ORIGINAL ARTICLE

Ricardo Cao · Jorge Alemany · Carmen Cabrero
Angel Carracedo · Ana Díez · Emilio Valverde

Semiparametric approach to match probability calculations using single locus probes

Received: 20 November 1995 / Accepted: 30 January 1996

Abstract A semiparametric approach to match probability calculations using single locus probes has been developed and compared graphically with other standard methods by a one-sample simulation. The density functions obtained using this method are closer to the real distributions than those obtained by conventional approaches. Our method does not need to establish an arbitrary match threshold, which has been a source of problems in practical applications of standard methods. Moreover, it can be adjusted to any particular conditions by setting the experimental error and correlation of each laboratory. To assess the practical performance of this method we carried out a comparison experiment using a sample of 229 individuals analysed in duplicate.

Key words Semiparametric approach · Single locus probes · Match probability

Introduction

DNA profiling has become a popular method in forensic investigation of crime and disputed paternity cases. Undoubtedly, it is a very powerful technique since DNA can be obtained from minimal amounts of any human tissue, even from degraded biological specimens that are frequently found at crime scenes. Considerable effort has been made to standardise the laboratory techniques and instruments used to carry out the analyses (Kearney et al. 1989; Schneider et al. 1991; Gill et al. 1992). However, the statistical methods used to evaluate the information contained in DNA profiles still cause controversy among forensic scientists (Gill et al. 1990; Evett and Gill 1991; Budowle et al. 1991; Berry 1991; Berry et al. 1992; Devlin et al. 1992; Evett et al. 1992; Monson and Budowle 1993). When conventional genetic markers are used, the comparison between two samples is straightforward. The alleles of these markers are discrete variables and a match is declared when two samples have identical phenotypes across all loci, otherwise an exclusion is considered because the probability that the same individual produces two different genotypes for the same locus is zero. When two samples match, the probability that they come from the same individual can be given on the basis of the frequency distribution of the alleles of the different loci used in the analysis. With VNTR systems, the statistical inference is more difficult. In a VNTR locus, alleles are DNA fragments of different lengths which are measured indirectly using gel electrophoresis and compared with standard weight markers. This results in some measurement error associated with the fragment size estimation which is often greater than the repeat unit size. Therefore, it is not possible to distinguish close alleles with absolute certainty. Hence there is a continuous allele distribution in which discrete allelic frequencies cannot be assigned. When analysing VNTR polymorphisms it is necessary to adopt a statistical method which allows the analyst to give a probability of a match between two bands appearing to have the same origin. A number of methods have been proposed to determine this probability (Baird et al. 1986; Gill et al. 1990; Budowle et al. 1991; Pascali et al. 1991), most of which are based on the subdivision of the continuous distribution of fragment sizes into arbitrary intervals. The frequency of a fragment whose molecular size falls into a specific interval is taken to be the frequency of the overall number of fragments that are contained in that subdivision. These discrete frequencies can be used in a very similar way to the conventional methods when analysing blood group loci and serum proteins, thus making it easier to understand by non-expert personnel involved in a trial. However, these kinds of approaches have led to controversy, mainly for the following reasons:

R. Cao
Department of Mathematics, Faculty of Computer Sciences,
E-15701 La Coruña, Galicia, Spain

J. Alemany · C. Cabrero · A. Díez · E. Valverde (✉)
Department of Molecular Biology, Pharmagen, Calera 3,
E-28760 Madrid, Spain

A. Carracedo
Institute of Legal Medicine, Faculty of Medicine,
E-15705 Santiago de Compostela, Galicia, Spain

– the existence or not of a match is declared in a subjective way because the alleles cannot be determined unambiguously

– a match is defined in a switch-like fashion that changes to non-match in arbitrarily established points

– the probability values obtained are higher than those obtained with conventional markers, but the probability is underestimated since the scoring system used is very conservative

Recently, some approaches have made use of density functions to estimate match probabilities in VNTR loci (Devlin et al. 1991; Berry et al. 1992; Evett et al. 1992). In this paper, we present a model based on a semiparametric estimation of density functions, and a subsequent calculation of the probability of a match between two bands by means of a reformulation of the Bayes' theorem in terms of the conditional density functions. This method overcomes the disadvantages discussed above. The method is later extended to the comparison of two-banded profiles, taking account of the correlation observed in the measurement errors of the two bands, further complicating the calculations. The practical performance of this method has been compared by means of a one-sample simulation with some other standard approaches such as the histogram, sliding window or fixed bin methods. The theoretical model (from which the data were simulated) gives a similar distribution of frequencies to those appearing in practice. Finally, we carried out an experiment similar to that described by Evett et al. (1992) using a data set of 229 individuals analysed in duplicate, from which we have extracted information about comparisons both between and within persons.

Estimation of allele frequencies

In the construction of databases of VNTR polymorphisms, the major statistical concern is the problem of estimating the probability density function of the fragment lengths (X) when the observed variable is only an approximation.

The following model is considered:

$$\ln Y = \ln m\,(X) + \varepsilon$$

or in other terms:

$$Y = m(X) \cdot e^{\varepsilon}$$

where ε is a random error term (due to measurement errors) independent of X (the true fragment length) in the distance domain, assumed to be normal with zero mean and variance σ^2 and m is a real function which takes into account not only the linear dependence between X and Y (the observed fragment length), but also more general relationships. See Valverde et al. (1993) for motivation of the model and estimation of the density function of X.

In our practical casework the function m has been estimated and can be accepted as the identity function $m(x) = x$. Consequently, the previous model becomes considerably more simpler:

$$\ln Y = \ln X + \varepsilon \qquad (1)$$

Match probability

One-band case

We can consider a factual case where two bands corresponding to different samples (e.g., presumptive father and son) are so close that they can be considered to have the same origin. The experimentally determined size of each band will be denoted herein as Y_1 and Y_2, respectively. We assign A to the event *"the two fragments have the same origin"*. An *a priori* probability will be assigned to this event, $P(A)$, given the basis of previous evidence. If the event A is true this means, using equation (1), that

$$\ln Y_1 = \ln X + \varepsilon_1$$
$$\ln Y_2 = \ln X + \varepsilon_2$$

given that ε_1 and ε_2 are independent errors, with normal distribution and variances σ^2, and are also independent with respect to X (the common true fragment length).

If the event A is not true, then

$$\ln Y_1 = \ln X_1 + \varepsilon_1$$
$$\ln Y_2 = \ln X_2 + \varepsilon_2$$

where ε_1, ε_2, X_1 and X_2 are independent of each other. This means that $\ln Y_1$ and $\ln Y_2$ are two independent variables with the same distribution.

Denote $Z_1 = \ln Y_1$ and $Z_2 = \ln Y_2$. In practice we observe two values z_1 and z_2 ($z_1 = \ln y_1$ and $z_2 = \ln y_2$). In this case, the Bayes rule can be reformulated to calculate the *a posteriori* probability of event A, having observed y_1 and y_2, in terms of the conditional density functions:

$$P(A \mid {}^{*}_{y_1, y_2}) = P(A \mid {}^{*}_{Z_1 = z_1, Z_2 = z_2}) =$$
$$= \frac{g(z_1, z_2 \mid {}^{*}_{A}) \cdot P(A)}{g(z_1, z_2 \mid {}^{*}_{A}) \cdot P(A) + g(z_1, z_2 \mid {}^{*}_{A}) \cdot P(\bar{A})} \qquad (2)$$

where

$g(z_1, z_2 \mid {}^{*}_{A}) = \int g(z_1, z_2 \mid {}^{*}_{A,\, Q = q}) f^{Q}(q)dq$, $Q = \ln X$ with density f^{Q} and $g(z_1, z_2 \mid {}^{*}_{A,\, Q = q})$ is the conditional density of (Z_1, Z_2) to the event A and to $Q = q$.

$g(z_1, z_2 \mid {}^{*}_{\bar{A}}) = f^{Z}(z_1) \cdot f^{Z}(z_2)$, where f^{Z} is the density function of Z.

Note that, since ε_1, ε_2 and Q are independent (under A), then

$$g(z_1, z_2 \mid {}^{*}_{A,\, Q = q}) = \varphi(z_1 - q) \cdot \varphi(z_2 - q)$$

φ being the density of a normal distribution with zero mean and variance σ^2.

In this manner, the two terms to be estimated in (2) are:

$$g(z_1, z_2 \mid {}^{*}_{A}) = \int \varphi(z_1 - q) \cdot \varphi(z_2 - q)\, f^{Q}(q)dq \qquad (3)$$

$$g(z_1, z_2 \mid {}^{*}_{\bar{A}}) = f^{Z}(z_1) \cdot f^{Z}\,(z_2) \qquad (4)$$

Expression (4) can be directly estimated by means of a kernel method:

$$\hat{g}\left(z_1, z_2\big|_{\bar{A}}\right) = \hat{f}_h^{z}(z_1) \cdot \hat{f}_h^{Z}(z_2) =$$
$$\left[\frac{1}{n}\sum_{i=1}^{n} K_h(z_1 - Z_i)\right] \cdot \left[\frac{1}{n}\sum_{i=1}^{n} K_h(z_2 - Z_i)\right] \qquad (5)$$

where

$$K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$$

and K is the gaussian kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

To estimate term (3) we need to estimate f^Q (q) and σ^2. To do this, a semiparametric method based on kernel estimation (Valverde et al. 1993) has been used:

$$\hat{f}^Q(q) = \frac{1}{n}\sum_{i=1}^{n} K_{\sqrt{h^2 - \hat{\sigma}^2}}(q - Z_i) \qquad (6)$$

$$\hat{\sigma}^2 = \ln(1 + S^2) \qquad (7)$$

where S^2 is the empirical variance of the values $Y_i/\hat{m}(X_i)$ for a preliminary sample (X_i, Y_i) where the true fragment length is known and $\hat{m}$ is a nonparametric regression estimation of Y given X. This step becomes very important in calibrating the laboratory experimental error.

Assume we have a database $D_1, D_2, \ldots, D_n$ of logarithms of observed fragment lengths. Now, using (6) and (7):

$$\hat{g}\left(z_1, z_2\big|_A\right) = \int K_{\hat{\sigma}}(z_1 - q) K_{\hat{\sigma}}(z_2 - q) \frac{1}{n}\sum_{i=1}^{n} K_{\sqrt{h^2 - \hat{\sigma}^2}}(q - D_i)\, dq =$$
$$= \frac{1}{n}\sum_{i=1}^{n} \int K_{\hat{\sigma}}(z_1 - q) K_{\hat{\sigma}}(z_2 - q) K_{\sqrt{h^2 - \hat{\sigma}^2}}(q - D_i)\, dq =$$
$$= \frac{1}{n\left(\sqrt{2\pi}\right)^3 \hat{\sigma}^2 \left(h^2 - \hat{\sigma}^2\right)^{1/2}} \sum_{i=1}^{n} \int \exp\left(-\frac{(z_1 - q)^2}{2\hat{\sigma}^2}\right)$$
$$\exp\left(-\frac{(z_2 - q)^2}{2\hat{\sigma}^2}\right) \exp\left(-\frac{(q - D_1)^2}{2\left(h^2 - \hat{\sigma}^2\right)}\right) dq$$
$$= \frac{1}{n\left(\sqrt{2\pi}\right)^3 \hat{\sigma}^2 \left(h^2 - \hat{\sigma}^2\right)^{1/2}} \sum_{i=1}^{n} L(z_1, z_2, D_1) \qquad (8)$$

Given that:

$$L(z_1, z_2, Z_i) = \sqrt{2\pi} \sqrt{\frac{\hat{\sigma}^2\left(h^2 - \hat{\sigma}^2\right)}{2h^2 - \hat{\sigma}^2}}$$
$$\cdot \exp\left(\frac{-h^2(z_1 - z_2)^2 + 2\hat{\sigma}^2\left[D_i(z_1 + z_2) - z_1 z_2 - D_i^2\right]}{2\hat{\sigma}^2\left(2h^2 - \hat{\sigma}^2\right)}\right) \qquad (9)$$

Operating with (8) and (9), we obtain:

$$\hat{g}\left(z_1, z_2\,|_A\right) = \frac{1}{n 2\pi \hat{\sigma} \sqrt{2h^2 - \hat{\sigma}^2}}$$
$$\sum_{i=1}^{n} \exp\left(\frac{-h^2(z_1 - z_2)^2 + 2\hat{\sigma}^2\left[D_i(z_1 + z_2) - z_1 z_2 - D_i^2\right]}{2\hat{\sigma}^2\left(2h^2 - \hat{\sigma}^2\right)}\right) \qquad (10)$$

In this way, the estimation of $P(A|_{y_1, y_2})$, or the probability that the two fragments share the same origin, results in:

$$\hat{P}\left(A\big|_{y_1, y_2}\right) = \frac{\hat{g}\left(z_1, z_2\,|_A\right) \cdot P(A)}{\hat{g}\left(z_1, z_2\,|_A\right) \cdot P(A) + \hat{g}\left(z_1, z_2\,|_{\bar{A}}\right) \cdot P(\bar{A})} \qquad (11)$$

which can be obtained from equations (5) and (10).

If we consider a prior probability $P(A) = 0.5$, then $P(\bar{A}) = 1 - P(A) = 0.5$, and the formula (11) can be simplified as:

$$\hat{P}\left(A\big|_{y_1, y_2}\right) = \frac{\hat{g}\left(z_1, z_2\,|_A\right)}{\hat{g}\left(z_1, z_2\,|_A\right) + \hat{g}\left(z_1, z_2\,|_{\bar{A}}\right)} \qquad (12)$$

This formula is of general use in paternity cases, which are usually reported in terms of percentage probability. However, a simple transformation allows the likelihood ratio values to be obtained.

Two-band case

In a criminalistic context, we usually compare two specimens each containing two bands which are suspected to have the same origin. In this situation, we must take account of the correlation between the measurement errors of each pair of bands. The experimentally determined size of each pair of bands will be denoted herein as (Y_1, Y_2) and (Y_3, Y_4). The correlation between measurement errors means that if Y_1 is greater than Y_3, it is very possible that, when the samples match, Y_2 will be greater than Y_4. As in the one-band case, we assign A to the event *"the two fragments have the same origin."*

If the event A is true, this means, using equation (1), that

$$\ln Y_1 = \ln X_1 + \varepsilon_1 \qquad \ln Y_3 = \ln X_1 + \varepsilon_3$$
$$\ln Y_2 = \ln X_2 + \varepsilon_2 \qquad \ln Y_4 = \ln X_2 + \varepsilon_4$$

where $(\varepsilon_1, \varepsilon_2)$ and $(\varepsilon_3, \varepsilon_4)$ are independent errors, with bivariate normal distribution with zero mean and covariance matrix given by

$$\sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

where ρ is the correlation coefficient of the measurement errors, and are also independent of X_1 and X_2 (the common true fragment lengths of each allele).

If the event A is not true, then

$$\ln Y_1 = \ln X_1 + \varepsilon_1 \qquad \ln Y_3 = \ln X_3 + \varepsilon_3$$
$$\ln Y_2 = \ln X_2 + \varepsilon_2 \qquad \ln Y_4 = \ln X_4 + \varepsilon_4$$

where $(\varepsilon_1, \varepsilon_2)$, $(\varepsilon_3, \varepsilon_4)$, X_1, X_2, X_3, and X_4 are independent of each other and the error distribution is as specified above. As a consequence (Y_1, Y_2) and (Y_3, Y_4) are independent vectors with the same distribution. Denote $Z_i = \ln Y_i$, $i = 1, 2, 3, 4$. In practice we observe two pairs (z_1, z_2) and (z_3, z_4), where $z_i = \ln y_i$, $i = 1, 2, 3, 4$. In this case, Bayes' rule applies to calculate the posterior probability of the event A, having observed (y_1, y_2) and (y_3, y_4),

$$P\left(A\Big|_{y_1,y_2,y_3,y_4}\right) = P\left(A\Big|_{Z_i=z_i,i=1,2,3,4}\right)$$
$$= \frac{g\left(z_1,z_2,z_3,z_4|_A\right)\cdot P(A)}{g\left(z_1,z_2,z_3,z_4|_A\right)\cdot P(A) + g\left(z_1,z_2,z_3,z_4|_{\bar{A}}\right)\cdot P\left(\bar{A}\right)} \tag{13}$$

where g denotes the joint four-dimensional density function of (z_1, z_2, z_3, z_4) given A or given $\bar{A}$. An equivalent formula relies on the so-called likelihood ratio

$$P\left(A\Big|_{y_1,y_2,y_3,y_4}\right) = \frac{LR\cdot P(A)}{LR\cdot P(A) + P\left(\bar{A}\right)} \tag{14}$$

where

$$LR = \frac{g\left(z_1.z_2,z_3,z_4|_A\right)}{g\left(z_1.z_2,z_3,z_4|_{\bar{A}}\right)} \tag{15}$$

Given a database $D_1, D_2, \ldots, D_n$ of logarithms of observed lengths and using the kernel method with gaussian kernel, tedious calculations lead to some estimations of the terms in the numerator and denominator of (15):

$$\hat{g}\left(z_1.z_2,z_3,z_4|_A\right) = \frac{1}{2(2\pi)^2 n^2\hat{\sigma}^2\sqrt{\left(1-\hat{\rho}^2\right)\left(2h^2-\hat{\sigma}^2+\hat{\rho}\hat{\sigma}^2\right)\left(2h^2-\hat{\sigma}^2-\hat{\rho}\hat{\sigma}^2\right)}}\left(\sum_{i,j=1}^{n}\left(A_{ij}^{(3)}+A_{ij}^{(4)}\right)\right)$$

$$\hat{g}\left(z_1,z_2,z_3,z_4|_{\bar{A}}\right) = \frac{1}{(2\pi)^2 n^4\left(h^4-\hat{\rho}^2\hat{\sigma}^4\right)}\left(\sum_{i,j=1}^{n}A_{ij}^{(1)}\right)\left(\sum_{i,j=1}^{n}A_{ij}^{(2)}\right)$$

where

$$A_{ij}^{(1)} = \exp\left(-\frac{h^2\left(z_1-D_i\right)^2+h^2\left(z_2-D_j\right)^2-2\hat{\rho}\hat{\sigma}^2\left(z_1-D_i\right)\left(z_2-D_j\right)}{2\left(h^4-\hat{\rho}^2\hat{\sigma}^4\right)}\right)$$

$$A_{ij}^{(2)} = \exp\left(-\frac{h^2\left(z_3-D_i\right)^2+h^2\left(z_4-D_j\right)^2-2\hat{\rho}\hat{\sigma}^2\left(z_3-D_i\right)\left(z_4-D_j\right)}{2\left(h^4-\hat{\rho}^2\hat{\sigma}^4\right)}\right)$$

$$A_{ij}^{(3)} = \exp\left(-\frac{b_1\left(c_{1i}^2+c_{2j}^2+c_{3i}^2+c_{4j}^2\right)+2b_2\left(c_{1i}\cdot c_{2j}+c_{3i}\cdot c_{4j}\right)+2b_3\left(c_{1i}\cdot c_{3i}+c_{2j}\cdot c_{4j}\right)+2b_4\left(c_{1i}\cdot c_{4j}+c_{2j}\cdot c_{3i}\right)}{2\left(1-\hat{\rho}^2\right)\hat{\sigma}^4\left(2h^2-\hat{\sigma}^2+\hat{\rho}\hat{\sigma}^2\right)\left(2h^2-\hat{\sigma}^2-\hat{\rho}\hat{\sigma}^2\right)}\right)$$

$$A_{ij}^{(4)} = \exp\left(-\frac{b_1\left(c_{1i}^2+c_{2j}^2+c_{3j}^2+c_{4i}^2\right)+2b_2\left(c_{1i}\cdot c_{2j}+c_{3j}\cdot c_{4i}\right)+2b_3\left(c_{1i}\cdot c_{4i}+c_{2j}\cdot c_{3j}\right)+2b_4\left(c_{1i}\cdot c_{3j}+c_{2j}\cdot c_{4i}\right)}{2\left(1-\hat{\rho}^2\right)\hat{\sigma}^4\left(2h^2-\hat{\sigma}^2+\hat{\rho}\hat{\sigma}^2\right)\left(2h^2-\hat{\sigma}^2-\hat{\rho}\hat{\sigma}^2\right)}\right)$$

$$b_1 = h^2\hat{\sigma}^2\left(2h^2-\hat{\sigma}^2-\hat{\rho}\hat{\sigma}^2\right),$$
$$b_2 = -\hat{\rho}\hat{\sigma}^2\left(2h^2\left(h^2-\hat{\sigma}^2\right)+\hat{\sigma}^4\cdot\left(1-\hat{\rho}^2\right)\right),$$
$$b_3 = -\left(h^2-\hat{\sigma}^2\right)\hat{\sigma}^2\left(2h^2-\hat{\sigma}^2+\hat{\rho}^2\hat{\sigma}^2\right),$$
$$b_4 = 2h^2\hat{\rho}\hat{\sigma}^2\left(h^2-\hat{\sigma}^2\right),$$
$$c_{ki} = z_k - D_i,\ \text{for } i = 1,2,\ldots,n \text{ and } k = 1,2,3,4.$$

h is the bandwidth in the kernel method and $\hat{\sigma}$ and $\hat{\rho}$ are suitable estimators of σ and ρ. To be more precise, h has been chosen to be the smoothed cross-validation bandwidth, i.e. a plausible estimator of the bandwidth that minimizes the distance between the whole curve of the true density function and its kernel estimator. This distance is measured in terms of the mean integrated squared error (MISE). See Hall et al. (1992) for details.

Error and correlation estimation

Although no observed sample $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ is available for ε and only $y_1, y_2, \ldots, y_n$ are observed, repeated measures pertaining to the same true fragment length (i.e., $\ln y_i = \ln x + \varepsilon_i$, $i = 1, 2, \ldots, n$) are enough to estimate σ^2 by

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\varepsilon_i-\bar{\varepsilon}\right)^2 = \frac{1}{2n(n-1)}\sum_{i\neq j}\left(\varepsilon_i-\varepsilon_j\right)^2 =$$
$$= \frac{1}{2n(n-1)}\sum_{i\neq j}\left(\ln y_i-\ln y_j\right)^2$$

which is observable. This is easily extended to the case where different values of x are available. In this situation,

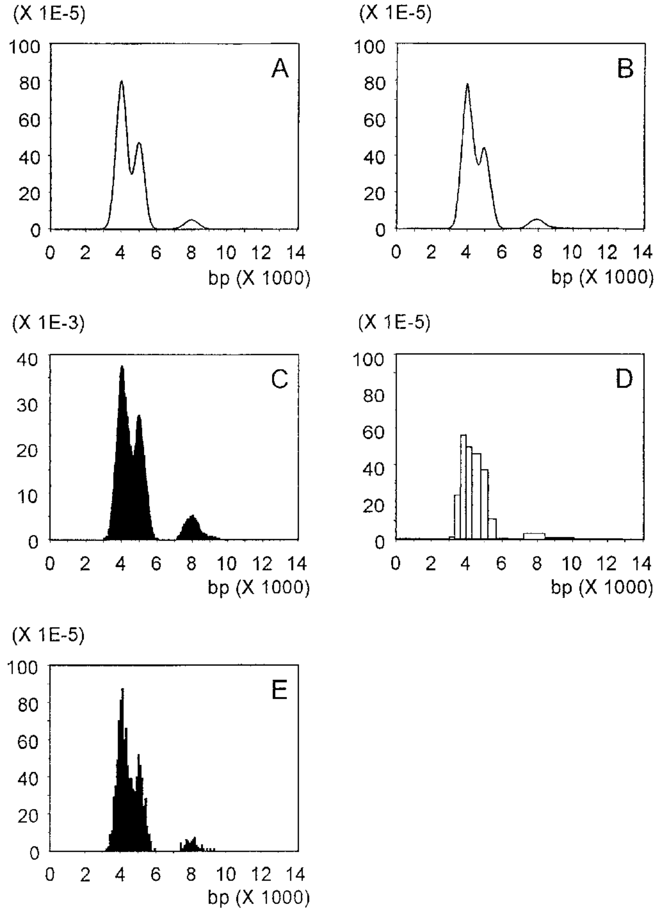

Fig. 1 Frequency distribution estimation of a 1000 one-sample simulation using different methods. A: true distribution. B: semiparametric method. C: sliding window. D: fixed bin. E: histogram (100 bp interval). Note that in plots C, D and E the frequency of each interval is equal to its bar area

the y_i values have to be grouped according to the different x values. If the true fragment lengths x_1, x_2, ..., x_n are known for a sample of given observed lengths y_1, y_2, ..., y_n, another way to estimate σ^2 is

$$\hat{\sigma}^2 = \ln(1 + S^2)$$

where S^2 is the empirical variance of the values $y_i/\hat{m}(x_i)$ and $\hat{m}$ is a non parametric regression estimator of Y given X. The correlation coefficient can be estimated using similar empirical methods as those presented above for σ^2.

Discussion

Frequency estimation

To asses the practical behaviour of the proposed method, a one-sample simulation was performed. This theoretical model is a mixture of normal densities which results in a similar shape to some of the distributions found in practice (Fig. 1). Frequency estimation of this sample using the sliding window method (Gill et al. 1990) distorts the true frequency values. This distortion is more pronounced in the high molecular weight ranges, however, each frequency value is overestimated by at least two degrees of magnitude. Something similar occurs when other bin-based frequency estimation methods, such as fixed bin (Budowle et al. 1991) are used. Thus, using these methods, conservativeness is not uniform along the whole size range and match probabilities will be clearly underestimated, even more so in systems with peaks in high molecular weight areas. Indeed, the need to account for measurement error will always induce such an effect. To minimize this kind of problem, the use of enzyme/probe combinations with small sized alleles and more appropriate frequency estimation methods must be considered.

Our approach to allelic frequency estimation is based on a more sophisticated "window method": the kernel estimation of density functions (Cao et al. 1994). The variance of the experimental measurement error is homogenised by applying a logarithmic transformation to the data (which is appropriate in this situation). Then the experimental error, in the logarithmic scale, may be assumed to be normally distributed and independent of the true fragment length. In practice the variance of this error is estimated with a preliminary sample, in which the true fragment lengths are known. The kernel method is applied to the observed fragment lengths, and the semiparametric estimator is found as a deconvolution of it and the error distribution (Valverde et al. 1993). The density functions obtained using this approach are very close to the real distributions, as can be seen in Fig. 1.

The above mentioned approaches differ completely in the way they handle the increasing error variance. These methods discretize the distribution by counting the data falling into intervals of increasing sizes (as the fragment length increases) around some predetermined fragment lengths. This leads to a loss of efficiency and poor estimation of frequencies, owing to the two transformations made on the variable. Firstly, as a result of the impossibility of determining the true fragment size, a discrete variable is translated into a continuous variable and the continuous distribution of fragment lengths is then forced into intervals to allow determination of the "bin frequencies". As explained, current technical limitations in the analytical procedure prevent an exact determination of fragment lengths, and this makes the first transformation unavoidable. However, the second transformation is not necessary, given that a continuous distribution can be treated as such, without discretization, and this leads to a finer adjustment to the real distribution. Our approach takes into account this consideration, and does not establish any artificial allelic class. Instead, the match probability is considered for each case independently.

Match probability

The density functions determined by this procedure are used to obtain the probability of a match between two bands. The calculations are carried out using a reformulation of the Bayes' theorem, in terms of the conditional density functions. This formula takes into account the probability that the evidence given by the two bands occurs when the two fragments have the same origin as well

Table 1 Correlation coefficients used for each probe in our simulation experiments

Probe	$\hat{\rho}$
YNH24	0.1842
MS43a	0.2120
MS31	0.3656

Table 2 Percentages of incorrect match assignments using 2 probes. Experimental error = 0.92%

Probe combination	% Incorrect matches
YNH24/MS43a	0.073
YNH24/MS31	0.080
MS43a/MS31	0.107

Table 3 Percentages of incorrect match assignments using 2 probes. Experimental error = 0.745%

Probe combination	% Incorrect matches
YNH24/MS43a	0.034
YNH24/MS31	0.031
MS43a/MS31	0.042

as when they are assumed to be independently observed. Both probabilities are computed by adding up the contribution that every item in the database (Z_i or D_i) gives to the total chance of observing the two fragments when either a match is assumed or not [see formulas (5), (10) and (12)]. However, it is necessary to point out that it is impossible to obtain an unbiased nonparametric estimator of a density function, and because of this, the probability of a match can be overestimated in regions where the density is small, and underestimated in more dense intervals. Nevertheless, the bias is typically small, and it is balanced with the variance when the bandwidth parameter is chosen in a correct way (Hall et al. 1992) and the sample size is reasonably large. Furthermore, the bias is not only affected by the fragment size, but also takes into account the density estimated at each point. Because of this, it is advisable to extend the database size, to guarantee a high reliability in the estimation of match probabilities. The extension of the method to the two-band case requires inclusion in the calculations of the correlation coefficient between the measurement errors of each pair of bands, which has been reported elsewhere (Evett et al. 1992). This coefficient has been calculated from a data set of 229 individuals analysed in duplicate and was considered as a constant for each probe for the sake of simplicity. However, other options, such as the use of a table of correlations (Evett et al. 1992) or a function modelling the behaviour of the coefficient with respect to the average size of the bands and the distance between them can also be considered and included in the formulas. Table 1 shows the correlation coefficient used for each probe in our experiments. The practical performance of our method was assessed for the two band case using a data set of 229 individuals analysed in duplicate in an experiment similar to that in Evett et al. (1992). Each of the 229 individuals was compared with its duplicate and, in turn, with the rest of subjects in the database. This gave a total of 229 within person comparisons and 26106 between person comparisons. The

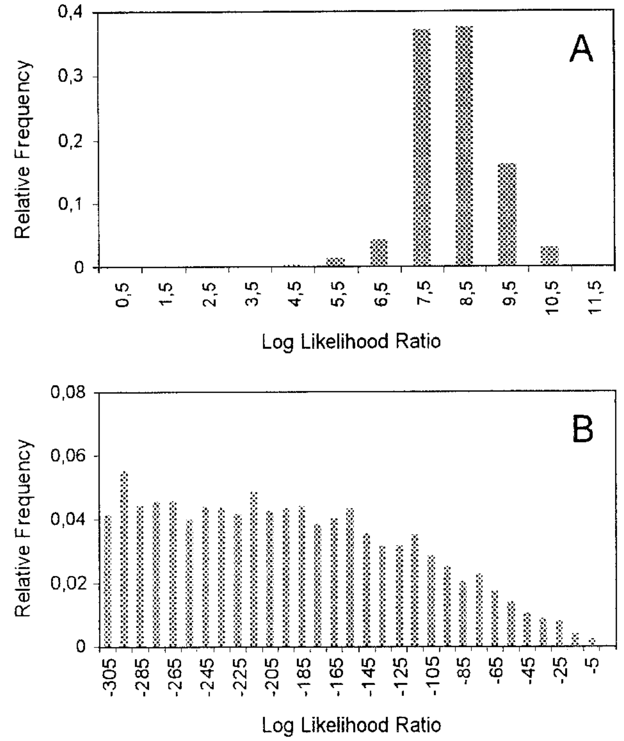

Fig. 2 (**A**) Distribution of likelihood ratio values from 229 within person comparisons using three probes (YNH24, MS43a and MS31). (**B**) distribution of likelihood ratio values from 26106 between person comparisons using the same probes. Experimental error = 0.92%

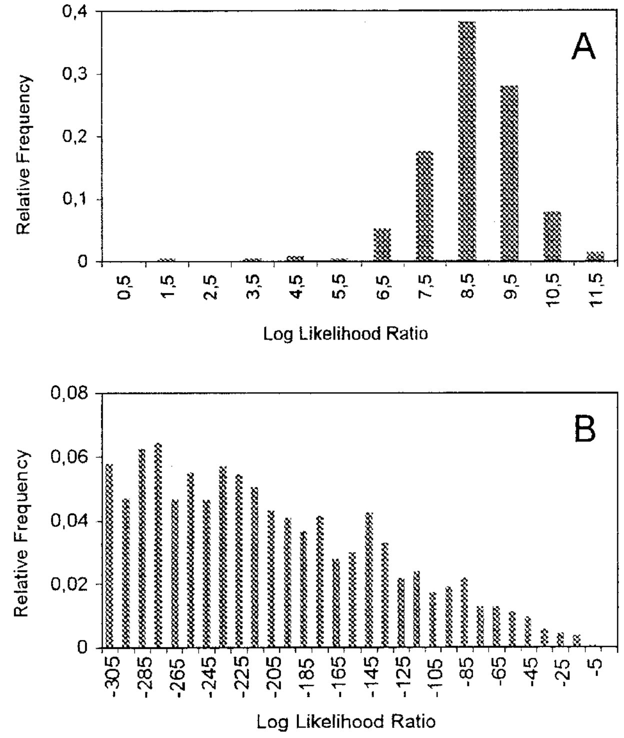

Fig. 3 (**A**) Distribution of likelihood ratio values from 229 within person comparisons using three probes (YNH24, MS43a and MS31). (**B**) distribution of likelihood ratio values from 26106 between person comparisons using the same probes. Experimental error = 0.745%

experimental error was 0.92% (Valverde et al. 1993) using fragments of the BRL 1 kb ladder measured against the ladder we use in our routine practice (Nice Ladder BRL). Given the high quality of the DNA samples and the similar electrophoretic conditions used, one-banded profiles were treated as having two fragments of the same length. For the 229 within person comparisons, we did not find any incorrect non-matches using one, two or three probes (Fig. 2 a). For between person comparisons, Table 2 shows percentages of incorrect match assignments using two probes. When we used three probes, no incorrect match assignments were found (Fig. 2 b). This experiment was repeated, this time calculating the experimental error from the allelic controls included in each gel analysed in our laboratory (see Error and correlation estimation). The value obtained was 0.745%. This change in the experimental error has a remarkable influence on the probability values (Fig. 3), however this effect does not lead to any incorrect assignment of matches or non-matches.

Conclusion

Our method offers some advantages over the currently used approaches. Firstly, it is theoretically justified, and it has been proved to be well adjusted to the reality in a great number of situations (e.g. Scott et al. 1978; Titterington et al. 1981). Secondly, it does not need any kind of binning to subdivide frequency distributions. The match probability is computed independently for each case, and the calculation is carried out with reference to the whole database. There is no subjective definition of a matching threshold, and match probability varies according to an approximately normal distribution depending on the length difference between the bands considered. Finally, this method is easy to implement in personal computers, it can be used with many kinds of databases, it can be adjusted to the specific conditions of each laboratory, and it is capable of subsequent correction as the accuracy in fragment length determination improves. It could be argued that our approach is more complicated than the currently used methods. However, it can be easily explained to personnel without specific mathematical training, and the final stages of the analysis are very similar to conventional methods. It should be borne in mind that those methods are highly conservative, making the estimate biased in favour of the defendant. Some publications have defended the advantages of conservative approaches for estimating frequencies in DNA analysis (National Research Council 1992; Monson and Budowle 1993). Extreme conservatism could be very dangerous, specially when we have samples coming from two or more suspects, one of whom could be innocent. In any case, for SLPs, conservatism must depend on the error of each laboratory. Here we propose a method to estimate this error and therefore to select an adequate degree of conservatism for each individual laboratory. Our aim is to achieve more realistic estimations of match probabilities, to take advantage of the great discrimination power of these markers and our argument has shown that this method can assist the forensic scientist in reaching this objective.

References

Baird M, Balazs Y, Giusti A, Miyasaki GL, Nicholas L, Wexler K, Kanter E, Glassberg J, Allen F, Rubinstein P, Sussman L (1986) Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its applications to the determination of paternity. Am J Hum Genet 39:489–501

Berry DA (1991) Inferences using DNA profiling in forensic identification and paternity cases. Stat Sci 6:175–205

Berry DA, Evett IW, Pinchin R (1992) Statistical inference in crime investigations using deoxyribonucleic acid profiling. Appl Stat 41:499–531

Budowle B, Giusti AM, Waye JS, Baechtel FS, Fourney RM, Adams DE, Presley LA, Deadman HA, Monson KL (1991) Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons. Am J Hum Genet 48:841–855

Cao R, Carracedo A, Valverde E (1994) Semiparametric density estimation with applications to DNA profiling. In: Bär W, Fiori A, Rossi U (eds) Advances in forensic haemogenetics 5. Springer-Verlag, Berlin Heidelberg New York, pp 444–446

Devlin B, Risch N, Roeder K (1991) Estimation of allele frequencies for VNTR loci. Am J Hum Genet 48:662–676

Devlin B, Risch N, Roeder K (1992) Forensic inference from DNA fingerprints. J Am Stat Assoc 87:337–350

Evett IW, Gill P (1991) A discussion of the robustness of methods for assessing the evidential value of DNA single locus profiles in crime investigations. Electrophoresis 12:226–230

Evett IW, Scranage JK, Pinchin R (1992) An efficient statistical procedure for interpreting DNA single locus profiling data in crime cases. J Forensic Sci Soc 32:307–326

Gill P, Sullivan K, Werrett DJ (1990) The analysis of hypervariable DNA profiles: problems associated with the objective determination of a match. Hum Genet 85:75–79

Gill P, Woodroffe S, Bär W, Brinkmann B, Carracedo A, Eriksen B, Jones S, Kloosterman AD, Ludes B, Mevåg B, Pascali VL, Schmitter H, Schneider PM, Thomson JA (1992) A report of an international collaborative experiment to demonstrate the uniformity obtainable using DNA profiling techniques. Forensic Sci Int 53:29–43

Hall P, Marron JS, Park B (1992) Smoothed cross-validation. Probab Theor Related Fields 92:1–20

Kearney JJ, Mudd JL, Hartmann JM, Kuo MC, Nelson MS, Presley LA, Stuver WC (1989) Guidelines for a quality assurance program for DNA restriction fragment length polymorphism analysis. Crime Lab Digest 16:40–54

Monson KL, Budowle B (1993) A comparison of the fixed bin method with the floating bin and direct count methods: effect of VNTR profile frequency estimation and reference population. J Forensic Sci 38:1037–1050

National Research Council (1992) DNA Technology in Forensic Science

Pascali VL, d'Aloja E, Dobosz M, Pescarmona M (1991) Estimating allele frequencies of hypervariable DNA systems. Forensic Sci Int 51:273–280

Schneider PM, Fimmers R, Woodroffe S, Werrett DJ, Bär W, Brinkmann B, Eriksen B, Jones S, Kloosterman AD, Mevåg B, Pascali VL, Rittner C, Schmitter H, Thomsom JA, Gill P (1991) Report of a European collaborative exercise comparing DNA typing results using a single locus VNTR probe. Forensic Sci Int 49:1–15

Scott DW, Gotto AM, Cole JS, Gorry GA (1978) Plasma lipids as collateral risk factors in coronary heart disease – a study of 371 males with chest pain. J Chronic Dis 31:337–345

Titterington DM, Murray GD, Murray LS, Spiegelhalter DJ, Skerre AM, Habbema JDF, Gelpke (GJ) (1981) Comparison of discrimination techniques applied to a complex data set of injured patients, J R Stat Soc A 144:145–174

Valverde E, Cabrero C, Cao R, Rodríguez-Calvo MS, Díez A, Barros F, Alemany J, Carracedo A (1993) Population genetics of three VNTR polymorphisms in two different Spanish populations. Int J Legal Med 105:251–256