note that especially n_k can take values i = 0,1,..., j. Thus

$$Q(k, j) = \sum_{i=0}^{j} \frac{a_k^i}{(i + 1)!} \sum_{\substack{n_1+\cdots+n_{k-1} = j-i \\ 0 \le n_r \le j - i}} \frac{a_1^{n_1}\cdots a_k^{n_{k-1}}}{(n_1 + 1)!\cdots(n_{k-1} + 1)!}.$$

The internal sum is nothing else than Q(k–1, j–i). Therefore we arrive at the following recursive formula for $Q(k, j)$:

$$Q(k, j) = \sum_{i=0}^{j} \frac{a_k^i}{(i + 1)!} Q(k - 1, j - i)$$

$$Q(k, 0) = 1 \qquad (11)$$

$$Q(0, j) = 0 \quad \text{for} \quad j > 0.$$

The formulae (7, 8) and (10, 11) are convenient since they contain explicitly only the frequencies of the alleles obligatory for the unknown persons included in the stain while the other alleles are present in a "cryptic" way – within the terms s and q and the sums are at the most n-fold.

Combining these formulae one can easily determine $p(n, k)$ for n = 0,1,2.

$$p(0, 0) = 1$$

$$p(1, 0) = s^2$$

$$p(1, 1) = s^2 - (s - a_1)^2$$

$$p(1, 2) = 2a_1 a_2$$

$$p(2, 0) = s^4 \qquad (12)$$

$$p(2, 1) = s^4 - (s - a_1)^4$$

$$p(2, 2) = s^4 - (s - a_1)^4 - (s - a_2)^4 + (s - a_1 - a_2)^4$$

$$p(2, 3) = 12a_1 a_2 a_3 (q + s)$$

$$p(2, 4) = 24a_1 a_2 a_3 a_4$$

Numerical example

In order to illustrate the application of the theory we address the situation described in the example 2 and consider results of the tests performed in 3 loci (3 systems).

Table 1 The entire test results – the stain and single person tests for the victim and two suspects – as well as the frequencies of the alleles found for the three systems

System	Stain	Frequency	V	*S*1	*S*2
DQa	1.1	0.134	+		+
	1.2	0.170		+	
	1.4	0.324	+	+	+
FES	10	0.327			+
	11	0.396	+	+	
	12	0.224	+		
	13	0.032			+
*F*13*A*1	3	0.082	+		
	5	0.174			+
	6	0.314	+		+
	7	0.341		+	
	15	0.026		+	

Table 2 Values of s and explicit n, k values together with the frequencies a_1, ..., a_k of corresponding alleles for each hypotheses H_i

System	s	H_1	H_2	H_3	H_4
DQa	0.628	$p(0,0)$	$p(1,0)$	$p(1,1)$ $a_1 = 0.170$	$p(2,1)$ $a_1 = 0.170$
FES	0.979	$p(0,0)$	$p(1,2)$ $a_1 = 0.327$ $a_2 = 0.032$	$p(1,0)$	$p(2,2)$ $a_1 = 0.327$ $a_2 = 0.032$
*F*13*A*1	0.937	$p(0,0)$	$p(1,1)$ $a_1 = 0.174$	$p(1,2)$ $a_1 = 0.341$	$p(2,3)$ $a_1 = 0.174$ $a_2 = 0.341$ $a_3 = 0.026$

These results are shown in Table 1, which presents the alleles found in the stain, the frequencies of these alleles in a population and also the single person tests of the three tested persons (one victim and two suspects *S*1 and *S*2) for the three investigated systems.

In Table 2 we give the s values and also the explicit numbers n and k in the form of $p(n, k)$, as well as the corresponding values of a_1, a_2, ..., a_k for each hypothesis H_i.

Table 3 Results of the calculations: the quantities $\tilde{X}_i$ for each single locus system and for the combination of the 3 systems; the likelihood ratios L_i (i = 2,3,4) for the hypotheses H_i as compared to the hypothesis H_1

System	$log_{10}\tilde{X}_i + 10$			
	H_1	H_2	H_3	H_4
DQa	10	9.5959	9.2663	9.0474
FES	10	8.3207	9.9816	8.9106
*F*13*A*1	10	9.4710	8.2488	8.3923
3 systems	10	7.3876	7.4967	6.3503
$\tilde{X}_i$	1	$2.44 \cdot 10^{-3}$	$3.14 \cdot 10^{-3}$	$2.24 \cdot 10^{-4}$
$L_i = \tilde{X}_1/\tilde{X}_i$		409.6	113.1	4463.8

Table 3 presents the quantities $\tilde{X}_i$ for each single locus and for the entire multilocus system and the likelihood ratios L_i (i = 2, 3, 4) for the hypotheses H_i as compared to the hypothesis H_1.

References

1. Evett IW, Buffery C, Willott G, Stoney D (1991) A guide to interpreting single locus profiles of DNA mixtures in forensic cases. J Forensic Sci Soc 31:41–47
2. Brenner CH, Fimmers R, Baur MP (1996) Likelihood ratios for mixed stains when the number of donors cannot be agreed. Int J Legal Med 109:218–219
3. The evaluation of forensic DNA evidence, National Academy Press. Washington DC (1996)
4. Hummel K, Bär W, Fukshansky N (1995) Identification of biological stains: probability of identity or of kinship. In: Carracedo A, Brinkmann B, Bär W (eds) Advances in Forensic Haemogenetics 6. Springer, Berlin Heidelberg New York, pp 90–92
5. Weir BS, Triggs CM, Starling L, Stowell LL, Watsh KAJ, Buckleton J (1997) Interpreting DNA Mixtures. J Forensic Sci 42: 213–219
6. Feller W (1967) An introduction to probability theory and its applications. New York John Wiley & Sons, Inc. London Chapman & Hall