To appear in: *The Journal of Forensic Sciences*

Linear Mixture Analysis:

A Mathematical Approach to Resolving Mixed DNA Samples

Mark W. Perlin*, PhD, MD, PhD
Beata Szabady, PhD
Cybergenetics, Pittsburgh, PA

Submitted for publication on: October 19, 2000
Resubmitted with revisions on: February 8, 2001
Accepted for publication: March 9, 2001

* Corresponding author contact Information:

Dr. Mark W. Perlin

Cybergenetics

160 North Craig Street, Suite 210

Pittsburgh, PA 15213 USA

412.683.3004

412.683.3005 FAX

perlin@cybgen.com

Running Header: "Linear Mixture Analysis"

Abstract

With the advent of PCR-based STR typing systems, mixed samples can be separated into their individual DNA profiles. Quantitative peak information can help in this analysis. However, despite such advances, forensic mixture analysis still remains a laborious art, with the high cost and effort often precluding timely reporting.

We introduce here a new automated approach to resolving forensic DNA mixtures. Our linear mixture analysis (LMA) is a straightforward mathematical approach that can integrate all the quantitative PCR data into a single rapid computation. LMA has application to diverse mixture problems. As demonstrated here on laboratory STR data, LMA can assess the quality and utility of its solutions. Such rapid and robust methods for computer-based analysis of DNA mixtures may help in reducing crime.

Keywords

forensic science, DNA typing, STR, DNA mixture, DNA database, criminal casework, mathematics, linear algebra, least squares, heuristic algorithm

In forensic science, DNA samples are often derived from more than one individual. In such cases, key objectives include elucidating or confirming a mixed DNA sample's component DNA profiles, and determining the mixture ratios. Current manual qualitative peak analysis of mixed DNA samples is slow, tedious, and expensive. These difficulties can generate considerable delay in the casework analysis of forensic DNA mixtures, underscored by the current USA backlog comprised of over 100,000 unanalyzed rape kits.

Under appropriate laboratory conditions, STR peak data can be quantitatively analyzed. Such quantitative approaches have spawned heuristic (1) and computer-based (2, 3) methods that can potentially resolve these complex data. These statistical computer programs typically analyze each STR locus separately, and may require human intervention when combining the locus results into a complete solution.

We have developed a quantitative analysis method that represents the mixture problem as a linear matrix equation. We call our approach "Linear Mixture Analysis," or "LMA." Unlike previous methods, the mathematical LMA model uses STR data from all the loci simultaneously for greater robustness. The linear mathematics permits rapid computer calculation, and provides a framework for statistical analysis. An associated error analysis can measure the quality of the overall solution, as well as the utility of each contributing locus.

In this paper, we introduce the linear LMA model, and then provide some illustrative examples. We describe several problem formulations, each one based on a particular subset of data available to the examiner. We then focus on laboratory data analysis results for one important mixture problem, before extending the method to other analyses. We conclude with some observations on the potential applications of LMA.

Linear Model

In the PCR amplification of a mixture, the amount of each PCR product scales in rough proportion to relative weighting of each component DNA template. This holds true whether the PCRs are done separately, or combined in a multiplex reaction. Thus, if two DNA samples A and B are in a PCR mixture with relative concentrations weighted as wA and wB ($0 \le wA \le 1$, $0 \le wB \le 1$, $wA + wB = 1$), their corresponding signal peaks after detection will generally have peak quantitations (height or area) showing roughly the same proportion. Therefore, by observing the relative peak proportions, one can estimate the DNA mixture weighting. Note that mixture weights and ratios are interchangeable, since the mixture weight $\frac{[A]}{[A]+[B]}$ is in one-to-one correspondence with the mixture ratio $\frac{[A]}{[B]}$.

To mathematically represent the linear effect of the DNA sample weights (wA, wB, wC, ...), we combine all the locus data into a single linear matrix equation:

$$\mathbf{d} = \mathbf{G} \cdot \mathbf{w},$$

Here, column vector **d** describes the mixture profile's peak quantitation data, matrix **G** represents the genotypes (column j gives the alleles for individual j), and **w** is the weight column vector that reflects the relative proportions of template DNA or PCR product. The quantitative data profile **d** is the product of genotype matrix **G** and the weight vector **w**. (A more complete data description would add an error term **e**; expected values suffice for our purposes.)

More precisely, we can write the vector/matrix equation $\mathbf{d} = \mathbf{G} \cdot \mathbf{w}$ for mixture coupling (of individuals and loci) as coupled linear equations that include the relevant data:

$$d_{ik} = \sum_j g_{ijk} w_j ,$$

where for locus i, individual j, and allele k:

- d_{ik} is the allele k proportion in the observed mixture data at locus i;
- g_{ijk} is the genotype of individual j at locus i in allele k, taking values 0 (no contribution), 1 (heterozyote or hemizygote contribution), or 2 (homozygote contribution), though with anomalous chromosomes other integer values are possible; and
- w_j is the weighting in the mixture of individual j's DNA proportion.

Illustrative Examples

This tutorial section motivates the use of vectors and matrices in modeling STR mixtures.

We first illustrate the coupling of DNA mixture weights with relative peak quantities. Suppose that there are three individuals A, B, C represented in a mixture, where 50% of the DNA is derived from individual A, 25% from individual B, and 25% from individual C. Mathematically, this corresponds to a weighting of wA=0.5, wB=0.25, and wC=0.25. Further suppose that at one locus the genotypes are:

A has allele 1 and allele 2,

B has allele 1 and allele 3, and

C has allele 2 and allele 3.

This information, and the predicted peak quantities, are laid out in Table 1.

The Table 1 information can be connected via the linear vector/matrix equation:

$$\begin{bmatrix} alleles \\ in \\ mixture \end{bmatrix} = \left[\begin{bmatrix} alleles \\ of \\ A \end{bmatrix} \begin{bmatrix} alleles \\ of \\ B \end{bmatrix} \begin{bmatrix} alleles \\ of \\ C \end{bmatrix} \right] \cdot \begin{bmatrix} wA \\ wB \\ wC \end{bmatrix}$$

Representing each allele as a position in a column vector, we have the linear relationship:

$$\begin{bmatrix} 0.75 \\ 0.75 \\ 0.50 \end{bmatrix} = \left[\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right] \cdot \begin{bmatrix} 0.50 \\ 0.25 \\ 0.25 \end{bmatrix}$$

which is the mathematical expression of Table 1. Note that the sum of alleles in each allele column vector (whether mixture or individual) is normalized to equal two, the number of alleles present.

With multiple loci, the weight vector **w** is identical across all the loci, since that is the underlying chemical mixture in the DNA template. This coupling of loci can be represented in the linear equations by extending the column vectors **d** and **G** with more allele information for additional loci.

To illustrate this coupling of DNA mixture weights across multiple loci, we add a second locus to the three individual mixture above. At locus two, suppose that the genotypes are:

A has allele 1 and allele 2,

B has allele 2 and allele 3, and

C has allele 3 and allele 4.

We can combine this vector information via the partitioned matrix equation:

$$\begin{bmatrix} locus1 \\ mixture \\ alleles \\ ---- \\ locus2 \\ mixture \\ alleles \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} locus1 \\ A's \\ alleles \\ --- \\ locus2 \\ A's \\ alleles \end{bmatrix} \begin{bmatrix} locus1 \\ B's \\ alleles \\ --- \\ locus2 \\ B's \\ alleles \end{bmatrix} \begin{bmatrix} locus1 \\ C's \\ alleles \\ --- \\ locus2 \\ C's \\ alleles \end{bmatrix} \end{bmatrix} \cdot \begin{bmatrix} wA \\ wB \\ wC \end{bmatrix}$$

Representing each allele as a position in a column vector, we have:

$$\begin{bmatrix} 0.75 \\ 0.75 \\ 0.50 \\ -- \\ 0.50 \\ 0.75 \\ 0.50 \\ 0.25 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \\ - \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ - \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ - \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \end{bmatrix} \cdot \begin{bmatrix} 0.50 \\ 0.25 \\ 0.25 \end{bmatrix}$$

Multiple loci produce more data and provide greater confidence in estimates computed from these linear equations.

Problem Formulations

Given partial information about equation $\mathbf{d} = \mathbf{G} \cdot \mathbf{w}$, other elements can be computed by solving the equation. Cases include:

- When **G** and **w** are both known, then the data profile **d** can be predicted. This is useful in search algorithms.
- When **G** and **d** are both known, then the weights **w** can be computed. This is useful in confirming a suspected mixture, and in search algorithms.
- When **d** is known, inferences can be made about **G** and **w**, depending on the prior information available (such as partial knowledge of **G**). This is useful in human identification applications.

The DNA mixture is resolved in different ways, depending on the case.

We assume throughout that the mixture profile data vector **d** has been normalized at each locus. That is, for each locus, let NumAlleles be the number of alleles found in an individual's genotype (typically NumAlleles = 2, one for each chromosome). For each allele element of the locus quantitation data, multiply by NumAlleles, and divide by the sum (over the observed alleles) of all the quantitation values for that locus. Then, the sum of the normalized locus quantitation data is NumAlleles, which totals 2 in the illustrative example above.

Resolving DNA mixtures using LMA entails (a) obtaining DNA profile data that include a mixed sample, (b) representing the data in a linear equation, (c) deriving a solution from the linear equation, and (d) resolving the DNA mixture from the solution. The LMA approach is illustrated in the following problem formulations.

Determining mixture weights

First consider the case where all the genotypes **G** and the mixture data **d** are known, and the mixture weights **w** need to be determined. This problem is resolved by solving the linear equations $\mathbf{d} = \mathbf{G} \cdot \mathbf{w}$ for **w** using a least squares matrix division method. One standard method is linear regression (4), which is often implemented using singular value decomposition (SVD) (5). In the MATLAB programming language, **w** can be estimated as:

w = G\d

using the built-in matrix division operation "\". With full rank matrices, matrix multiplication via the normal equations computes the weights as:

$$\mathbf{w} = \left(\mathbf{G}^{\mathbf{T}} \cdot \mathbf{G}\right)^{-1} \cdot \mathbf{G}^{\mathbf{T}} \cdot \mathbf{d}.$$

Others have computed mixture weights by minimizing parameters at single loci (3). From the LMA perspective, this pioneering work essentially minimizes at a single locus the sum of squares deviation $\left\|\mathbf{d} - \mathbf{G} \cdot \mathbf{w}\right\|^2$ over **w** for each feasible integer-valued

genotype matrix **G**. LMA improves on such earlier search methods by providing a mathematical basis that can use the data from *all the loci simultaneously* in a rapid numerically computed global minimization. Moreover, LMA permits the genotype matrix entries to assume any possible value, and not just integers.

Analogous mixture problems occur in other fields, and are similarly modeled using linear matrix equations. In chemometrics, the approach is termed “multivariate calibration” (MC) (6). These MC methods are quite different from computing genotypes (and mixture weights) from the data. For example, MC finds real-valued solutions but genotypes are whole numbers; calibration exploits signal continuity whereas locus patterns contribute combinatorially; and MC methods rely on multiple samplings whereas (with limited forensic samples) mixture data arise from a single multiplex PCR experiment. Therefore, our methods must be tailored to the needs of the STR mixture data, as described next.

Determining genotype profiles

Consider now the case of two individuals A and B where one of the two genotypes (say, A) is known, the mixture weights **w** are known, and the quantitative mixture data profile **d** is available. Expand $\mathbf{d} = \mathbf{G} \cdot \mathbf{w}$ in this case as:

$$\mathbf{d} = wA \cdot \mathbf{gA} + wB \cdot \mathbf{gB} \ ,$$

where **gA** and **gB** are the genotype column vectors of individuals A and B, and wA and wB = (1-wA) are their mixture weights. Then, to resolve the genotype, we can algebraically rewrite this equation as:

$$\mathbf{gB} = (\mathbf{d} - wA \cdot \mathbf{gA})/wB$$

or, equivalently, as:

$$\mathbf{gB} = (\mathbf{d} - wA \cdot \mathbf{gA})/(1 - wA)$$

and then solve for **gB** by vector arithmetic. The computed **gB** is the normalized difference of the mixture profile minus a fraction of A's genotype. The accuracy of the solution increases with the number of loci used, and the quality of the quantitative data. Typically, however, the mixture weights **w** are not known.

Consider now the critical case of making inferences about the genotype matrix **G** starting from a mixture data profile **d**. This case has practical applications for forensic science. In one typical scenario, a stain from a crime scene may contain a DNA mixture from the victim and an unknown individual, the victim's DNA is available, and the investigator would like to connect the unknown individual's DNA profile with a candidate perpetrator. This scenario typically occurs in rape cases. The perpetrator may be a specific suspect, or the investigator may wish to check the unknown individual's DNA profile against a DNA database of possible candidates. If the mixture weight wA were known, then the genotype **gB** could be computed immediately from the vector difference operation of the preceding paragraph.

Heuristic Search Algorithm: Mixture Deconvolution

Since wA is not known, one workable approach is to search for the best weight w in the [0,1] interval that satisfies additional constraints on the problem. By setting wA equal to this best w, we can compute the genotype **g**(wA) as a function of this optimized wA value, and derive **gB** = **g**(wA). A suitable constraint is the prior knowledge of the form that possible solution genotype vectors **g** can take. It is known that solutions must have a valid genotype subvector at each locus (e.g., having alleles taking on values 0, 1 or 2, and summing to 2). One may also consider null alleles, corresponding to failed PCR amplifications. This knowledge can be translated into a heuristic function of **g**(w) which evaluates each candidate genotype solution **g** against this criterion. The result of this "mixture deconvolution" algorithm is a computed genotype **gB** and the mixture weights **w**.

The heuristic we apply is a function of the unknown weight w, the observed data profile **d**, and the known genotype **gA**. Since **d** and **gA** are fixed for any given problem, in this case the function depends only on the optimization variable w. For any given w in (0,1), compute the vector:

$$\mathbf{g}(w) = (\mathbf{d} - w \cdot \mathbf{gA})/(1-w).$$

Then, at each locus, compute and record the deviation $dev_{locus}(\mathbf{g}(w))$.

The dev_{locus} function at one locus is defined as:

- Assume the genotype comprises one allele. Compute the deviation by finding the index of the largest peak, and forming a vector **oneallele** that has the value 2 at this index and is 0 elsewhere. Let dev1 be the sum of squares difference between **g**(w) and **oneallele**.
- Assume the genotype comprises two alleles. Compute the deviation by finding the index of the two largest peaks, and forming a vector **twoallele** that has the value 1 at each of these two indices and is 0 elsewhere. Let dev2 be the sum of squares difference between **g**(w) and **twoallele**.
- Return the the lesser of the two deviations as *minimum*(dev1, dev2).

To compute $dev(\mathbf{g}(w))$, we sum the component $dev_{\text{locus}}(\mathbf{g}(w))$ at each locus. That is, the heuristic function is the scalar value

$$dev(\mathbf{g}(w)) = \sum_{\text{loci}} dev_{\text{locus}}(\mathbf{g}(w)) \quad .$$

We can appropriately optimize (e.g., minimize, or detect local minimum peaks for) this function over w in [0,1] to find wA, and estimate **gB** from the computed **g**(wA). If desired, the summation terms can be normalized to reflect alternative weightings of the loci or alleles, e.g., based on variance. One useful reweighting, $(1\text{-}w)^2 \cdot dev(\mathbf{g}(w))$, is derived from the data error. Other heuristic functions can be used that reflect reasonable constraints on the genotype vectors (3).

To assess the quality of the computed STR profile, we can use information from the heuristic search. Rule checking can identify potentially anomalous allele calls, particularly when peak quantities or sizes do not conform to expectations (7). Quality

measures can be computed on the genotypes, which may suggest problematic calls even when no rule has fired. A most useful quality score in our mixture analysis is the deviation *dev*(**gB**) of the computed genotype. Low deviations indicate a good result, whereas high scores suggest a poor result. It may be helpful to partition the deviations by locus, using the locus deviation function dev_{locus}(**gB**). When a locus has an unusually high deviation, it can be removed from the profile, and the resulting partial profile then used for human identity matching.

Data Results

We analyzed two anonymous human DNA samples (A and B) both individually and in different mixture proportions (1:9, 3:7, 5:5, 7:3, 9:1). We PCR amplified the samples on a PCT-100 thermocycler (MJ Research, Waltham, MA) using the ten STR locus SGMplus multi-mix panel (PE BioSystems, Foster City, CA). We then size separated the fluorescently labeled PCR products with internal size standards on an ABI/310 Genetic Analyzer capillary electrophoresis instrument (PE Biosystems). Our manual GeneScan analysis included comparison with allelic ladder runs for allelic size designation, and recording of the peak heights and areas.

Our mixture analysis used the mixed DNA profile data **d**, along with the reference profile genotype **gA**. We implemented the LMA heuristic search algorithm in MATLAB (The MathWorks, Natick, MA), and analyzed the data on a Macintosh PowerBook G3 (Apple Computer, Cupertino, CA). We applied the automated heuristic algorithm to each data

case, with the program searching for local minima to compute the mixture weight w and the unknown genotype profile **gB**. The computation time for each problem was less than 0.1 second. We recorded the total deviation *dev*(**gB**), along with the deviations at each locus and allele. We also compared our computed profile with the actual profile for individual B. (While known in advance for assessment purposes, neither the mixture weight w nor B's profile were used in the calculations.)

For each mixture proportion, for both height and area, the computed mixture weights and sum of squares deviations (between the estimated and actual genotypes) are shown (Table 2). There is good agreement between the estimated weights and the known proportions. When the unknown proportion (B) becomes small (e.g., at 10% in the 9:1 case), the low relative signal can lead to less certain results, as measured by the deviation.

We examine the data analysis for the 3:7 (30% A to 70% B) case in more detail. Using peak area data, the search (Figure 1) for weight w by minimization of *dev*(**g**(w)) gave a weighting of 29.18%; this value is close to the true 30% DNA mixture. The total sum of squares deviation *dev*(**g**(w)) of the computed genotype from the closest (and correct) feasible solution was 0.1000. A summary diagram (Figure 2) shows the locus-by-locus profiles in separate rows for (1) the mixture data **d**, (2) the reference profile **gA**, and (3) the numerically derived unknown profile **gB**. Quality assessment of the computed profile **gB** shows uniform peak heights that are consistent with a correct genotype.

Data and results are tabulated for each locus (Table 3). "Mixture" is the normalized peak quantity data from the mixed sample. "Geno A" is the known genotype of individual A. "Profile" is the numerical estimate of B's genotype computed by the mixture deconvolution heuristic search algorithm. "Geno B" is the resulting integer genotype (and, in this case, identical to B's actual genotype) obtained by rounding Profile to the nearest integer. "Sq Devs" are the sum of squares deviations of the Profile from Geno B. Examination of the squared deviation components for each allele revealed no major outliers. The largest within-locus sum of squares deviation was the nominal value 0.0272 at locus D2S1338; this locus has relatively long DNA fragment lengths, which is consistent with finding larger variation.

We applied our automation methods to data from other laboratories, obtaining accurate results. For example, we reanalyzed the original six locus STR data (provided by Dr. Peter Gill) underlying the quantitative analysis of mixture sample MT/NO in (3). Taking individual MT as the known reference profile, for each approximate mixing ratio (1:10, 1:5, 1:2, 1:1, 2:1, 5:1, 10:1), we derived exact mixture weights and estimated individual NO's genotype. The respective computed weights (10.02%, 13.83%, 27.87%, 41.89%, 58.43%, 77.25%, 86.66%) are in close agreement with the four allele locus weights that the authors had estimated (Table 6 for 5ng DNA in (3)).

To assess three person mixture deconvolution, we analyzed three anonymous human DNA samples (A, B and C) in different mixture proportions. We generated SGMplus STR data on these mixed samples using the protocols described above, and recorded

the peak measurements (height, area, size, designation). The (very approximate) 4:1:1 DNA combination experiment generated 44 alleles across the 10 STR loci. Specifying all three known genotypes, we estimated the true mixture weights using LMA, and determined that the weights were wA = 70.56%, wB = 11.43%, and wC = 18.01%.

We then performed mixture deconvolution on the three person mixture data **d**. We used genotypes **gA** and **gB** as known references, but left genotype **gC** (and the mixture weights) as unknown parameters. Mixture deconvolution explored the 44 dimensional allele measurement space by searching for the best two dimensional (wA, wB) weighting pair, and estimated the weights as wA = 70%, wB = 11%, and wC = 19%. This weighting result is in good agreement with the "all knowns" calculation, and suggests that LMA may be useful on data containing more than two contributors.

Other Analyses

Stutter peaks are often a concern in mixture analysis. One clean analysis method is to mathematically remove the stutter artifact from the quantitative signal using stutter deconvolution methods (8) prior to the mixture analysis. Other forensic scientists have used Bayesian approaches to account for stutter (9). However, direct stutter removal from the data signal can be highly robust, since it is working directly at the level of the stutter artifact, prior to any mixture computation.

In the reporting of mixture analysis, some courts are interested in likelihood ratio formulations. Bayesian methods have been developed to provide such likelihoods (2). However, these reporting methods require a reasonable estimate of the conditional probability *Prob*(**d** | **G**, **w**) of the observed mixture data, given an hypothesized genotype and mixture weight. Our LMA can help supply such estimates, since the linear algebra provides a geometric framework for measuring the Euclidean distance $\|\mathbf{d} - \mathbf{G} \cdot \mathbf{w}\|$ or its square (which is the sum of squares deviation) between an observed mixture profile **d**, and a profile estimate $\mathbf{G} \cdot \mathbf{w}$. One can compute the requisite conditional probabilities by correlating these distances with genotype correctness on empirical mixture data, or by using linear statistical models (4).

The LMA model is also useful for resolving mixtures when there are no reference profiles available. In this situation, the computer considers all feasible genotype pairs $\mathbf{H}_i$ at a locus subset, and then determines the weight **w** (and genotype pair $\mathbf{H}_i$) that provides the best possible fit to the data by minimizing $\|\mathbf{d} - \mathbf{H}_i \cdot \mathbf{w}\|$. Progressing in this way from the most informative loci (e.g., those with the most alleles in their data), the computer can ascertain the full genotype profiles of both individuals.

Once large DNA databases have been constructed, there will be an alternative LMA approach to resolving mixtures without reference profiles. With such a database, one could iterate through an entire convicted offender database, testing each offender profile in turn as a possible **gA**, and then compute **gB**. If a **gB** profile of sufficient quality were derived, this could implicate both individuals (having DNA profiles **gA** and

gB) as the contributors to the mixture. In this way, the mathematical LMA method, coupled with knowledge of criminal profiles from a database, would effectively search for the individual component profiles.

Conclusion

STR profiling of human DNA is proving to be an effective mechanism for reducing crime. However, DNA mixtures have become a key bottleneck impeding the rapid resolution of cases. Interestingly, the underlying PCR amplification step, as well as the fluorescent detection step, show a quantitatively linear response in the presence of DNA mixtures. This suggests the use of linear algebraic models to explain mixture problems and compute their solutions.

We have introduced linear mixture analysis (LMA), a straightforward mathematical method for resolving DNA mixture problems. The underlying linear mathematics permits rapid and robust solutions on real quantitative data. LMA uses all the data in a single combined computation, which contributes to its robustness and accuracy – the method is unlikely to find an incorrect solution. Moreover, heuristic algorithms based on LMA have built-in approaches for determining error, identifying suspect loci, and establishing confidence.

Under reasonable PCR conditions, multiplex STR data appear to demonstrate linear additivity, once DNA concentrations have been renormalized within each locus. Our

linear analysis of each experiment produced a mixture weight having only small deviations across the loci. Based on 6-plex STR data, others have conjectured that DNA mixtures amplify linearly (3); our 10-plex data and linear analysis concur. Ongoing experimentation will assess the linearity of newer multilocus multiplex panels.

LMA may see broad application in rape cases. Applying the LMA-based mixture deconvolution method to the mixed DNA crime profile, together with a reference profile from the victim, may enable rapid and automated determination of the perpetrator's DNA profile. When coupled with the anticipated large offender DNA databases, perpetrator identities could be revealed in a matter of hours. This technological "DNA surveillance" capability may have a deterrent effect on some subpopulation of potential offenders.

Acknowledgements

Conversations with Drs. Peter Gill and Ian Evett of the British Forensic Science Service (FSS) were most helpful. Discussions with Dr. Cecelia Crouse of the Palm Beach County Sheriff's Office in Florida helped focus the method on useful mixture problem areas.

References

1. Clayton TM, Whitaker JP, Sparkes R, Gill P. Analysis and interpretation of mixed forensic stains using DNA STR profiling. Forensic Sci. Int. 1998;91:55-70.
2. Evett IW, Gill P, Lambert JA. Taking account of peak areas when interpreting mixed DNA profiles. J. Forensic Sci. 1998;43(1):62-69.
3. Gill P, Sparkes R, Pinchin R, Clayton TM, Whitaker JP, Buckleton J. Interpreting simple STR mixtures using allele peak area. Forensic Sci. Int. 1998;91:41-53.
4. Seber GAF. Linear Regression Analysis. New York: John Wiley & Sons, 1977
5. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in C: The Art of Scientific Computing. Second ed. Cambridge: Cambridge University Press, 1992
6. Martens H, Naes T. Multivariate Calibration. New York: John Wiley & Sons, 1992
7. Perlin M. Computer automation of STR scoring for forensic databases. In: First International Conference on Forensic Human Identification in The Millennium; 1999 Oct 25-27; London, UK: The Forensic Science Service; 1999.

8. Perlin MW, Lancia G, Ng S-K. Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. Am. J. Hum. Genet. 1995;57(5):1199-1210.

9. Gill P, Sparkes R, Buckleton JS. Interpretation of simple mixtures when artefacts such as stutters are present - with special reference to multiplex STRs used by the Forensic Science Service. Forensic Sci. Int. 1998;95:213-224.

TABLE LEGENDS

Table 1. The relative data quantity is calculated for each allele at the locus as shown. For example, allele 1's relative data value of 0.75 is calculated from (a) the genotype values of <1, 1, 0> (i.e., the allele is <present, present, absent>) at allele 1 for individuals A, B, and C, and (b) the individuals' DNA mixture weight contributions of <0.50, 0.25, 0.25>. The computation is performed by computing the inner product of these two vectors as (1x0.50) + (1x0.25) + (0x0.25) = 0.75.

Table 2. The DNA mixtures were combined in the proportions shown, and the DNA profiles were generated. For each proportion, the quantitative peak heights and areas were measured. From these data, the mixture weight and sum of squares deviation from the correct answer were computed.

Table 3. The detailed quantitation results for a 3:7 mixture of two DNA samples processed with the SGMplus panel. The computed profile (Profile) is a reasonable numerical estimate of the actual genotype (Geno B), as indicated by the small sum of squares deviations (Sq Dev) listed. Deviations are listed for alleles, loci (subtotals, shown in italics), and the sample (grand total, shown in bold). Please refer to the text for a detailed description of the other quantities shown.

Table 1.

			Individuals		
			A	B	C
			Genotypes G		
Alleles	Data d		1,2	1,3	2,3
1	0.75		1	1	0
2	0.75	=	1	0	1
3	0.50		0	1	1
			0.50	0.25	0.25
			wA	wB	wC
			Weights w		

Table 2.

Known Proportions		Derived Weight and Profile Deviations (Height)		(Area)	
A:B	%	Weight	Sq Dev	Weight	Sq Dev
1:9	10%	10.9%	0.0900	9.5%	0.1142
3:7	30%	29.3%	0.1112	29.2%	0.1000
5:5	50%	48.0%	0.3222	48.4%	0.2493
7:3	70%	69.2%	0.5303	69.5%	0.4111
9:1	90%	84.6%	4.3907	86.0%	6.3853

Table 3.

Locus-Allele	Mixture	Geno A	Profile	Geno B	Sq Dev
D3S1358-14	1.0365	1	1.0516	1	0.0027
D3S1358-15	0.9635	1	0.9484	1	0.0027
					0.0053
vWA-17	1.4755	0	2.0835	2	0.0070
vWA-18	0.5245	2	-0.0835	0	0.0070
					0.0140
D16S539-11	1.4452	0	2.0406	2	0.0017
D16S539-13	0.2889	1	-0.0041	0	0.0000
D16S539-14	0.2660	1	-0.0365	0	0.0013
					0.0030
D2S1338-16	0.3190	1	0.0384	0	0.0015
D2S1338-18	0.6339	0	0.8951	1	0.0110
D2S1338-20	0.3713	1	0.1122	0	0.0126
D2S1338-21	0.6758	0	0.9543	1	0.0021
					0.0272
D8S1179-9	0.7279	0	1.0278	1	0.0008
D8S1179-12	0.2749	1	-0.0239	0	0.0006
D8S1179-13	0.6813	0	0.9620	1	0.0014
D8S1179-14	0.3160	1	0.0341	0	0.0012
					0.0040
D21S11-27	0.2787	1	-0.0185	0	0.0003
D21S11-29	0.7876	0	1.1121	1	0.0126
D21S11-30	0.9337	1	0.9064	1	0.0088
					0.0217

D18S51-12	0.3443	1	0.0741	0	0.0055
D18S51-13	0.6952	0	0.9816	1	0.0003
D18S51-14	0.6755	0	0.9538	1	0.0021
D18S51-17	0.2850	1	-0.0096	0	0.0001
					0.0081
D19S433-12.2	0.6991	0	0.9872	1	0.0002
D19S433-14	0.6060	2	0.0316	0	0.0010
D19S433-15	0.6949	0	0.9813	1	0.0004
					0.0015
THO1-6	0.3178	1	0.0366	0	0.0013
THO1-7	1.0074	1	1.0104	1	0.0001
THO1-9	0.6749	0	0.9530	1	0.0022
					0.0037
FGA-19	1.0580	1	1.0819	1	0.0067
FGA-24	0.2830	1	-0.0124	0	0.0002
FGA-25.2	0.6589	0	0.9304	1	0.0048
					0.0140
					0.1000

FIGURE LEGENDS

Figure 1. Five curves are shown, each plotting the squared deviation against the mixture weight w. From left to right, these curves correspond to the heuristic functions of the 1:9 (plus), 3:7 (solid), 5:5 (cross), 7:3 (dash), and 9:1 (dot) mixture ratios. The minima of these curves are located near 10%, 30%, 50%, 70%, and 90%, respectively, demonstrating that mixture deconvolution correctly infers the true mixture weight. The shape of the 9:1 (dot) curve reflects the trajectory through allele space as the weight changes from 0 to 1.

Figure 2. The quantitative data **d** of the 3:7 mixture experiment is shown at every SGMplus locus (first row). Also shown is the known reference profile of individual **a** (second row). Using mixture deconvolution, the computer estimates the unknown genotype **b** (third row) and the mixture weight w. Note that the estimated genotype is the same as the actual genotype **b** (fourth row).

Figure 1.

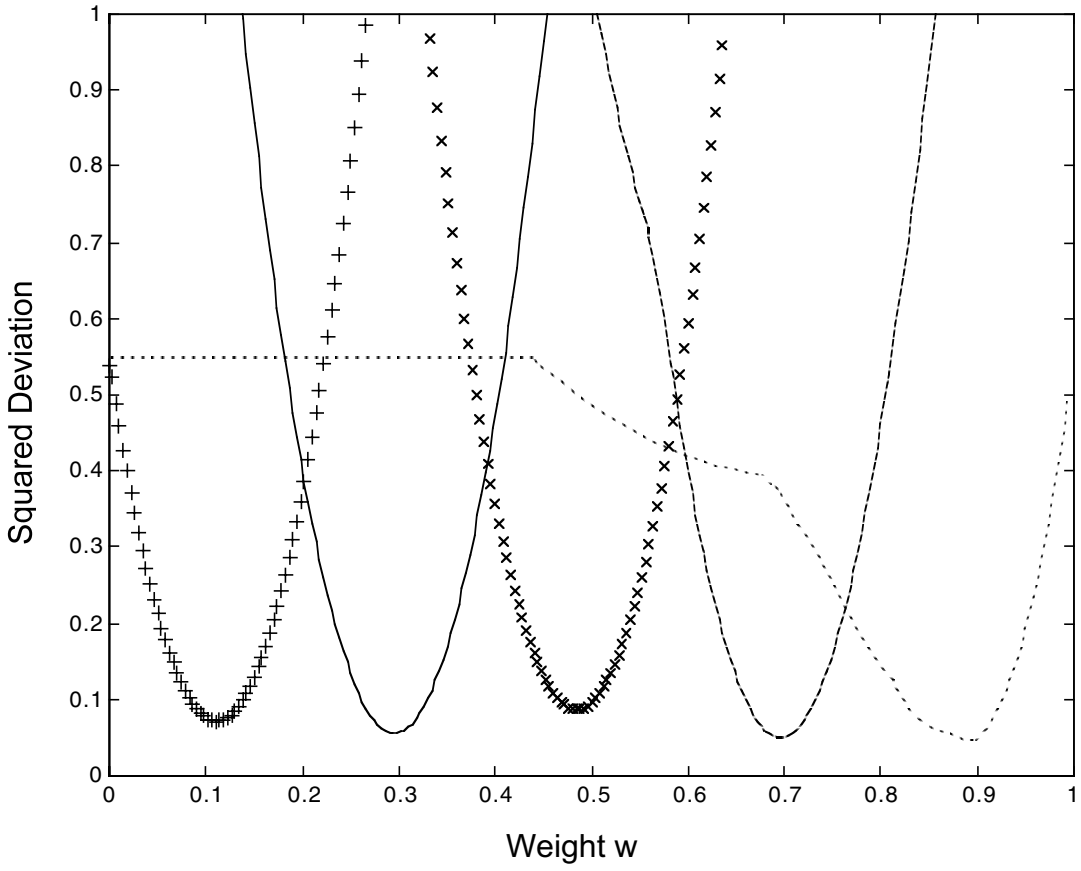
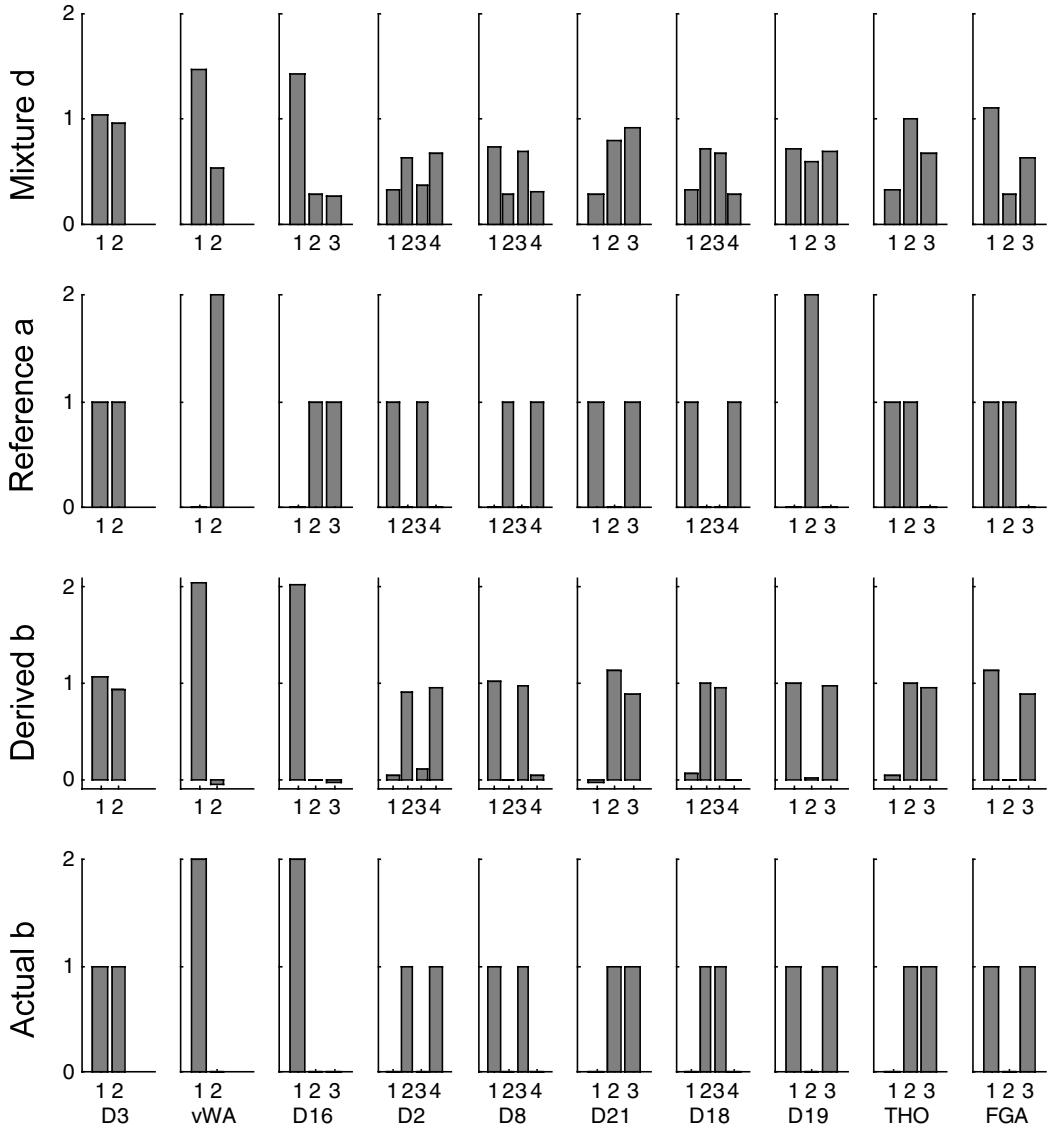


Figure 2.


Mixture d
Reference a
Derived b
Actual b
2
1
0
1 2
1 2 3
1 2 3 4
D3
vWA
D16
D2
D8
D21
D18
D19
THO
FGA