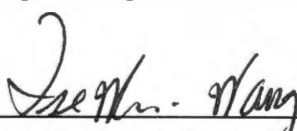
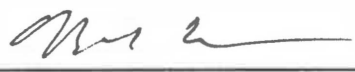
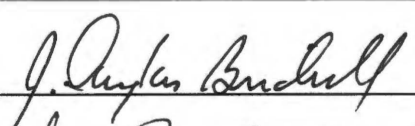
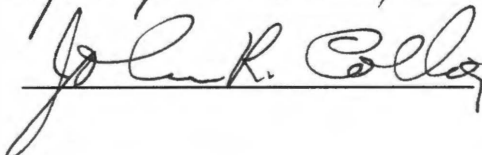


To the Graduate Council:

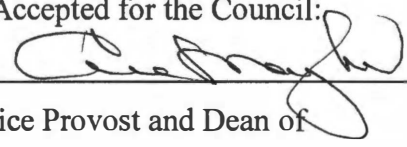
I am submitting herewith a thesis written by Ning Xue entitled "Least-Square Deconvolution (LSD): a New Method To Resolve STR-DNA Mixtures." I have examined the final paper copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Chemical Engineering.


Tsewei Wang, Major Professor

We have read this thesis and
recommend its acceptance:

Accepted for the Council:


Vice Provost and Dean of
Graduate Studies

**LEAST-SQUARE DECONVOLUTION (LSD): A NEW METHOD TO
RESOLVE STR-DNA MIXTURES**

A Thesis
Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Ning Xue
December 2002

THESIS
2002
.x84

Copyright © 2002 by Ning Xue

All rights reserved.

ACKNOWLEDGEMENTS

I wish to express my appreciation to all those people who helped me in completing the Master of Science degree in Chemical Engineering. First and foremost, I would like to thank my advisor, Dr. Tsewei Wang for mentoring me during my program and thesis writing in the last two years. Her professionalism, patience, and encouragement allowed me to grow immensely as a student and researcher. I also want to thank the other members of my thesis committee, Dr. J.R. Collier, Dr. J.D. Birdwell and Dr. Mark Rader for their continued guidance and support during the research and the preparation of this thesis. I thank Melinda Coker, a good partner in the course of English 462, and Dr. Russel Hirst in English Department, who read my thesis and gave me some good advice. Finally, I wish to thank all the members in the Laboratory for Information Technologies in the University of Tennessee for their advice and help.

This work has been supported by the Federal Bureau of Investigation under contract J-FBI-98-083. The DNA mixture data was provided by the Promega Corporation, the Acadiana Criminalistics Laboratory in New Iberia, LA, and Texas Department of Public Safety in Corpus Christi. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. government.

ABSTRACT

Mixed DNA samples are very common in crime scenes, especially in rape and murder cases. This kind of DNA evidence essentially includes the perpetrator's DNA sample superimposed on that of another, usually the victim. As a result, the genotype of each individual contributor cannot be obtained directly from the output of a genetic analyzer. The availability of allele peak area information from mixed STR/DNA samples makes it possible to use the least-square algorithm to find the most fitting genotype profile for each contributor. A new method, the Least-Square Deconvolution (LSD) method, is presented in this thesis to resolve DNA mixture samples. The estimation is based on finding the least-square fit of mass ratio coefficients that come 'closest' to the allele peak area ratios of each locus, given by the PCR products after amplification. The top-ranked genotype combination from each locus with consistent mass ratios can be pulled out to form a composite DNA profile at all the available loci. This approach is applicable for mixture loci containing 1-4 alleles per locus. Both real DNA peak area data from forensic samples, provided by different crime laboratories, as well as simulated peak area sample data were used to test this new method. The LSD results consistently show that the most fitting genotype combinations selected by LSD almost always correspond to the true genotypes of the victim and the suspect, except in some known circumstances. The assumed perpetrator's genotype profile yielded by LSD can then be used to check against that of a suspect in custody, or be used to search for a matching profile in a DNA database. The limitations of LSD and the comparison between LSD and other methods to resolve DNA mixtures are also presented in this thesis.

TABLE OF CONTENTS

CHAPTER	PAGE
1. INTRODUCTION.....	1
1.1. Motivation for Work Undertaken.....	2
1.2. Anticipated Importance of LSD in Forensic Work.....	4
1.3. Problem Statement.....	5
2. BACKGROUND.....	7
2.1. Genetic and Technical Basis of DNA Typing.....	7
2.1.1. Fundamentals of Genetics.....	7
2.1.2. Introduction of DNA Identification.....	8
2.1.2.1. Gene Amplification by the Polymerase Chain Reaction (PCR).....	9
2.1.2.2. Short Tandem Repeat (STR) Markers.....	12
2.1.2.3. Instrumentation for the Detection of Alleles of STR/DNA Loci.....	17
2.2. Introduction to the Problem of Resolving DNA Mixtures.....	21
2.2.1. Identification of DNA Mixtures.....	21
2.2.2. Identification of the Number of Contributors.....	24
2.2.3. Characteristics of Allele Peaks of STR/DNA Mixtures.....	25
2.3. The Linear Least-Square Solution to Data Fitting.....	26
2.3.1. Introduction to the Least-Square Approach.....	26

2.3.2.	Multivariable Least-Square Solution to Inconsistent Linear System of Equations, with More Equations Than Unknowns	29
3.	PRIOR EFFORTS TO RESOLVE MIXTURE DNA SAMPLES USING ALLELE PEAK DATA.....	34
3.1.	Estimation of the Mass Ratio in a DNA Mixture Sample.....	34
3.2.	Quantitative Approaches in Resolving Mixture DNA.....	39
3.2.1.	A First-Reported Method to Resolve DNA Mixture.....	39
3.2.2.	A Recently Published Mathematical Method to Resolve DNA Mixture.....	44
3.2.2.1.	Linear Model of LMA.....	45
3.2.2.2.	Problem Formulation and Solution Approach of LMA.....	47
4.	LEAST-SQUARE PROBLEM FORMULATION AND SOLUTION APPROACH TO RESOLVE TWO-PEOPLE STR/DNA MIXTURE SAMPLES.....	53
4.1.	Least-Square Deconvolution (LSD), an Overview.....	53
4.1.1.	Representing All Possible Genotype Combinations at Each Locus with Matrices.....	54
4.1.2.	Using the Least-Square Approach to Calculate the Best-Fit Mass Coefficients and the Residuals at Each Locus.....	59
4.1.3.	Finding the Most Likely Genotype Combination for Each Locus.....	60
4.1.4.	Reviewing the Ranking Result and Forming the Composite Profiles.....	63
4.2.	Examples of Using LSD in Resolving Real Forensic Mixture DNA...	65
4.2.1.	Interpretation of the LSD Result for a Three-Allele Locus ...	65

4.2.2.	Interpretation of the LSD Result for a Four-Allele Locus....	73
4.2.3.	Interpretation of the LSD Result for a Two-Allele Locus.....	78
5.	RESULTS AND DISCUSSION.....	85
5.1.	Sensitivity of LSD to Noise in the Allele Peak Data.....	86
5.2.	LSD Interpretation Results Using Simulated DNA Mixture Data.....	89
5.2.1.	LSD Interpretation Result Using Simulated DNA Mixture Data with a Mass Ratio of 1:5.....	91
5.2.2.	LSD Interpretation Result Using Simulated DNA Mixture Data with a Mass Ratio of 1:2.....	96
5.2.3.	LSD Interpretation Result Using Simulated DNA Mixture Data with a Mass Ratio of 1:1.....	101
5.3.	Some Limitations of LSD Algorithm.....	106
5.4.	LSD Interpretation Results of DNA Mixture Data from Literature Reports.....	109
5.5.	LSD Interpretation Results of Real DNA Mixture Samples.....	120
5.5.1.	LSD Interpretation Results of DNA Mixture Samples from the Acadiana Crime Lab.....	120
5.5.1.1.	LSD Interpretation Result of Acadiana DNA Mixture Data No. 5.....	122
5.5.1.2.	LSD Interpretation Result of Acadiana DNA Mixture Data No. 1.....	130
5.5.1.3.	LSD Interpretation Result of Acadiana DNA Mixture Data No. 2.....	137
5.5.2.	LSD Interpretation Result of DNA Mixture Samples from the Texas Department of Public Safety in Corpus Christi....	146

5.6. Comparison Between LSD and LMA.....	162
6. CONCLUSION AND FUTURE WORK.....	182
LIST OF REFERENCES.....	185
VITA.....	189

LIST OF TABLES

TABLE	PAGE
1.1 All Possible Genotype Combinations for the Two Contributors at a Three-Allele Locus as Shown in Figure 1.1 for a Mixed DNA Sample..	6
2.1 Numbers of Copies of Target DNA Segment After Different Rounds of PCR Amplification Process.....	11
2.2 Frequency Distributions of Alleles at Locus CSF1PO for Caucasians...	14
2.3 Frequency Distributions of Alleles at Locus D7S820 for Caucasians....	15
2.4 Frequency Distributions of Alleles at Locus D8S1179 for Caucasians...	16
3.1 All Allele Assignment Patterns for a Two-People DNA Mixture.....	35
3.2 Allele Designation of a Mixture DNA Sample at Six Loci, As Well As the True Genotypes of Its Two Contributors.....	41
3.3 The Expected Proportions of Allele Peak Areas for All Possible Genotype Combinations If Two Alleles Were Detected at a Locus.....	41
4.1 All Possible Genotype Combinations and Their Corresponding Matrices at a Three-Allele Locus for the Two Contributors.....	56
4.2 All Possible Genotype Combinations and Their Corresponding Matrices at a Two-Allele Locus for the Two Contributors.....	58
4.3 All Possible Genotype Combinations and Their Corresponding Matrices at a Four-Allele Locus for the Two Contributors.....	58
4.4 The Allele Peak Areas at Nine Loci and the Genotypes of Both Contributors for the DNA Mixture Sample Shown in Figure 4.3.....	67
4.5 All Possible Genotype Combinations for the Two Contributors and Their Corresponding Matrices for Locus VWA with Three alleles 16, 17, 18, for the DNA Mixture Data Shown in Table 4.4.....	72

4.6	Normalized Fitting Errors and Calculated Mass Ratios for All Possible Genotype Combination Cases for Locus VWA with Three Alleles as Shown in Table 4.5.....	74
4.7	Ranking Result for Locus VWA in the DNA Mixture Data Shown in Table 4.4, According to the Magnitudes of fitting Errors.....	76
4.8	All Possible Genotype Combinations and the Corresponding Matrices for the Two Contributors at Locus D21S11 with Four Alleles 30, 30.2, 31.2, 32.2 in the DNA Mixture as Shown in Table 4.4.....	77
4.9	Ranking Result According to the Magnitudes of Fitting Errors for Locus D21S11 in the DNA Mixture Data shown in Table 4.4.....	79
4.10	All Possible Genotype Combinations and the Corresponding Matrices for the Two Contributors at Locus D7S820 with Two Alleles 10, 11 in the DNA Mixture Data Shown in Table 4.4.....	81
4.11	Ranking Result for Locus D7S820 in the DNA Mixture Data Shown in Table 4.4 According to the Magnitudes of Fitting Errors.....	81
4.12	LSD Suggested Genotype Resolution Results at All Nine Loci for the DNA Mixture Data Shown in Table 4.4.....	82
5.1	Maximum Gaussian Noise Present in the Allele Peak Data Tolerable by LSD in Resolving DNA Mixture Samples.....	88
5.2	Simulated Allele Peak Data and Genotypes for the Two Contributors at Six Loci for Mixture DNA Samples with the Mass Ratios of 1:5, 1:2, and 1:1, Respectively.....	90
5.3	LSD Suggested Genotype Resolution Result of a Set of Simulated DNA Mixture Data Shown in Table 5.2, with a Mass Ratio of 1:5.....	92
5.4	Ranking Result at the Two-Allele Locus D8S1179 in the Simulated DNA Mixture Data Shown in Table 5.2, with a Mass Ratio of 1:5.....	95
5.5	LSD Suggested Genotype Resolution Result of a Set of Simulated DNA Mixture Data Shown in Table 5.2, with a Mass Ratio of 1:2.....	97
5.6	Ranking Result at the Two-Allele Locus FGA in the Simulated DNA Mixture Data Shown in Table 5.2, with a Mass Ratio of 1:2.....	100

5.7	LSD Suggested Genotype Resolution Result of a Set of Simulated DNA Mixture Data Shown in Table 5.2, With a Mass Ratio of 1:1.....	102
5.8	Example of a Possible Reversed Genotype Assignment for the Two Contributors at a Four-Allele Locus With Mass Ratio Close to 1:1.....	106
5.9	Two Possible Genotype Combinations at a Three-Allele Locus with the Ratio of the Allele Peak Areas of 2:1:1, and the Mass Ratio of the DNA Mixture is Close to 1:1, as Predicted from Other Loci.....	108
5.10	Three Possible Genotype Combinations at a Three-Allele Locus with the Ratio of the Allele Peak Areas of 1:1:1, and the Mass Ratio of the DNA Mixture is Close to 1:2, as Predicted from Other Loci.....	108
5.11	Two Possible Genotype Combinations at a Two-Allele Locus with the Peak Area Ratio of 1:1, and the Mass Ratio of the DNA Mixture is Also Close to 1:1, as Predicted from Other Loci.....	108
5.12	Special Interpretation Rule for a 2-Allele Locus, When Both the Mass Ratio and the Peak Area Ratio are Close to 1:2.....	109
5.13	Allele Peak Data and the Genotypes for the Two Contributors at Six Loci from a Mixed DNA Sample Presented in Reference [3].....	111
5.14	LSD Suggested Genotype Resolution Result of a Mixed DNA Sample Presented in Reference [3].....	112
5.15	Allele Peak Data and the Genotypes of the Two Contributors at Six Loci of a Mixed DNA Sample with the Mass Ratio of 1:10 Prior to PCR Amplification, as Presented in Reference [21].....	116
5.16	LSD Suggested Genotype Resolution Result of a Mixed DNA Sample Obtained from Reference [21].....	117
5.17	Allele Peak Data and the Genotypes of Both Contributors at Thirteen Loci in Acadiana DNA Mixture Data No. 5.....	123
5.18	LSD Suggested Genotype Resolution Result for Acadiana DNA Mixture Data No. 5.....	128
5.19	Detailed LSD Ranking Result at Locus D21S11 in Acadiana DNA Mixture Data No. 5, in Which the Fitting Errors for the Top Two Ranked Combinations are Comparable.....	129

5.20	Allele Peak Data and the True Genotypes for Both Contributors at Thirteen Loci in Acadiana DNA Mixture Data No. 1	131
5.21	LSD Suggested Genotype Resolution Result for Acadiana DNA Mixture Data No. 1	136
5.22	Allele Peak Data and the True Genotypes for the Two Contributors at Thirteen Loci in Acadiana DNA Mixture Data No. 2	138
5.23	LSD Suggested Genotype Resolution Result for Acadiana DNA Mixture Data No. 2	142
5.24	LSD Ranking Result at a Two-Allele Locus D7S820 in Acadiana DNA Mixture Data No.2	143
5.25	LSD Ranking Result at D8S1179, a Three-Allele Locus, in Acadiana DNA Mixture Data No.2, According to the Fitting Error Associated with Each Case	145
5.26	Allele Peak Data and the Genotypes of Both Contributors at Nine Loci in Texas DNA Mixture Data No. 3	150
5.27	LSD Suggested Genotype Resolution Result for Texas DNA Mixture Data No. 3	154
5.28	Allele Peak Data and the True Genotypes of Both Contributors at Nine loci in Texas DNA Mixture Data No. 4	157
5.29	LSD Suggested Genotype Resolution Result for Texas DNA Mixture Data No. 4	161
5.30	Allele Peak Data and the Genotypes of the Two Contributors at Ten Loci of a Mixed DNA Sample with the Mass Ratio of 3:7 Prior to PCR, as Presented in Reference [18]	166
5.31	Validation of Our Implementation of LMA [18]. Exactly the Same Results are Generated	170
5.32	Allele Peak Data and the Genotypes of the Two Contributors at Six Loci of a Mixed DNA Sample with the Mass Ratio of 1:10 Prior to PCR Amplification, as Presented in Reference [21]	172

5.33	LSD Suggested Genotype Resolution Result of a Mixed DNA Sample Presented in Reference [21].....	173
5.34	LMA Suggested Genotype Resolution Result of the DNA Mixture Data Given in Table 5.32.....	174
5.35	Allele Peak Data and the True Genotypes of Both Contributors of a DNA Mixture Sample, Provided by the Promega Corporation, With the Known Mass Ratio of 9:1, Prior to Amplification.....	177
5.36	LSD Suggested Genotype Resolution Result of the DNA Mixture Data Given in Table 5.35.....	178
5.37	LMA Suggested Genotype Resolution Result of the DNA Mixture Data given in Table 5.35.....	180
5.38	Summary of Comparison Between LMA and LSD in Resolving 2-People DNA Mixtures.....	181

LIST OF FIGURES

FIGURE	PAGE
1.1 Simulated Superimposed Allele Peaks at One Locus for a Mixed DNA Profile with the Contributors' Genotypes of [3, 15] and [7, 15], Mixed at a Mass Ratio of 1:0.4 Respectively.....	6
2.1 A Simulation of the PCR Process Used to Amplify Each DNA Segment of Interest 20-35 times.....	10
2.2 Bar Chart of Frequency Distributions of Alleles at Locus CSF1PO for Caucasians.....	14
2.3 Bar Chart of Frequency Distributions of Alleles at Locus D7S820 for Caucasians.....	15
2.4 Bar Chart of Frequency Distributions of Alleles at Locus D8S1179 for Caucasians.....	16
2.5 Typical Processing Steps of Using ABI 310 Genetic Analyzer to Identify Alleles Present at Each Locus.....	19
2.6 Typical Analysis Output of a DNA Profile Using the ABI 310 Genetic Analyzer.....	20
2.7 Illustration of Allele Peaks.....	22
2.8 Projection of b Vector (4, 6) onto A Vector (9, 5).....	28
2.9 Vector b is on the Column Space of A (3×2 Matrix).....	31
2.10 Projection of b Vector onto the Column Space of a 3 by 2 Matrix A ...	31
3.1 Typical Allele Peaks at Ten Loci from a DNA Mixture of a Male and a Female.....	36
3.2 Typical Curves of the Residuals for All Possible Genotype Combinations Across the Range of Mass Proportion of Person 1 in the Mixture for a 4-Allele Locus.....	43

3.3	Flow Sheet of LMA to Resolve Two-People DNA Mixture.....	49
3.4	The Calculated Squared Deviation $dev(g(w))$ Against the Mixture Weight w_A for a DNA Mixture Sample with the Mass Ratios of 9:1....	51
4.1	Flow Sheet of LSD to Resolve Mixture DNA Samples.....	55
4.2	Flow Sheet of Using LSD to Resolve Mixed DNA at a 2-Allele Locus.....	64
4.3	The Allele Peaks at Nine Loci of Texas DNA Mixture Data No. 1.....	66
4.4	Output of Web-LSD for Texas DNA Mixture Data No. 1.....	68
4.5	Normalized Fitting Residuals of All Possible Genotype Combinations at the Locus VWA in the DNA Mixture Data Shown in Table 4.4.....	75
4.6	Normalized Fitting Residuals of All Possible Genotype Combinations at the Locus D21S11 in the DNA Mixture Data Shown in Table 4.4...	78
5.1	Maximum Gaussian Noise Present in the Allele Peak Data Tolerable by LSD in Resolving DNA Mixture Data.....	88
5.2	Output of Web-LSD for the Simulated DNA Mixture Data with a Mass Ratio of 1:5.....	93
5.3	Output of Web-LSD for the Simulated DNA Mixture Data with a Mass Ratio of 1:2.....	98
5.4	Output of Web-LSD for the Simulated DNA Mixture Data with a Mass Ratio of 1:1.....	103
5.5	Output of Web-LSD for the DNA Mixture Data Presented in Reference [3].....	113
5.6	Output of Web-LSD for the DNA Mixture Data Presented in Reference [21].....	118
5.7	Output of Web-LSD for Acadiana DNA Mixture Data No. 5.....	124
5.8	Output of Web-LSD for Acadiana DNA Mixture Data No. 1.....	132

5.9	Output of Web-LSD for Acadiana DNA Mixture Data No. 2.....	139
5.10	The Allele Peaks at Nine Loci of Texas DNA Mixture Data No. 3.....	149
5.11	Output of Web-LSD for Texas DNA Mixture Data No. 3.....	151
5.12	The Allele Peaks at Nine Loci of Texas DNA Mixture Data No. 4.....	156
5.13	Output of Web-LSD for Texas DNA Mixture Data No. 4.....	158
5.14	Reproduced from [18]: LMA Genotype Resolution Result of the DNA Mixture Data Shown in Table 5.30.....	167
5.15	Genotype Resolution Result of the DNA Mixture Data in Table 5.30, Using LMA Algorithm Modified from the Original Paper Implemented by the Author.....	168
5.16	Sum of Squared Deviation as a Function of the Set of Assumed Mass Fractions of LMA on Mixture Data Presented in Table 5.30.....	169
5.17	Sum of Squared Deviation as a Function of the Set of Assumed Mass Fractions of LMA on Mixture Data Presented in Table 5.32.....	176
5.18	Sum of Squared Deviation as a Function of the Set of Assumed Mass Fractions of LMA on Mixture Data Presented in Table 5.35.....	181

GLOSSARY

Locus	A region of a gene with a specific sequence of interest
Allele	A possible DNA base sequence at a STR locus; the name of the allele represents the number of the tetra base repeat units within the allele sequence of interest
Genotype	Specification of the alleles found at DNA loci, for human identification purposes
DNA Typing	The process of identifying the types of alleles at the different loci of a STR-DNA sample

Abbreviations

LSD	Least-Square Deconvolution; a novel method to resolve DNA mixtures of two people using allele peak areas or heights
DNA	Deoxyribonucleic acid
bp	Base Pairs
PCR	Polymerase-Chain-Reaction; a process for amplifying a selected DNA segment from a minute quantity of original DNA
STR	Short-Tandem-Repeat; DNA markers found to be different among the human population; used for human identification purposes; such regions are composed of short repeated sequences arranged in tandem with flanking sequences
VNTR	Variant Number of Tandem Repeat; a DNA segment with medium-length repeat units; an older set of DNA markers used for human identification purposes
LMA	Linear Mixture Analysis; a newly reported method to resolve DNA mixture, presented in [18]
CODIS	Combined DNA Index System; a national DNA database and search system used in forensics in the US

CHAPTER 1

INTRODUCTION

DNA identification has become one of the most important application tools in forensic science since DNA typing methodologies were introduced around 1985, due to its high degree of discrimination of forensic human identification [1]. It may be regarded as one of the most important discoveries in the forensic field since the introduction of fingerprinting. DNA typing, with its extremely high capability to differentiate one individual from another [2], has become widely used in courts around the country and in the world.

In recent years, with the improvement of DNA typing technology and the availability of a national DNA database, the DNA typing method has become a popular and effective tool in forensic applications. The DNA analysts and other law enforcement agencies have been trying to understand the hidden information contained in DNA through an understanding of the molecular biology, genetics, and statistics involved to provide justice to trials. Since the probability that one person has the same genotype at a set of prescribed DNA loci as another person is very small, it is widely used in forensic identification, especially when unidentified criminals leave testable evidence such as semen, blood, and saliva at crime scenes. These important stains provide the extracted DNA samples that are to be used for criminal identification.

DNA typing for forensic applications is based on applying the statistical tools to

the fundamental principles of diagnosis and gene characteristic analysis [2]. The DNA profile extracted from the criminal evidence maintains its unique identity, and the characteristic of the profile is analyzed using these methods. The objective of DNA typing is to identify the genotype of the individual who left the evidence. After the perpetrator's genotype is obtained after DNA analysis, the forensic caseworkers can compare the genotype of the criminal with that of the suspect or search for a matching DNA profile in the national CODIS database for a possible suspect [2]. Therefore, as an early step in investigation, DNA typing results of forensic samples should be obtained.

1.1. Motivation for Work Undertaken

In many cases, especially in rape cases, when a DNA sample is extracted from a biological stain containing body fluids or tissues from more than one person, the result is often a mixed DNA profile. This kind of DNA profile is essentially composed of one contributor's DNA sample superimposed on that of another [3]. Much of the DNA evidences obtained from crime scenes are mixtures. Generally, the genotype of the victim is known, but the genotype of the perpetrator cannot be obtained clearly and directly due to the presence of DNA of another person in the mixed DNA profile. The genotype of each contributor to the DNA mixture must be deciphered first before further investigation.

Until now, the deconvolution of mixed DNA profiles contributed by multiple people has been one of the most challenging tasks facing forensic scientists. Part of the difficulty is due to the large number of possible genotype combinations from the multiple

contributors [4] that would result in the same mixed DNA profile. So far, no analytical and reliable method has been published for the resolution of DNA mixture into its components. It is desirable that an efficient method be formulated to resolve the DNA mixture sample and to provide the most likely genotype of each contributor. Therefore, this thesis work was undertaken with the goal of developing an analytical method, well supported by mathematical theory, to resolve DNA mixtures based on quantitative measurements of the DNA allele peak signals, either area or height.

Applying automated DNA allele sequencers in typing and analyzing the Short Tandem Repeat (STR) loci regions of human DNA provide not only qualitative information about the profile, but also quantitative information of the relative amount of amplified DNA alleles in the sample reflected by the relative peak area or height of the measured alleles. This quantitative information allows for the interpretation of DNA mixtures in a mathematical way not previously possible with other STR/DNA mixture interpretation approaches. The most important and most useful quantitative information about the DNA mixture is the allele peak areas or heights obtained from the electropherogram. The peak areas or heights can be used to determine the most likely genotypes of the mixture sample's contributors. This is based on the principals of superposition of the allele peaks in a mixed DNA profile [5]. Dr. Tsewei Wang (Department of Chemical Engineering) and Dr. J.D. Birdwell (Department of Electrical and Computer Engineering) have proposed a new approach, the Least-Square Deconvolution (LSD), to resolve DNA mixture samples [6]. This new approach uses a least-square optimization method to obtain the genotype combination that provides the

best fit to the observed allele peak areas or heights in a mixed DNA profile.

The use of allele peak area and height is interchangeable in the least square fit, as long either is used consistently throughout. Henceforth, allele peak areas and heights shall be used interchangeably throughout this document.

1.2. Anticipated Importance of LSD in Forensic Work

LSD is a systematic mathematical routine that gives the best-fit mass proportions and the genotype combination profiles of the two contributors, from a mixture sample data, based on the given allele peak data. To date, STR mixture DNA is attempted to be resolved only by experienced DNA analysts. Usually they are able to resolve with confidence at only a few of the 13 loci. In general, 4-allele loci with clear major/minor peaks, and some 3-allele loci whose allele peak data indicate a certain ratio relationship are attempted to be resolved. The remaining loci are left as mixture loci. LSD, on the other hand, can lead to a clear resolution at most of the 13 core loci with full confidence. Thus a composite genotype profile for the perpetrator can be developed for follow-up investigation as outlined below. Two avenues of pursuit can follow a confident mixture resolution by LSD. First, the caseworkers can compare the LSD identified perpetrator's profile with that of a suspect, if available, or second, to search for a matching DNA profile in DNA databases if no suspect exists. If a suspect's genotype is found to be exactly the same as that of the perpetrator identified from LSD, then the match probability can be calculated to access the odds that the suspect is indeed the perpetrator. If no suspect is available or the suspect is found not to match that of the LSD identified

perpetrator's profile, the genotype of the perpetrator can then be searched in the convicted offenders DNA databases or in the forensic DNA databases for a matching profile. If a matching DNA profile is found, the owner of this matching profile can further be investigated to see if he can indeed be the perpetrator. It is envisioned that LSD would aid greatly in those forensic cases, in which mixture DNA samples are hard to resolve by traditional means and by a more conservative approach (than LSD).

1.3. Problem Statement

For many DNA mixture profiles, the allele peak area is the only quantitative information available for the caseworkers. Figure 1.1 shows a set of simulated allele peaks at one locus for a mixed DNA sample contributed by two individuals. Three alleles, 3, 7, and 15 are present at this locus with the peak height ratio of 1:0.4:1.4. The genotypes of the two contributors at this locus are [3, 15] and [7, 15]. If this genotype combination is not available, there would be a total of six possible genotype assignments for the two contributors at this locus, as shown in Table 1.1. The problem is to decide which one is the best fit to the observed allele peak heights, based on the relative mass of each contributor's DNA in the mixture.

The goal of this research work is the formulation, development, and evaluation of the least-square approach using quantitative allele peak areas, in arriving at the best-fit genotype combination at all available loci for the DNA mixture sample contributed by two people. The least-square deconvolution algorithm will be tested by first using simulated, and then real mixture DNA sample profiles.

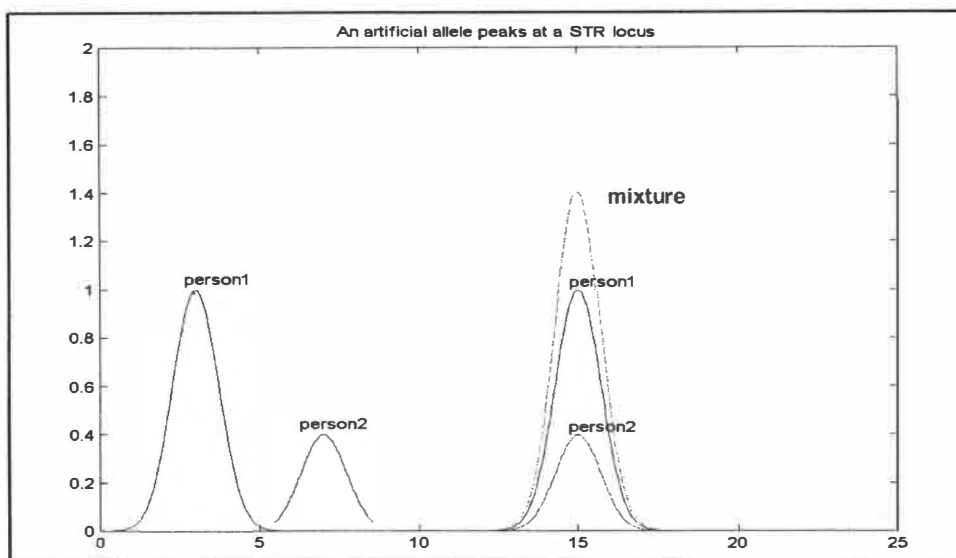


Figure 1.1: Simulated superimposed allele peaks at one locus for a mixed DNA profile, with the contributors' genotypes of [3, 15] and [7, 15], mixed at a mass ratio of 1:0.4 respectively. Note that allele 15 is the shared allele.

Source: T. Wang, N. Xue, M. Radar, and J.D. Birdwell, "Least Square Deconvolution (LSD) of Mixture STR/DNA Using Peak Area / Height Data," The 7th CODIS User's conference, Arlington, VA, October 2001.

Table 1.1: All possible genotype combinations for the two contributors at a three-allele locus as shown in Figure 1.1 for a mixed DNA sample.

Case	Genotype Combination	
1	3, 3	7, 15
2	7, 7	3, 15
3	15, 15	3, 7
4	3, 7	3, 15
5	3, 7	7, 15
6	3, 15	7, 15

CHAPTER 2

BACKGROUND

2.1. Genetic and Technical Basis of DNA Typing

2.1.1. Fundamentals of Genetics

In higher organisms, the genetic material is organized into microscopic structures called chromosomes. A fertilized human egg has 46 chromosomes (23 pairs), which, with appropriate microscopic techniques, are visible in the cell nucleus. The two members of each pair are homologous and come from different sources: one from the sperm, and the other from the egg. Through the process of chromosomal duplication during cell division, the chromosomal contents are identical in the two daughter cells as that in the parent. With a few exceptions, all the cells in the body have chromosomes identical with those of the original fertilized egg.

Usually, chromosome is not visible under an ordinary microscope. With the help of an electron microscope, DNA was found to be organized into the form of a double helix, with two strands coiled around each other linked together by hydrogen bonds between the two strands resembling a twisted rope ladder with connected wooden steps. The basic chemical unit of DNA is a nucleotide, consisting of a base (a half-step in the ladder) and a sugar-phosphate complex. There are four kinds of bases, designated A for adenine, G for guanine, T for thymine, and C for cytosine [7]. The nucleotides of one DNA strand pair up in hydrogen bonds in a specific fashion with those of the other strand

to form the ladder; because of their size, chemical properties, and complementary structure, T always pairs with A, and G with C [8]. If the sequence of nucleotides on one strand is available, one automatically knows the sequence on the other strand, but with the two strands running in opposing directions. A gene is a sequence of DNA consisting of base pairs, arranged in some specified order, and the position that a gene occupies along the DNA is referred to as a locus [2]. Genes are differentiated by their specific sequences of base pairs. The type of the specific gene sequence of a locus is called an allele. There are two possible alleles present at one locus of a chromosome: one from the father, and the other from the mother [8]. If these two copies are the same, the DNA profile is referred to as homozygous at that locus. If these two copies are different, the DNA profile is heterozygous at the locus [8].

2.1.2. Introduction of DNA Identification

DNA typing for human identification purposes was first introduced around 1985 [1]. Compared with other identification methods, DNA typing has three advantages:

- Every person has his/her own genotype. At certain DNA marker sites, the probability that one person has the same genotype as another person is very small [2];
- DNA is strongly resistant to surrounding conditions, making DNA more stable than other biological compounds [8];
- Only a tiny amount of DNA sample is needed for human identification when employing the PCR amplification method. The DNA profile maintains its identity after amplification. Usually a DNA sample on the order of 10^{-9} grams collected from a crime scene is enough for subsequent DNA analysis and identification [9].

For such reasons, DNA typing has been widely used in identification, in forensics and in searching for missing people. The advent of Polymerase Chain Reaction (PCR), a method of amplifying DNA, and the identification of Short-Tandem-Repeat (STR) DNA loci markers improve the application of DNA typing and reduces costs, while achieving a high level of discrimination.

2.1.2.1. Gene Amplification by the Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction (PCR), first introduced around 1985 [7], is a laboratory process for duplicating a selected short segment of DNA millions of times [2]. It has rapidly gained acceptance as a tool in many fields, such as molecular biology, population genetics, and medical diagnostics; it can also be viewed as the breakthrough technology that enabled the human genetic mapping project's successes [8]. This PCR process is similar to the mechanism by which DNA duplicates itself in normal condition. The PCR amplification process includes a three-step cycle (Figure 2.1) [2]:

1. The double-stranded DNA segment is first separated into two single strands by incubation at high temperature [2].
2. After being lowered to a lower temperature, the oligonucleotide fluorescent-labeled DNA primers are added and they become attached to their complementary sequences of the template strand, to start the DNA replication step [7].
3. In the presence of the enzyme DNA polymerase, and the four DNA bases (A, C, G, and T), each separated strand serves as a template to replicate the target sequence, by elongating the primer sequence, between two adjacent primer binding sites, according to the rules that A pairs with T, and G with C only [2].

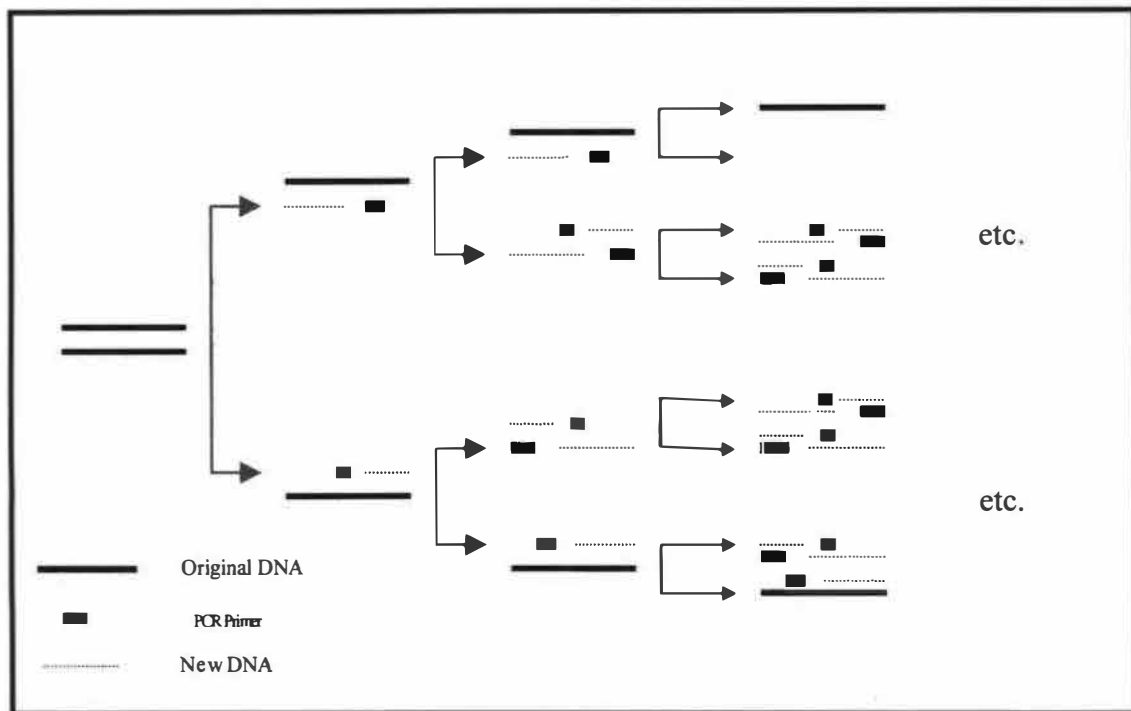


Figure 2.1: A simulation of the PCR process used to amplify each DNA segment of interest 20-35 times; 3 cycles are shown in the figure.

Source of information: M. Farley and J. Harrington, "Forensic DNA Technology," Lewis Publishers, INC, 1991.

This three-step process is repeated. Usually each DNA template is duplicated through 20-35 cycles. In one round, the two strands of each DNA segment yield four copies; the four into eight copies; and so on, until the number of copies of the original DNA segment becomes enormous after tens of cycles, to generate tens of millions of copies of the sequence of interest, as illustrated in Table 2.1. The main difference between this procedure and the normal DNA duplication process is that the PCR process only amplifies a small DNA region of specific interest. The length of the region is usually not more than 1,000 nucleotide base pairs long [2]. Thus, PCR methods cannot be used to amplify large DNA regions. Short tandem repeats (STR) are usually a few hundred nucleotide base pair units in length [2], and STRs are ideally suited for PCR. This system is growing in use as more STR markers are recognized by forensic scientists [2].

The PCR process can be easily carried out in laboratories. Results can be obtained within 24 hours [2]. Since there is almost no limitation to the number of cycles of

Table 2.1: Numbers of copies of target DNA segment after different rounds of PCR amplification process.

Number of PCR Cycles	Number of DNA Segments
1	2
2	4
5	32
10	1, 024
15	32, 768
20	1, 048, 576
25	33, 544, 432

amplification, it is possible to analyze very tiny amounts of extracted DNA samples with PCR-based methods. This advantage makes PCR methods widely used in forensic analysis, because in some forensic samples, such as a small piece of fingernail, or a tiny amount of saliva, only a small amount of DNA can be extracted. At present, no method, other than the PCR process, can amplify such a small amount of DNA [2]. In addition, the small amount of DNA required for the PCR process makes it possible to amplify samples that contain degraded DNA; this allows DNA typing of old and highly decayed samples, such as remains of accident victims, and decayed bodies [2]. Another consequence of PCR-based methodologies is that they usually yield exact copies of each allele present at each locus [2]. Because of these advantages, PCR-based DNA typing is widely and increasingly used in forensic laboratories in many countries [2].

2.1.2.2. Short Tandem Repeat (STR) Markers

Some regions of human genomes include repeated DNA sequences of different types and sizes [10]. These DNA sequences can be differentiated by either the length of the core repeat units, and the number of continuous repeat units, or by the full length of the repeat segment [10]. Long repeat units sometimes have several hundred to several thousand base pairs. A medium-length repeat may be referred to as Variant Number of Tandem Repeats (VNTR) [10]. Usually a VNTR has approximately 10-100 repeat units [10]. Short Tandem Repeats (STR) stands for the DNA segments with repeat units of 2-6 bp in length [10]. Each STR member site is called a 'locus'. Each STR locus of a human chromosome has two alleles: one from the father, and the other from the mother. Even for a heterozygous locus, both alleles are almost the same in size because the size of the

repeat region is small. Therefore, up to date, STRs are the most versatile and popular DNA repeat markers that are amplified with PCR process without the problems of differential amplification that occur with VNTRs [10].

The number of different possible alleles for each STR locus in a population varies from 7 to 30, depending on the type of locus [11]. The name of the allele represents the number of the repeat units at that locus. DNA typing refers to the identification of the allele types at several STR loci of a DNA sample. The alleles present at each locus can exhibit great variability among the individuals of a population group, with different probability of occurrence, thus making them highly effective for human identification.

To date, a large number of STRs have been recognized by the academic and forensic laboratories for identification purposes [10]. For example, in order to develop human genetic maps, genotype data of more than 8000 STRs (loci), scattered across the 23 pairs of human chromosomes, have been collected in the Marshfield Medical Research Foundation in Marshfield, Wisconsin [10]. In 1997, based on the population statistics of the United States, 13 STR loci were selected as the core loci for the FBI CODIS national database [10]. The 13 core loci are CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, and D21S11 [11]. Scientists also study the frequency distributions of the alleles at all 13 loci for different populations. Examples of distribution are shown in Figures 2.2-4 [11]. The corresponding probability data are listed in Tables 2.2-4 [11]. These figures show the

Table 2.2: Frequency distributions of alleles at locus CSF1PO for Caucasians.

Alleles	Probabilities
7	0
8	0.003
9	0.028
10	0.268
10.3	0.003
11	0.258
12	0.358
13	0.07
14	0.015
15	0

Source of data: B. Budowle, “ Probabilities of alleles’ distribution at 13 loci,” Department of Justice, Federal Bureau of Investigation, October 1997.

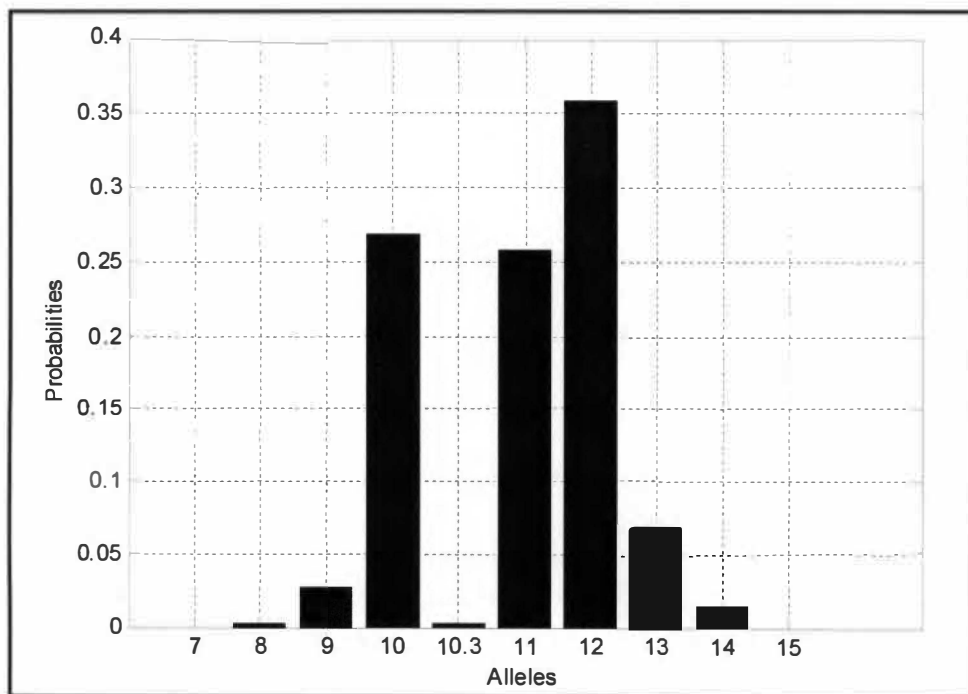


Figure 2.2: Bar chart of frequency distributions of alleles at locus CSF1PO for Caucasians.

Source of data: B. Budowle, “ Probabilities of alleles’ distribution at 13 loci,” Department of Justice, Federal Bureau of Investigation, October 1997.

Table 2.3: Frequency distributions of alleles at locus D7S820 for Caucasians.

Alleles	Probabilities
6	0
6.3	0.003
7	0.025
8	0.175
9	0.13
10	0.24
11	0.23
12	0.16
13	0.028
14	0.008
15	0.003

Source of data: B. Budowle, "Probabilities of alleles' distribution at 13 loci," Department of Justice, Federal Bureau of Investigation, October 1997.

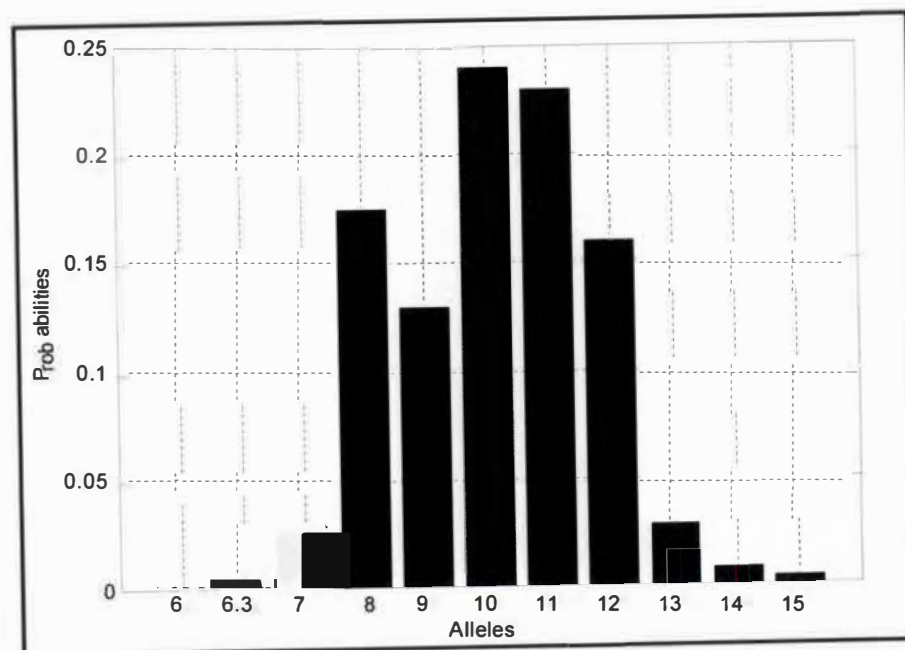


Figure 2.3: Bar chart of frequency distributions of alleles at locus D7S820 for Caucasians.

Source of data: B. Budowle, "Probabilities of alleles' distribution at 13 loci," Department of Justice, Federal Bureau of Investigation, October 1997.

Table 2.4: Frequency distributions of alleles at locus D8S1179 for Caucasians.

Alleles	Probabilities
8	0.018
9	0.013
10	0.094
11	0.066
12	0.143
13	0.333
14	0.209
15	0.088
16	0.031
17	0.004
18	0

Source of data: B. Budowle, “ Probabilities of alleles’ distribution at 13 loci,” Department of Justice, Federal Bureau of Investigation, October 1997.

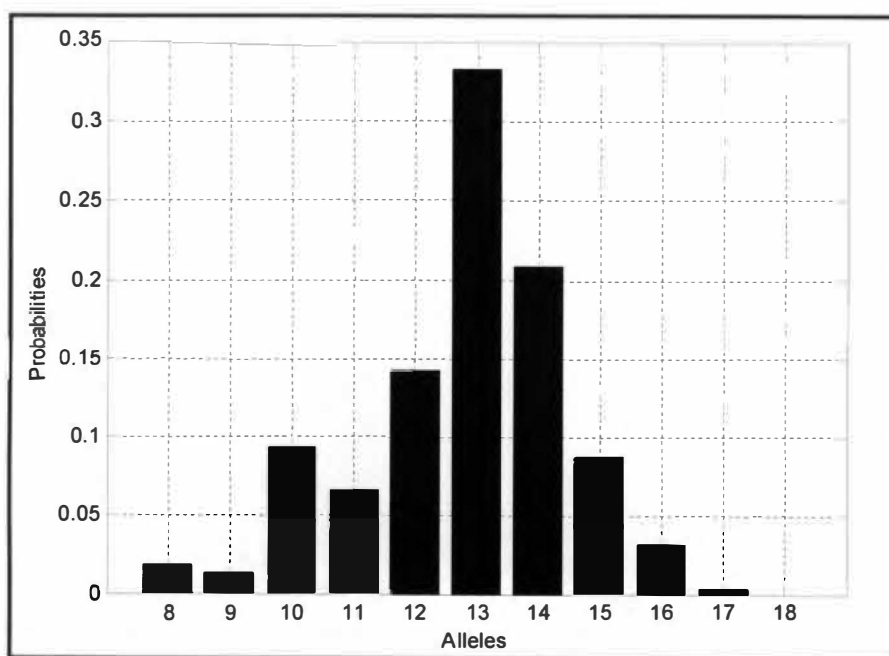


Figure 2.4: Bar chart of frequency distributions of alleles at locus D8S1179 for Caucasians.

Source of data: B. Budowle, “ Probabilities of alleles’ distribution at 13 loci,” Department of Justice, Federal Bureau of Investigation, October 1997.

frequency distributions of the alleles at each of the three loci: CSF1PO, D7S820, and D8S1179 for the Caucasian population. From the bar charts, one can see that the probabilities that a person has the different alleles at a locus are not distributed evenly across all allele values. According to Reference [1], the power of discrimination using all 13 loci is about 1.74×10^{-15} , or one in 575 trillion, which means that the number of all possible genotypes using all 13 loci is about 575 trillion. However, this does not imply that in a DNA database with the size of 575 trillion or even one with a much smaller size (say, one with 10^6 profiles), the probability that at least two random people not related sharing exactly the same genotype profile is also 1.74×10^{-15} . This probability is much higher than 1.74×10^{-15} (B. Budowle and B. Brown of the FBA Laboratories, personal communication).

2.1.2.3. Instrumentation for the Detection of Alleles of STR/DNA Loci

Identification and quantification of allele peaks are the most important tasks in STR/DNA typing of profiles. In order to automatically detect and measure the fluorescent signals of the amplified STR/DNA, two fluorescence detection systems and related software have been developed by commercial companies in the U.S., one is the ABI Prism 310 Genetic Analyzer, the other is the HITACHI FMBIO II Fluorescence Imaging System [10].

Introduced in 1995 by Perkin Elmer Applied Biosystems, ABI Prism 310 Genetic Analyzer is widely used in American forensic laboratories, including the FBI laboratory in Washington, D.C. [10]. In this kind of genetic analyzer, capillary electrophoresis (CE)

is used to separate the different DNA alleles [10]. The process steps are presented in Figure 2.5. Figure 2.6 shows a typical output result of a DNA profile from this type of genetic analyzer. One advantage of this instrument is that, in addition to sizing the detected allele signals, the related software can also display their peak heights and automatically calculate the area under each peak [10]. The most important characteristic of the allele peak is that the measured peak area or height is roughly proportional to the mass of the corresponding allele in the DNA sample [10]. Compared with the peak height, peak area is a better indicator of the relative amount of allele present in a DNA sample, because the former can be affected by the shape of the allele peak [3]. Therefore, the peak area of each allele provides the scientists with more valuable quantitative information about the DNA sample being tested. Such characteristics of allele peaks can also make artifact peaks easily identifiable, and eliminated from the genotype of the DNA sample. Moreover, no background information about capillary electrophoresis is required for caseworkers to operate such a system [10].

Compared with the ABI Prism 310 Genetic Analyzer, the HITACHI FMBIO II Fluorescence Imaging System uses gel electrophoresis instead of capillary electrophoresis to separate the alleles of a DNA sample [10]. This system requires much more sample and a longer time to completion. In this genetic analyzer, each allele corresponds to a specific band in a gel lane. The band size for each allele is compared with a well-calibrated allelic ladder to identify the corresponding allele [10]. This kind of genetic analyzer can only detect the genotype of a DNA sample; the quantitative peak information of each allele cannot be measured. In addition, artifact alleles cannot be

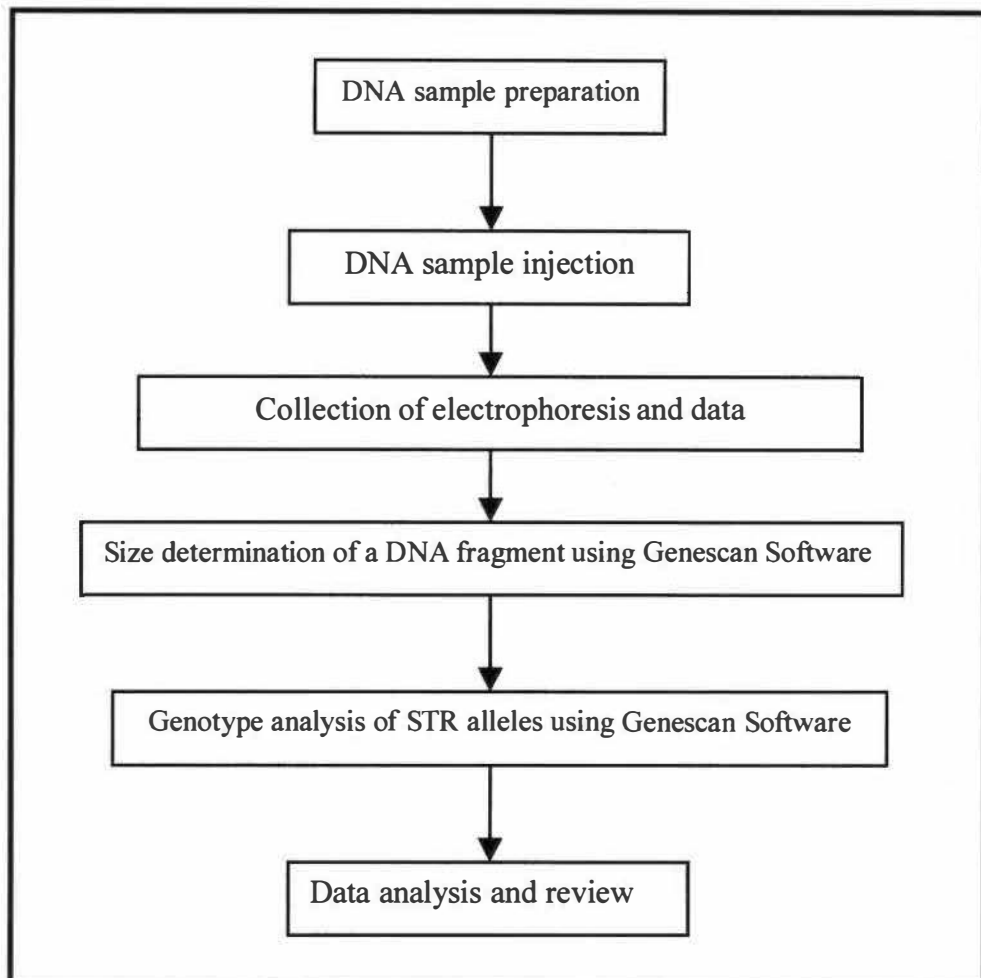


Figure 2.5: Typical processing steps of using the ABI 310 Genetic Analyzer to identify alleles present at each locus.

Source of information: J. Butler, "Forensic DNA Typing," Academic Press, 2001.

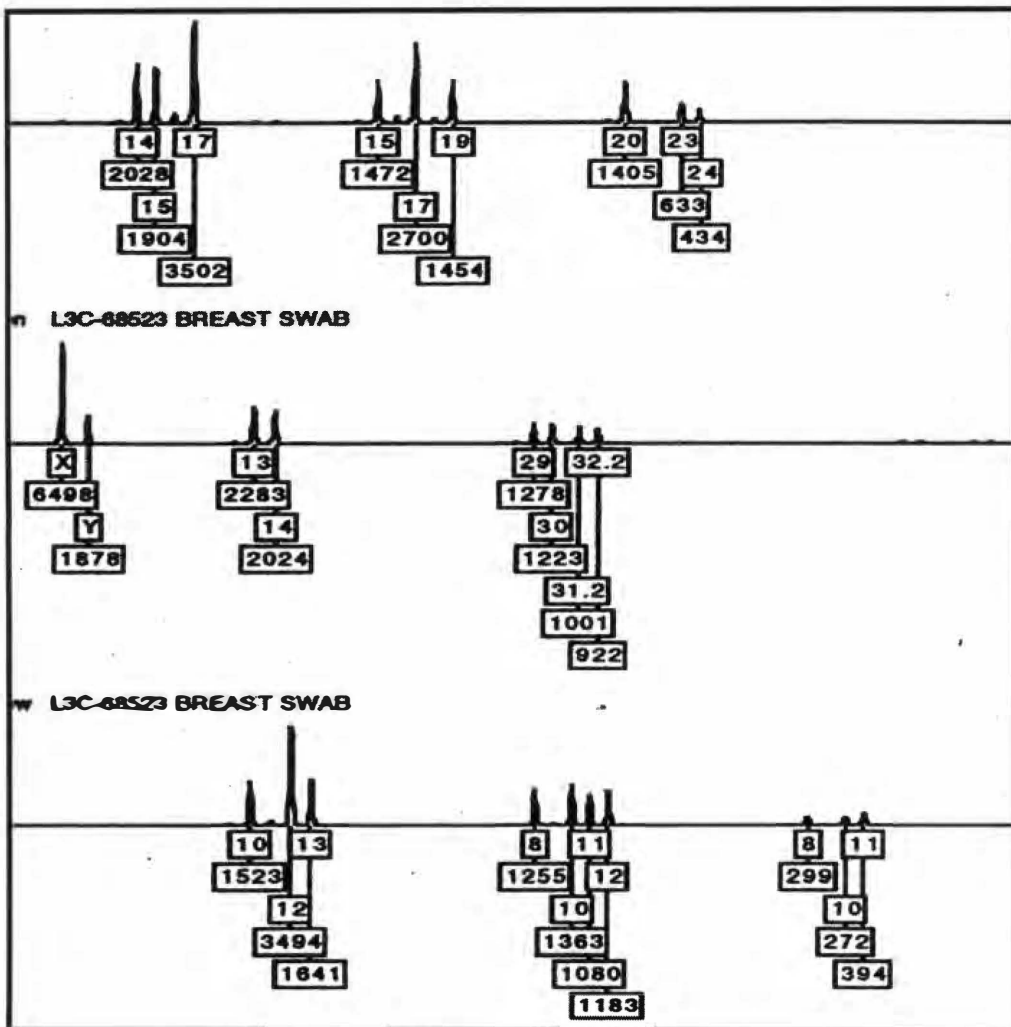


Figure 2.6: Typical analysis output of a DNA profile using the ABI 310 Genetic Analyzer. Each peak corresponds to one allele, and the peak area is proportional to the mass of the allele it represents.

Source: P. Graham, "Allele Peak Data for DNA Mixture Samples from Real Cases," Texas Department of Public Safety, Corpus Christi, Personal Communication, 2001.

easily differentiated from the analysis result. Therefore, this kind of system is not widely used in forensic laboratories in the U.S. [10].

2.2. Introduction to the Problem of Resolving DNA Mixtures

Mixture DNA samples consist of DNA from more than one person with possible allele sharing among the contributors' genotypes [3]. This kind of DNA profile is common in rape and murder cases, where the collected DNA sample is a mixture containing DNA profile from both the perpetrator and the victim.

2.2.1. Identification of DNA Mixtures

According to studies reported in the literature [10], if only one person contributes to a DNA sample, then no more than two alleles will be detected at a locus. Therefore, a DNA/STR sample is highly likely to be a mixture if more than two alleles are present consistently at multiple loci. For a locus with two heterozygous alleles from a single contributor, if the higher peak is set as 100%, the heterozygous companion allele, due to the co-amplification of the two alleles, usually has a peak area or peak height greater than 70% of the higher allele [10]. A severe peak-height imbalance between any two heterozygous alleles at a locus strongly suggests that the DNA sample being tested is a mixture. Figure 2.7(a) is an illustration of a typical heterozygote DNA/STR profile from a single contributor. Note that only two 'true' alleles are present; the two minor peaks are considered to be stutter peaks [10], an artifact of the DNA amplification procedure, whose characteristics will be explained later. In the following paragraphs, some

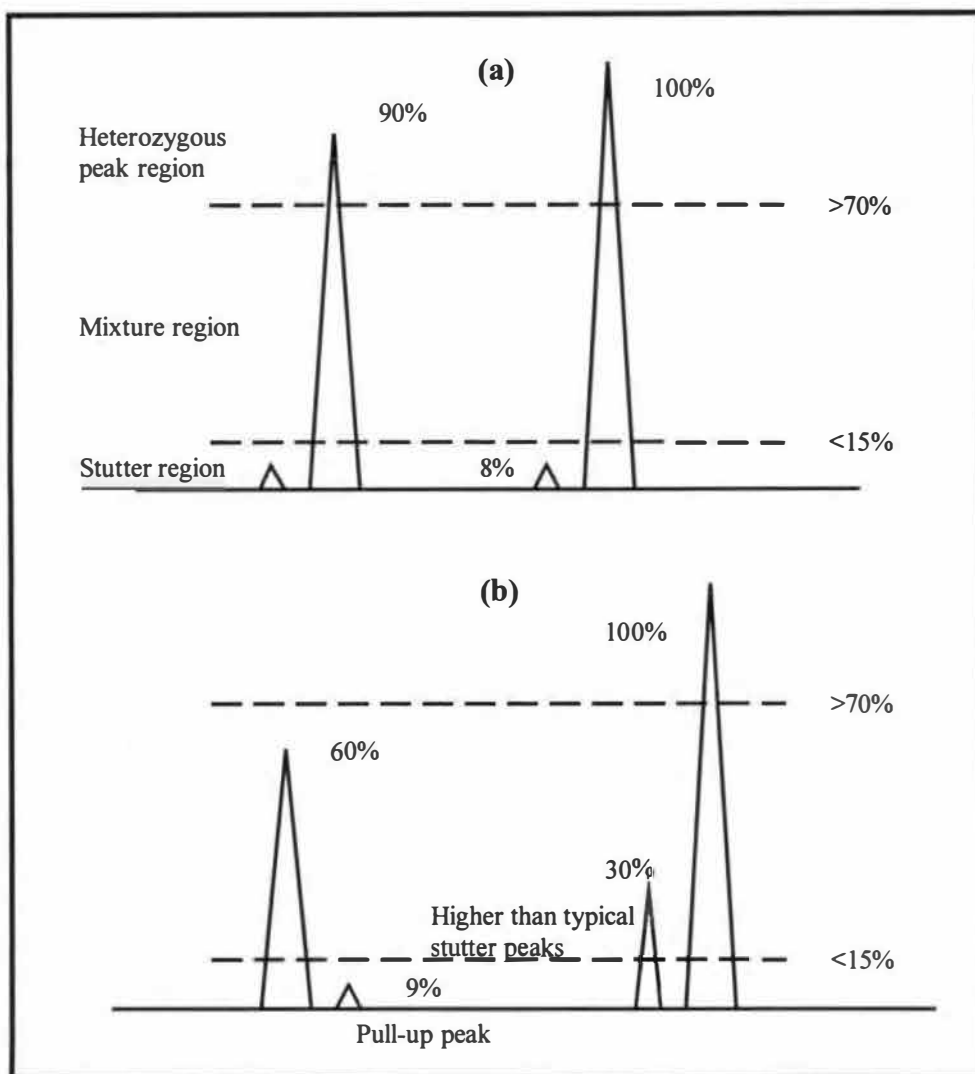


Figure 2.7: Illustration of allele peaks. (a) Illustration of the allele peaks of a typical heterozygous locus from a single source. (b) Illustration of the allele peaks of a DNA sample from multiple sources containing three true alleles at one locus; stutter peak always appears to the left of a true allele peak and is less than 15% of the major peak height.

Source of information: J. Butler, "Forensic DNA Typing," Academic Press, 2001.

guidelines will be presented for the determination of whether a minor peak is to be classified as an artifact.

There are several kinds of peaks, other than the peaks of 'true' alleles that affect the identification of the DNA mixtures: stutter peaks, artifact peaks, and pull-up peaks. The following is a brief review of these three types of peaks [10, 12].

Stutter peaks are generated by the enzyme's slippage during the DNA PCR amplification process [12]. In most cases, stutter peaks are located on the left side of the associated alleles, and the gene distance between the stutter peak and the associated allele peak is usually less than 4 bp [12]. The height of the stutter peak is usually less than 15% of the height of the corresponding true allele peak [12]. Figure 2.7 (a) shows two typical stutter peaks - the peak height is about 9-10% of the highest peak. (Besides the stutter peak, there are only two peaks present at this locus. Therefore, it cannot be concluded with certainty that the peaks in Figure 2.7 (a) come from a DNA mixture sample.)

Artifact peaks are due to impurities in the DNA samples. Generally, the artifact peaks have the following three characteristics:

- About 53% of them are less than 5% of the nearest allele peak's height [12];
- Some artifact peaks consist of multiple peaks. The distances among them are always less than 1 bp [12];
- Some artifact peaks are within 0.5bp of an allelic ladder marker [12].

If a peak satisfies any of the above three rules, the peak can be defined as an artifact peak,

and the peak's effect can be eliminated.

A minor peak directly to the right of or below (in the gel position) a 'true' allele peak is defined as a pull-up peak. Usually, a pull-up peak is always located on the right side of a 'true' allele peak with a distance less than 3/8 bp, and the height is less than 50% of the major peak. Figure 7(b) shows a typical pull-up peak with a peak height of 9% of the highest peak. After eliminating the effect of the pull-up peak, there are still three peaks present at this locus. Therefore, it can be concluded that the peaks in Figure 7(b) come from a DNA mixture sample, especially if some other loci also contain more than two alleles

If a sample comes from only one person, usually no more than two true alleles can be detected at one locus. If there are still more than two peaks appearing at several loci, even after eliminating the effects of the stutter peaks, the pull-up peaks, and the artifact peaks, it can be concluded that the sample contains a mixture from at least two separate contributors.

2.2.2. Identification of the Number of Contributors

Since usually no more than two alleles can be detected at one locus for a one-person DNA sample, the maximum number of alleles that can be detected at one locus for a two-person mixed DNA/STR profile is four. Five or more alleles present at several loci would be strongly indicative of three or more contributors [3]. In the national CODIS DNA database in 2002, approximately 9% of DNA profiles are mixture samples, but the DNA

analysts spend 90% of their efforts and time trying to interpret them [13]. In most forensic cases, especially in rape cases, a great number of mixed DNA samples are contributed by two individuals, although three-people mixtures are also encountered occasionally [3]. For this reason, this thesis work concentrates only on resolving two-people mixtures.

2.2.3. Characteristics of Allele Peaks of STR/DNA Mixtures

Although existing statistical models may aid in interpreting mixtures qualitatively based on the alleles present [14, 15], using automated sequencers opens the opportunity to account for quantitative aspects embodied by the peak area of each allele. This is because each peak corresponds to one allele at a locus in the DNA/STR profile and the peak area measured is theoretically proportional to the mass of the allele it represents. For DNA mixtures, some alleles may be shared by the contributor who each has the shared alleles. Thus, the peaks in the mixture sample can be regarded as the superposition, or sum of the two individuals' allele peaks. These important characteristics allow one to resolve the DNA mixture profile by using the allele peak information. Figure 1.1 shown earlier is an artificial example of allele peaks contributed by two people. Note that only two heterozygous alleles associated with the same contributor have equal peak areas (in the ideal case), and allele 15 is shared between the two individuals. In addition, the peak area of alleles from person 2 is about 40% of that from person 1. The question to be answered in interpretation of the given composite profile is: Given the measurements of the relative areas of the three allele peaks at this locus $\{1:0.4:1.4\}$, what is the most likely

genotype profiles of the two contributors? All possible genotype combination profiles for the two contributors, when superimposed together at this locus, that can yield the allele pattern of [3, 7, 15] are listed in Table 1.1, disregarding the relative amount of DNA at these three alleles. However, the relative allele peak area information supports only the conclusion that the allele combination of {[3, 15] [7, 15]} in a mixture ratio of 1:0.4 provides a high-quality fit to (or explains the measurements of) the observed allele area information of {1:0.4:1.4}, respectively.

2.3. The Linear Least-Square Solution to Data Fitting

In this research work, multivariable linear equations ($Ax = b$) are to be solved for data fitting. Most overdetermined multivariable linear equations do not have exact solutions because they are inconsistent with each other. In those cases, the least-square approach can be used to obtain a best-fit solution [16], which yields a fitted $\hat{b}$ comes closest to the given b .

2.3.1. Introduction to the Least-Square Approach

Let A be an m by n ($m > n$) matrix and b be a column vector composed of m elements. For a linear equation of $Ax = b$, an exact solution for this equation, x , can be obtained only when b is in the column space of A . (The vector b can be expressed as a linear combination of the columns of the matrix, A . The elements in the vector, x , determine the linear combination.) The multiplication of A and x can be viewed as:

$$Ax = b \Rightarrow \begin{bmatrix} | & | & | & | \\ a_1 & a_2 & \dots & a_n \\ | & | & | & | \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \Rightarrow x_1 \begin{bmatrix} | \\ a_1 \\ | \end{bmatrix} + x_2 \begin{bmatrix} | \\ a_2 \\ | \end{bmatrix} + \dots + x_n \begin{bmatrix} | \\ a_n \\ | \end{bmatrix} = \begin{bmatrix} | \\ b \\ | \end{bmatrix}$$

For example: if A is $\begin{bmatrix} 1 & 2 \\ 6 & 5 \\ 3 & 4 \end{bmatrix}$, and b is $\begin{bmatrix} 8 \\ 27 \\ 18 \end{bmatrix}$, the linear equation: $Ax=b$ ($\begin{bmatrix} 1 & 2 \\ 6 & 5 \\ 3 & 4 \end{bmatrix} x = \begin{bmatrix} 8 \\ 27 \\ 18 \end{bmatrix}$)

has an exact solution x , because b can be expressed as a linear combination of the columns of A as shown below:

$$\begin{bmatrix} 1 \\ 6 \\ 3 \end{bmatrix} 2 + \begin{bmatrix} 2 \\ 5 \\ 4 \end{bmatrix} 3 = \begin{bmatrix} 2+6 \\ 12+15 \\ 6+12 \end{bmatrix} = \begin{bmatrix} 8 \\ 27 \\ 18 \end{bmatrix} \Rightarrow \begin{cases} x_1 = 2 \\ x_2 = 3 \end{cases}$$

Therefore, the solution for this multivariable linear equation is $x = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$, where A and b are given as stated.

When b cannot be expressed as a linear combination of the columns of A , as shown in the example described below, then no solution, x , would exist that exactly satisfies the equation $Ax=b$. Such system is termed inconsistent and has no solution [16].

For example, the simultaneous equation: $\begin{cases} 9x = b_1 \\ 5x = b_2 \end{cases}$, which can also be written as: $Ax = b$

($A = \begin{bmatrix} 9 \\ 5 \end{bmatrix}$), can be solved with exact solution only when b_1 and b_2 are in the ratio of 9:5,

which means that the exact solution x exists only when b is on the same line as the vector

A (such as $\begin{bmatrix} 9 \\ 5 \end{bmatrix}$) drawn from the origin. If b is $\begin{bmatrix} 4 \\ 6 \end{bmatrix}$, it is clear that no exact solution x can fit both of these two equations simultaneously. Seemingly unsolvable equations such as this are often encountered in practice from modeling experimental results, and must still be solved. In this case, an ‘approximate’ solution is to be sought, such that A times the approximate solution $\hat{x}$, would give a $\hat{b}$ that is ‘closest’ to the given b . For the example given above, from Figure 2.8, one can see that the distance between the given b vector and $\hat{b}$, the projection of b vector onto the A vector is the shortest among all the routes from b to the line A . This implies that the error vector $(b - \hat{b})$ for the optimal $\hat{b}$ must be perpendicular to the A vector. Since $A\hat{x} = \hat{b}$, we require the inner product between the A vector and the error vector to be 0:

$$A^T(b - \hat{b}) = A^T(b - A\hat{x}) = 0 \Rightarrow A^T A \hat{x} = A^T b. \quad (2.1)$$

The transpose of the matrix A is denoted by A^T . The solution is given by $\hat{x} = (A^T A)^{-1} A^T b$

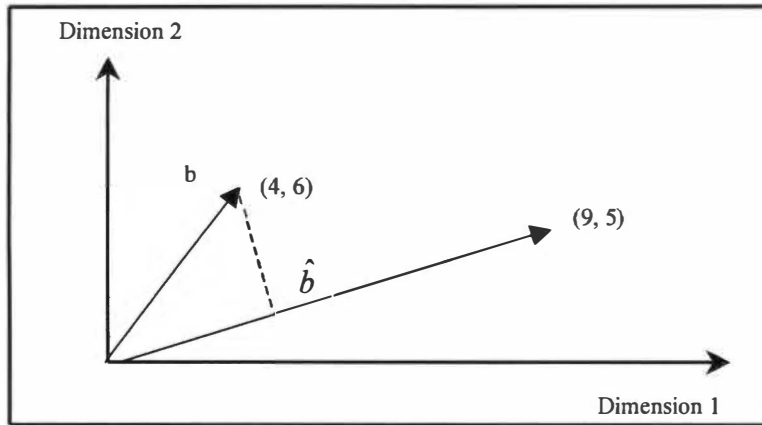


Figure 2.8: Projection of b vector (4, 6) onto A vector (9, 5). The error vector $(b - \hat{b})$ is perpendicular to the A vector and gives the shortest distance from b to A .

if $(A^T A)^{-1}$ exists. If not, the pseudoinverse of A must be used (Section 2.3.2.). In a single variable case, $\hat{x}$ is given by $\hat{x} = A^T b / (A^T A)$ because $A^T A$ is a scalar. For the example above, the $\hat{x}$ is given by:

$$\hat{x} = A^T b / A^T A = \frac{9 * b_1 + 5 * b_2}{9 * 9 + 5 * 5}.$$

Another approach to arrive at the optimum $\hat{x}$ minimizes the sum of the squares of the fitting error between the calculated $\hat{b}$ and the measured data b using calculus. For the example above, the error can be expressed as:

$$E^2 = (9 \hat{x} - b_1)^2 + (5 \hat{x} - b_2)^2.$$

The minimum error E^2 would be zero, if an x can be found to fit both of these two equations exactly. In most cases, the vectors b and a are not in the same line, and the value of E^2 would not be equal to zero. E^2 is expressed as a function of $\hat{x}$, and the minimum value of E^2 can be obtained by taking the derivative of E^2 with respect to $\hat{x}$, i.e. $d(E^2)/d\hat{x}$, and setting it to 0, followed by solving for $\hat{x}$. $\hat{x}$ is found to be:

$$\hat{x} = \frac{9 * b_1 + 5 * b_2}{9 * 9 + 5 * 5}.$$

The $\hat{x}$ that minimizes E^2 is referred to as the least-square solution to the inconsistent system $Ax = b$. The solution $\hat{x}$, arrived by the two approaches, is exactly the same.

2.3.2. Multivariable Least-Square Solution to Inconsistent Linear System of Equations, With More Equations Than Unknowns

The multivariable system of equations, $Ax = b$, which has n variables and m observations, is solvable only when the vector b is a linear combination of the columns of

A . Figure 2.9 shows an example where the vector b is in the column space of A (a 3 by 2 matrix). In this case, the multivariable equations $Ax = b$ is solvable. When n is less than m , it is very possible that $Ax = b$ is inconsistent, which means that the given vector b does not lie in the space spanned by the columns of A , as shown in Figure 2.10. In this case, an exact solution x , to the multivariable systems of equations $Ax = b$, does not exist. However, an approximate solution can be obtained by choosing $\hat{x}$, such that $A\hat{x}$ comes ‘closest’ to the given vector b . In other words, $\hat{x}$ would minimize the distance between the given b vector, and the fitted point of $(A\hat{x})$, which is expressed as:

$$E^2 = ||A\hat{x} - b||^2. \quad (2.2)$$

Here, $||\cdot||$ denotes the norm of a vector (or the magnitude of a vector), and can be defined as the sum of squares of the elements in a vector. This least-square approach is widely accepted as a convenient way for multivariable data fitting [16]. From Figure 2.10, one can see that the perpendicular distance between the vectors b and p , which is the projection of b onto the column space of A , is the minimum of all distances from b to the plane spanned by the two columns of A . In this case, the result of $A\hat{x}$ gives the projection of the vector b onto the column space of A , or the point p , which is indeed closest to b of all points on the plane. The vector $(b - A\hat{x})$ is the error vector e and is perpendicular to the column space of A . This implies:

$$A^T(b - A\hat{x}) = 0. \quad (2.3)$$

So, the least-square solution, $\hat{x}$, of $Ax = b$ is one that satisfies the so-called ‘normal equation’:

$$A^T A \hat{x} = A^T b. \quad (2.4)$$

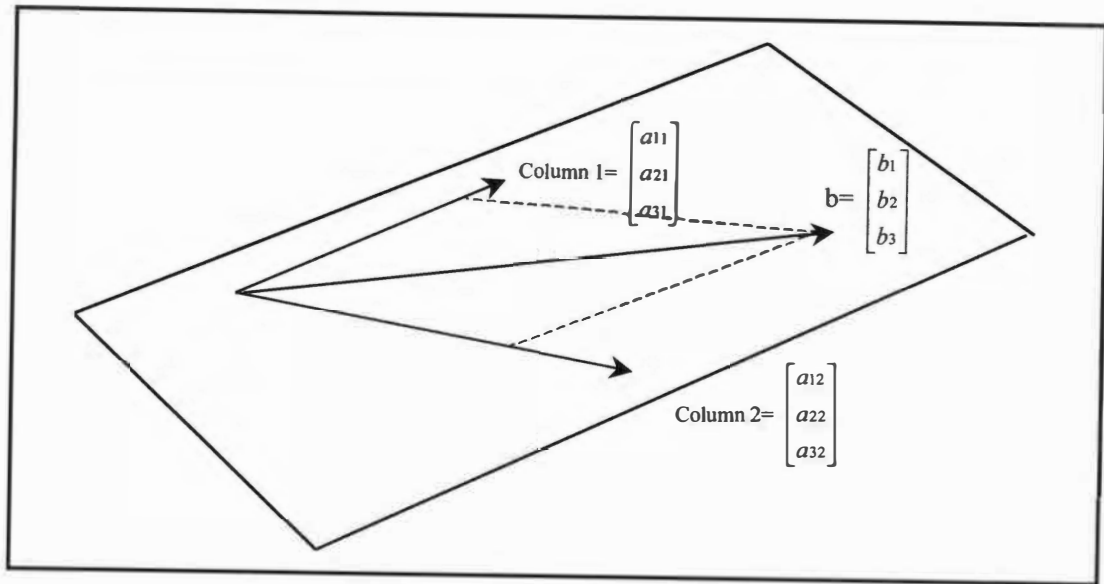


Figure 2.9: Vector b is on the column space of A (3×2 matrix). Therefore, b can be expressed as the linear combination of columns of A .

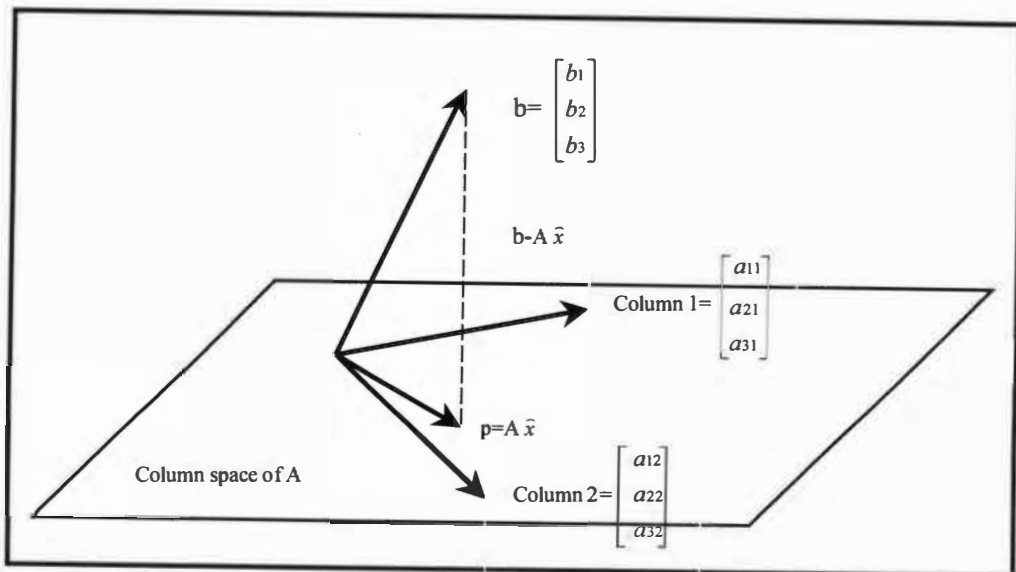


Figure 2.10: Projection of b vector onto the column space of a 3 by 2 matrix A . p is the projection of b vector onto the column space of A .

Source of ideas for illustration: G. Strang, "Linear Algebra and Its Applications (Third edition)," Harcourt Brace Jovanovich College Publishers, 1986.

If the columns of A are linearly independent, then $(A^T A)$ is invertible, and $(A^T A)^{-1}$ exists, and the least-square solution for such equation is given by:

$$\hat{x} = (A^T A)^{-1} A^T b. \quad (2.5)$$

If the columns of A are linearly dependent and $(A^T A)$ is not invertible, then the best-fit $\hat{x}$ is to be computed by the pseudoinverse of x , or x^+ :

$$\hat{x} = A^+ b, \quad (2.6)$$

where A^+ represents the pseudoinverse of the matrix A , which is to be computed using the singular-value-decomposition (SVD) of A [16]:

$$A = U \Sigma V', \quad (2.7)$$

where U (m by m matrix) and V (n by n matrix) are orthogonal matrices, and Σ (m by n matrix) is a diagonal matrix with non-negative diagonal elements arranged in a decreasing order. The diagonal elements of Σ , $\sigma_1, \sigma_2, \dots, \sigma_n$, are called the singular values of A , and the rank of A is determined by the number of nonzero σ 's. The pseudoinverse of A is computed as:

$$A^+ = V \Sigma^+ U', \quad (2.8)$$

where Σ^+ is an n by m matrix with $1/\sigma_1, 1/\sigma_2, \dots, 1/\sigma_r$ on its main diagonal, where r denotes the number of nonzero σ 's; the remaining diagonal elements of Σ^+ , corresponding to those 0 σ 's in Σ , are also 0. This is one of the hallmarks of the pseudoinverse which enables the least-square solution of an inconsistent system of linear equation to be obtained in one direct step.

If the columns of A are linearly independent, we can also calculate the least-

square solution by Equation 2.6, which will give the same result as that in Equation 2.5:

$$\hat{x} = (A^T A)^{-1} A^T b = (V \Sigma^T \Sigma V)^{-1} V \Sigma^T U^T b = V (\Sigma^T \Sigma)^{-1} V^T V \Sigma^T U^T b = V \Sigma^+ U^T b = A^+ b.$$

If A is a square and full-rank matrix, implying A is invertible, then the equation $Ax = b$ is consistent with the exact solution, x , which is given by:

$$x = A^{-1} b.$$

In this case, U , V , and Σ are all square matrices, and Σ is invertible, and the inverse of A can be calculated as:

$$A^{-1} = (U \Sigma V)^{-1} = V \Sigma^+ U^T,$$

which is exactly the same as A^+ . Therefore, for all multivariable systems of equations $Ax = b$, the least-square solution, $\hat{x}$, can always be calculated from $\hat{x} = A^+ b$ using the SVD.

CHAPTER 3

PRIOR EFFORTS TO RESOLVE MIXTURE DNA SAMPLES USING ALLELE PEAK DATA

During recent years, literature reports emerged on different groups' attempts to interpret DNA mixtures. When a DNA sample is presented as forensic evidence, generally two questions need to be answered in order to resolve it, as presented in [3], [5], and [10]:

1. What is the approximate mass ratio of the two contributors in the DNA mixture?
2. What is the genotype of each contributor?

3.1. Estimation of the Mass Ratio in a DNA Mixture Sample

The mass ratio in the mixed DNA sample can lead to the determination of each contributor's genotype [3]. Therefore, determination of the mass ratio is prerequisite to individual identification of a possible suspect. At a single locus, a two-people mixture can exhibit one, two, three, or four allele peaks leading to all possible genotype combinations as listed in Table 3.1. Figure 3.1 shows an example of the allele peaks at 10 loci of a DNA mixture profile contributed by a male and a female. From Figure 3.1, one can see that there are four alleles present at loci FGA and D21S11. Loci D3S1358, VWA, D8S1179, D18S51, D5S818, and D13S317 have three alleles each. Two alleles were detected at locus D7S820.

Table 3.1: All allele assignment patterns for a two-people DNA mixture.

Allele Type	Genotype Combination	Examples	Comment
Four Alleles [A, B, C, D]	heterozygote + heterozygote	[A, B]+[C, D]	No overlapping allele
Three Alleles [A, B, C]	heterozygote + heterozygote	[A, C]+[B, C]	One overlapping allele
	heterozygote + homozygote	[A, A]+[B, C]	No overlapping allele
Two Alleles [A, B]	heterozygote + heterozygote	[A, B]+[A, B]	Two overlapping alleles
	heterozygote + homozygote	[A, B]+[A, A]	One overlapping allele
	homozygote + homozygote	[A, A]+[B, B]	No overlapping allele
One Allele [A]	homozygote + homozygote	[A, A]+[A, A]	One overlapping allele

Source of information: J. Butler, "Forensic DNA Typing," Academic Press, 2001.

Since the allele peak area is proportional to the mass of the corresponding allele in the DNA mixture sample, it can be used to estimate the ratio of the mass portions contributed by each person in a mixed sample [3]. For a locus with four detected alleles, each contributor has to have two different alleles with no shared allele between the two contributors. Therefore, only one allele assignment structure is possible (two heterozygotes). For this reason, the loci with four alleles were usually used first by DNA analysts to estimate the mass ratio between the two contributors [5]. Since the mass ratio present in the original mixture is approximately preserved during PCR amplification [10], it should be more or less similar for all loci. Even when gross asymmetry in the two pairs of allele peaks is evident, it is still relatively easy to separate the four alleles into that of the major and minor contributor [5], as long as the mass ratio is not close to 1:1. For other loci with only two or three alleles, it is more complicated in that there is more than one possible allele assignment structure at each locus. So, an initial-guess mass ratio derived

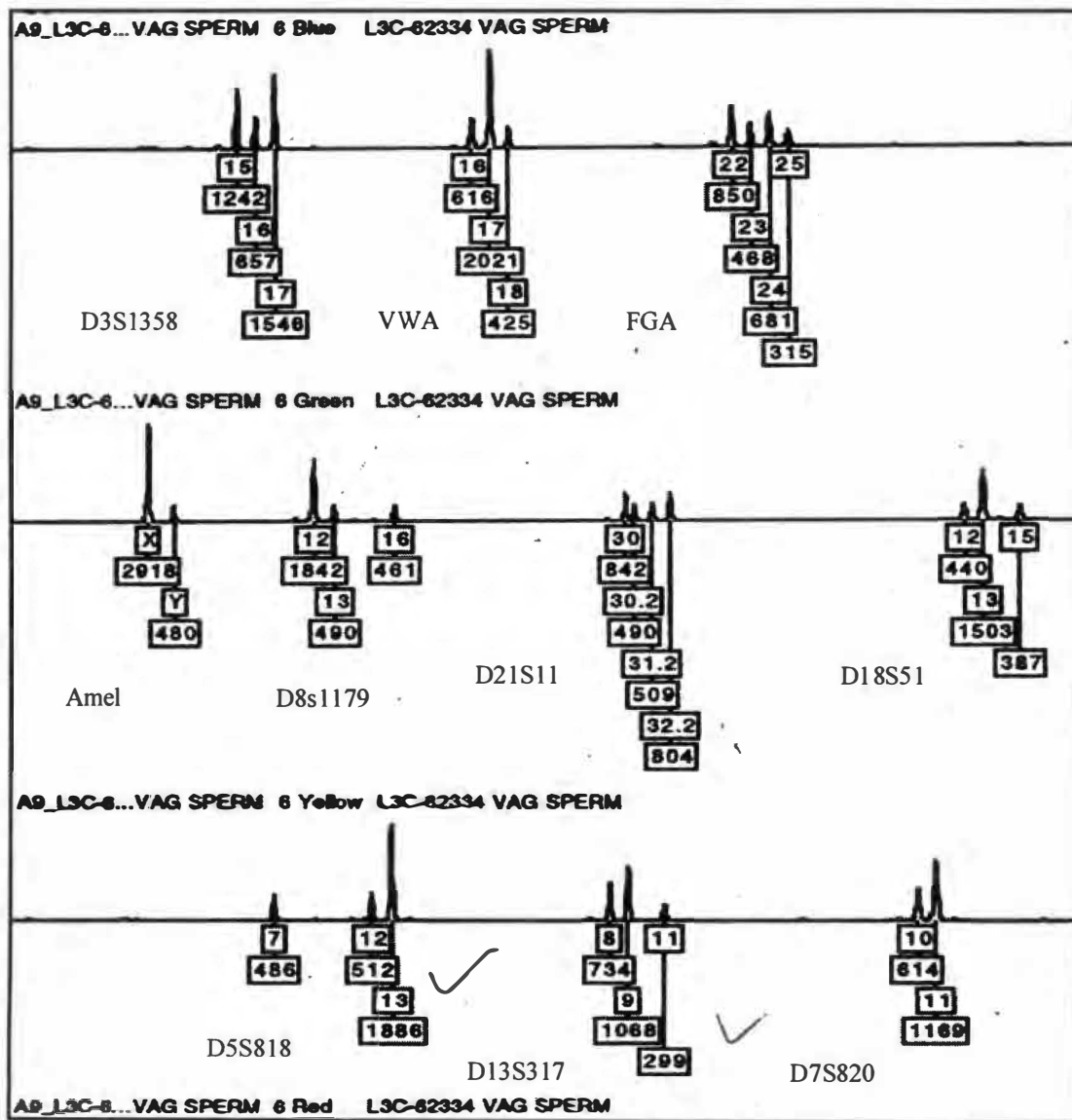


Figure 3.1: Typical allele peaks at ten loci from a DNA mixture of a male and a female.

Source: P. Graham "Allele Peak Data for DNA Mixture samples from real cases," Texas Department of Public Safety, Corpus Christi, 2001, Personal Communication.

from the four-allele loci can be used to estimate and evaluate all the possible allele assignment combinations to find out which structure is the most likely one to fit the observed relative allele peak areas.

Estimation of the mass ratio using loci with four detected alleles was first reported in [5]. The following is a brief review of the methodology presented in [5]. In that method, each of the four alleles is labeled 'A', 'B', 'C', or 'D' in increasing molecular weight order. The two alleles with larger peak areas are designated to be from one person, and the other two alleles are designated to be from the other person. Person 1 is designated to have alleles 'A' and 'B' and person 2 is designated with alleles 'C' and 'D'. The mass proportion of the DNA sample that comes from person 1 in the mixture is designated as ' Mx ', which is to be calculated from Equation 3.1:

$$Mx = \frac{PaA + PaB}{PaA + PaB + PaC + PaD}, \quad (3.1)$$

where PaA , PaB , PaC , and PaD denote the peak areas of the four alleles, respectively.

For the locus with three alleles, if person 1 has alleles 'A' and 'B' and person 2 has alleles 'B' and 'C', Mx is theoretically calculated as:

$$Mx = \frac{PaA + PaB1}{PaA + PaB + PaC}, \quad (3.2)$$

where $PaB1$ denotes the peak area of allele B contributed by person 1. In all cases, the peak area of a shared allele, such as the 'B' allele contributed by person 1 alone, cannot be measured directly from that of a combined peak. As a result, the mass proportion of

the DNA sample for person 1 or 2 cannot be obtained directly by knowing the peak areas of the three alleles detected at this locus.

If a DNA mixture sample is contributed by a male and a female, the mass ratio or proportion can be estimated from the allele peak areas at the Amelogenus gene locus [3]. The female's genotype is [XX] and the male's genotype is [XY]. Assuming the mass ratio of the female to the male's DNA is $MR:1$, then the ratio of the X allele peak area to that of Y is $(2MR+1):1$ or $r:1$ ($r=2MR+1$), where r denotes the peak area ratio of the X allele to Y allele. Therefore, the mass ratio of female to male DNA in the mixture is found to be:

$$(MR = \frac{r-1}{2}). \quad (3.4) [17]$$

If the peak areas of the Amelogenus alleles are not available or the two contributors are both female, or both male, then it is relatively hard to get an initial estimate of the mass ratio in the DNA mixture sample. A mass ratio estimated from a 4-allele loci, if available, would have to be employed to serve as a starting point to resolve other 2-allele and 3-allele loci of the same sample. Currently, resolving either a 2-allele or 3-allele locus into its individual contributors is performed using this approach of cycling through all possible genotype assignment patterns using the estimated mass ratio derived from 4-allele loci, to see which pattern fits the best to the observed allele peak areas. This procedure becomes a labor-intensive process, yielding a very conservative resolution result, at best.

The methodology of LSD, reported here, bypasses the need for an initial estimate of the mass ratio for the mixture sample before resolving it. In LSD, the most fitting genotype combination for the two contributors and the best-fit mass ratio are to be obtained in one step directly. The complete method of LSD will be detailed in Chapter 4.

3.2. Quantitative Approaches in Resolving Mixture DNA

The ultimate goal of resolving DNA mixtures is to identify the genotypes of the individual contributors. A couple of methods have been reported in recent literature. Earlier methods use the match probabilities and likelihood ratios approach [4, 14, 15]. These methods were not widely applicable because they are more ambiguous in nature, intended only to begin to show that some combinations of individual contributors' genotypes are more likely than others, given the observed mixed profile. Further, the statistical theory and applications behind these methods are not easy to formulate, apply, and interpret. Recently, reports presenting quantitative methods based on using the allele peak data of a mixture sample have emerged [5, 18]. A review of these approaches is now presented below.

3.2.1. A First-Reported Method to Resolve DNA Mixture

In 1998, the British group of P. Gill, *et al*, of the Forensic Science Services, presented a novel method to resolve DNA mixtures using quantitative allele peak data. Their method involves a rather long and tedious approach to realize the goal, and also involves an iterative search for the optimum mass ratio. The following is a summary of the method reported in [5].

DNA was extracted from the blood of five people, and the genotype of each of them was subsequently typed. Thirty-nine mixture samples of known content and mass ratios were prepared. Each mixture sample contained DNA from two individuals. The total mass of each sample was either 1 ng or 5 ng, and the mixture ratio was between 1:10 and 10:1 among the samples. ABD 373A automated sequencers were used to detect and analyze the alleles at six different loci of the DNA mixture after amplification: TH01, D21S11, D18S51, D8S1179, VWA, and FGA. At each locus, alleles were designated 'A', 'B', 'C', or 'D' in increasing molecular weight. The allele peak area was also recorded for each allele of every locus. The details of the data fit analysis algorithm are summarized below.

Table 3.2 shows the allele designation at six loci for a DNA mixture from individuals MT and NO, as well as the genotype for each of the two contributing people. Analysis proceeds taking each locus in turn. First, for the given number of alleles detected at the locus being analyzed, the authors of [5] list all possible genotype assignments of the two individuals as in Table 3.3 for a 2-allele locus. The expected proportions for the two allele peak areas corresponding to each possible allele assignment pattern are also listed. M_x denotes the proportion of DNA in the mixture contributed by person 1. A range of mass ratio is then established, and a discrete set of mass ratios within that range is selected for testing. The ' M_x ' that corresponds to each selected mass ratio is then calculated, and the expected proportion of each allele peak area can be

Table 3.2: Allele designation of a mixture DNA sample at six loci, as well as the true genotypes of its two contributors. MT and NO are the two individual contributors.

Locus	THO1	D21S11	D18S51	D8S1179	VWA	FGA
MT	8, 9.3	67, 70	11, 11	9, 17	16, 18	13, 15
NO	9.3, 9.3	59, 65	17, 19	9, 11	17, 19	13, 17
MT	A, B	C, D	A, A	A, C	A, C	A, B
NO	B, B	A, B	B, C	A, B	B, D	A, C
Mixture Type	A, B	A, B, C, D	A, B, C	A, B, C	A, B, C, D	A, B, C
Type of Model	2 - allele	4 - allele	3 - allele	3 - allele	4 - allele	3 - allele

Source: P. Gill, R. Sparkes, R. Pinchin, T. Clayton, J. Whitaker, and J. Buckleton, "Interpreting Simple STR Mixtures Using Allele Peak Areas," Forensic Science International, 91(1998): 41-53.

Table 3.3: The expected proportions of allele peak areas for all possible genotype combinations if two alleles were detected at a locus.

All Possible Allele Assignment Patterns		Expected Proportions of Peak Area (Proportion of A + proportion of B = 1)	
Person 1	Person 2	Proportion of A	Proportion of B
AA	AB	$(Mx/2) + 0.5$	$(1 - Mx)/2$
AB	AB	0.5	0.5
AA	BB	Mx	$1 - Mx$
AB	AA	$1 - (Mx/2)$	$Mx/2$
BB	AA	$1 - Mx$	Mx
AB	BB	$Mx/2$	$1 - Mx/2$
BB	AB	$(1 - Mx)/2$	$(Mx/2) + 0.5$

Note: Alleles A and B were observed at one locus in the DNA mixture. Mx is the proportion of the mass from person 1 to the total mass of the DNA mixture sample.

Source: P. Gill, R. Sparkes, R. Pinchin, T. Clayton, J. Whitaker, and J. Buckleton, "Interpreting Simple STR Mixtures Using Allele Peak Areas," Forensic Science International, 91(1998): 41-53.

obtained from the corresponding calculated Mx . For every mass ratio, associated with every genotype combination the residuals between the expected allele peak area proportions, and those calculated from the observed allele peaks for a locus can be calculated with Equation 3.5:

$$Residual = \sum_i (proportion\ of\ PA\ observed - proportion\ of\ PA\ expected)_i^2 \quad (3.5)$$

where ‘ i ’ denotes the i th allele of a locus. PA represents the peak area. Equation 3.5 is used to calculate the residuals for each possible genotype combination with the mass ratio varying in a selected range. An example of the calculated proportion of each allele peak area based on the mass proportion of person 1 for a 2-allele locus is listed in Table 3.3. Mx , as used in Table 3.3, is obtained from the mass ratio of person 1 to person 2 of each of the assumed values within the selected range of mass ratio. The overall residual at that locus is to be summed from the individual residual calculated for each allele. The overall residuals of the locus for all the mass ratios in the chosen range can be plotted. The smallest residual is picked out and compared with the minimum residuals associated with other genotype combinations similarly derived. There will be several curves resulting from the various genotype combinations as shown in Figure 3.2. Thus among the set of curves, the genotype combination that gives the lowest minimum-residual is selected to be the best-fit genotype combinations. Then the corresponding mass proportion of person 1 can be identified from the x-axis. Analysis then proceeds to the next locus. Assuming the preservation of mass proportion across all loci being tested, the most fitting mass fraction, identified by choosing that which gives the smallest minimum residual, should correspond, more or less, to the same mass proportion identified at other loci.

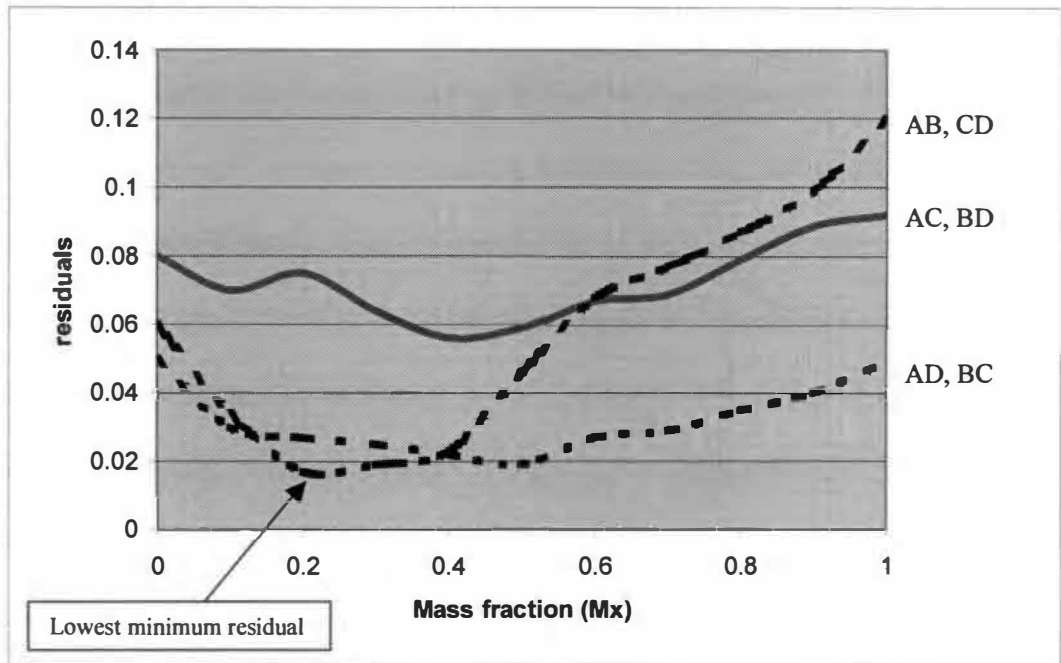


Figure 3.2: Typical curves of the residuals for all possible genotype combinations across the range of mass proportion of person 1 in the mixture for a 4-allele locus. The range of the mass proportion of person 1 in the mixture sample is from 0 to 1; curves shown in this figure are generated by us to mimic curves that would result from the computations presented in [5]. Each curve represents that of a particular genotype combination.

The method presented above will yield an estimated mass ratio, which best fits the observed peak area ratios. About 85% of the 270 loci were estimated successfully using this method. When the mass ratio in the DNA mixture is close to 1:1, several possible genotype combinations can rank optimally causing substantial difficulty in determining which genotype is a better fit than that of the others. In applying this method, if the estimated mass ratio is available beforehand, then comparing the residuals for every possible genotype combination at this fixed mass ratio can lead to the final result in one step. If the allele peak areas at a 4-allele locus suffer from a signal imbalance, perhaps due to unequal PCR amplification across alleles, then there may be considerable deviation of the estimated mass ratio from the real mass ratio. In this case, the final results of the most fitting genotype combination may be wrong. If the peak areas of the amelogenin locus are not available or no locus has 4 alleles, then the residuals for every possible genotype combination at every selected mass ratio should be calculated. Once a set of mass ratios is chosen over a range, the calculation will be repeated for each of the mass ratios in the set. If the actual mass ratio exceeds the assumed range used in the calculation indicated by an absence of a minimum in plot similar to that of Figure 3.2, then a wider range would have to be used, and the scheme has to be rerun to accommodate this wider range. As can be seen, such an approach is rather laborious, complicated, and tedious, and not powerful enough to gain popular acceptance.

3.2.2. A Recently Published Mathematical Method to Resolve DNA Mixture

In 2001, Mark Perlin and Beata Szababy introduced another method: the Linear Mixture Analysis (LMA) method that attempts to resolve DNA mixtures also using

quantitative allele peak data [18]. In the opinion of the paper's authors, the advantage of LMA is that it integrates all the quantitative allele peak data into a single rapid matrix computation [18]; therefore, the authors argue that it is more straightforward compared with other methodologies, in which each locus is analyzed separately [18]. However, in our opinion, it is this very feature that renders LMA inflexible, because LMA imposes the same mass ratio to all loci and claims it is the best fit for all. This is in direct contradiction to the observation that the best-fit mass ratio varies from locus to locus, due to unequal amplification and other nonidealities [24]. LMA is the method that most resembles our LSD method. The details and features of LMA method are now reviewed.

3.2.2.1. Linear Model of LMA

The basic formulation of LMA, according to Reference [18], is explained in this section. After PCR amplification of a DNA sample, the peak area of each allele is more or less proportional to its mass in the sample [10]; therefore, observing the relative peak area proportions of alleles at a locus of a mixed DNA sample can help the caseworkers to estimate the relative mass of the DNA of the two contributors in the mixture. For a DNA mixture contributed by two people, the mixture weight percent of person A can be defined as:

$$w_A = \frac{\text{weight of DNA sample from person } A}{\text{total weight of DNA mixture sample}}; \quad 0 < w_A < 1. \quad (3.6)$$

The linear relationship between the weights of contributors' DNA to a DNA mixture sample and the allele peak areas at all loci can be combined into a single linear matrix equation:

$$d = G \cdot w \quad (3.7)$$

where the vector d is composed of the relative allele peak areas at **all** loci to be analyzed, matrix G represents the genotypes of all contributors at **all** loci, and has as many columns as the number of contributors to the mixture, and w is a column vector of DNA weights of all contributors as defined in Equation 3.6. Note that the sum of the component of w is 1. If there are three people, A, B, C contributing to a DNA mixture, where 60 % of the DNA is from A, 20 % from individual B, and the other 20 % from individual C, then the corresponding mixture weights are:

$$w_A = 0.6, w_B = 0.2, w_C = 0.2.$$

Suppose that at two loci, containing 3 alleles and 4 alleles respectively, the genotypes of the three contributors are [18]:

- A has alleles 1 and 2 at locus 1 and has alleles 1 and 2 at locus 2;
- B has alleles 1 and 3 at locus 1 and has alleles 2 and 3 at locus 2;
- C has alleles 2 and 3 at locus 1 and has alleles 3 and 4 at locus 2.

This information can be combined into Equation 3.8:

$$d = G \cdot w$$

$$d = \begin{bmatrix} \text{Locus 1} \\ \text{mixture} \\ \text{allele} \\ \text{peak areas} \\ \hline \text{Locus 2} \\ \text{mixture} \\ \text{allele} \\ \text{peak areas} \end{bmatrix} = \begin{bmatrix} \text{Locus 1} \\ A's \\ \text{alleles} \\ \hline \text{Locus 2} \\ A's \\ \text{alleles} \end{bmatrix} \begin{bmatrix} \text{Locus 1} \\ B's \\ \text{alleles} \\ \hline \text{Locus 2} \\ B's \\ \text{alleles} \end{bmatrix} \begin{bmatrix} \text{Locus 1} \\ C's \\ \text{alleles} \\ \hline \text{Locus 2} \\ C's \\ \text{alleles} \end{bmatrix} \cdot \begin{bmatrix} w_A \\ w_B \\ w_C \end{bmatrix} = \begin{bmatrix} g^A & g^B & g^C \end{bmatrix} \cdot \begin{bmatrix} w_A \\ w_B \\ w_C \end{bmatrix} \quad (3.8)$$

For each locus, the sum of allele peak data in each column vector of G and d is to be

normalized to two. For the DNA mixture sample simulated above, Equation 3.8 can be written as:

$$\begin{bmatrix} 0.80 \\ 0.80 \\ 0.40 \\ \vdots \\ 0.60 \\ 0.80 \\ 0.40 \\ 0.20 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ - & - & - \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0.60 \\ 0.20 \\ 0.20 \end{bmatrix}. \quad (3.9)$$

Equation 3.8 is a linear mixture model for three individuals' DNA mixture at two loci. If allele peak areas are available at all thirteen loci, then all the peak data will be compiled into a linear model similar to Equation 3.9. The number of rows of the matrix G is equal to the total number of alleles at all thirteen loci of the mixture sample. Note that a prominent feature of the LMA formulation is that the **same** set of relative mixture weights ($\{w_A, w_B, w_C\}$) is imposed for DNA at all thirteen loci of the mixture.

3.2.2.2. Problem Formulation and Solution Approach of LMA

In order to use Equation 3.8, some information must be known, *a priori*, then other elements can be computed using the equation $d = G * w$. As presented in [18], resolving DNA mixtures using LMA includes four steps:

- Obtaining the allele peak data of a DNA profile contributed by multiple individuals.
- Integrating all the allele peak data of this mixture sample into the vector d in the linear model shown in Equation 3.8.

- Depending on which part of Equation 3.8 is known, solve the linear equation formulated in the previous step for the remaining pieces.
- Resolving the DNA mixture from the solution of the linear equation.

If the genotypes of all contributors (G) are known or assumed, and the allele peak data (d) are also known, the authors of [18] show that the mixture weights are obtained by solving the linear equation $d = G \cdot w$ for w using the least-square method:

$$w = (G^T \cdot G)^{-1} \cdot G^T \cdot d. \quad (3.10)$$

Assume the matrix G is full rank, and therefore $(G^T \cdot G)^{-1}$ exists. For DNA mixture from two individuals A and B, where one of the two individuals' genotypes is known, if the allele peak data d and the mixture weights w are also known or assumed known, the genotype of the other person, B, can then be calculated. In this case, the equation $d = G \cdot w$ can be expanded as:

$$d = w_A \cdot g_A + w_B \cdot g_B. \quad (3.11)$$

The genotype of individual B can be obtained easily by solving g_B from Equation 3.11 and rounding the elements of g_B to the nearest integers:

$$g_B = (d - w_A \cdot g_A) / w_B = (d - w_A \cdot g_A) / (1 - w_A). \quad (3.12)$$

However, for most DNA mixture samples, the mixture weights w are not known beforehand to the caseworkers. In this case, the problem of calculating the genotype of person B and the mixture weights w will be more complicated. The authors of [18] suggest an iteration approach in searching for the optimal w_A , that fits the observed peak areas the best. Figure 3.3 is the flow sheet of LMA to resolve a DNA mixture contributed

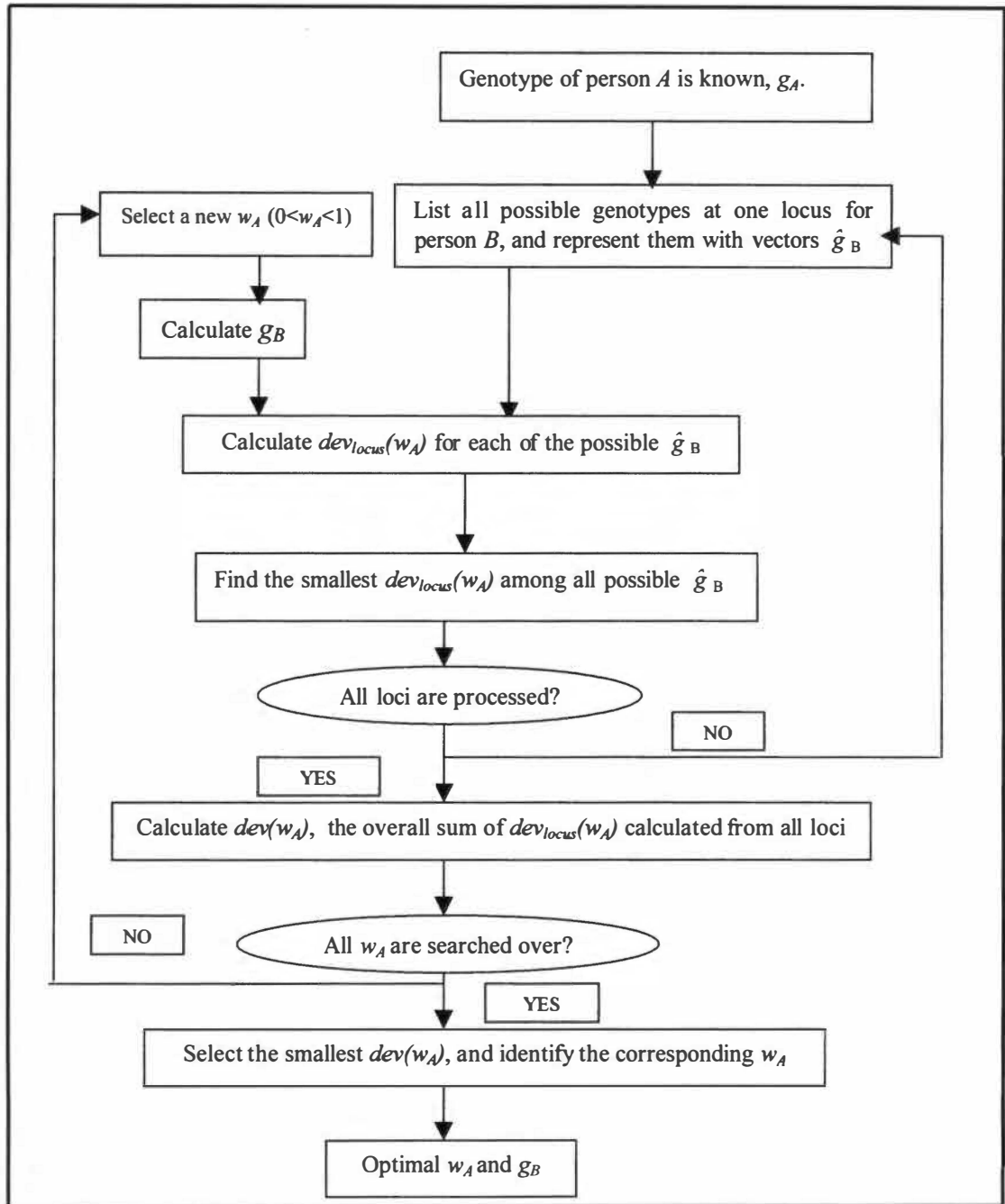


Figure 3.3: Flow sheet of LMA to resolve two-people DNA mixture; the genotype of one of the two people (designated as person A here) has to be known [18].

by two individuals based on their proposed iterative search approach. The overall procedures are summarized as follows:

- The best weight w_A is searched over the interval $[0, 1]$. A discrete set of weights within this range is chosen for testing. Since the genotype of person B can be expressed as a function of w_A , for every chosen w_A , one can compute the genotype g_B from Equation 3.12.
- Choose a w_A , and at each locus, all possible genotypes for person B will be listed and represented by a vector v . For every possible genotype, the deviation, which is the sum of squares of the difference between the calculated g_B and the corresponding v , is calculated. For each locus at the indicated w_A , the smallest deviation, derived from all the possible genotypes, is selected and recorded as $dev_{locus}(g(w))$.
- To compute the overall $dev(g(w))$ for the entire profile, at the same chosen w_A , one needs to sum the individual $dev_{locus}(g(w))$ at each locus for this DNA mixture sample, which can be expressed as:

$$dev(g(w)) = \sum_{loci} dev_{locus}(g(w)). \quad (3.13)$$

- Another w_A is selected, and a new $dev(g(w))$ is computed, until all $\{w_A\}$ have been processed.
- By plotting $dev(g(w))$ against the set of chosen w_A 's, the w_A that gives the smallest deviation can be identified, such as from the graph shown in Figure 3.4. After obtaining the optimal w_A , the corresponding genotype vector g_B can be identified and isolated. Person B's exact genotype can be obtained by rounding the elements of vector g_B to the nearest integers.

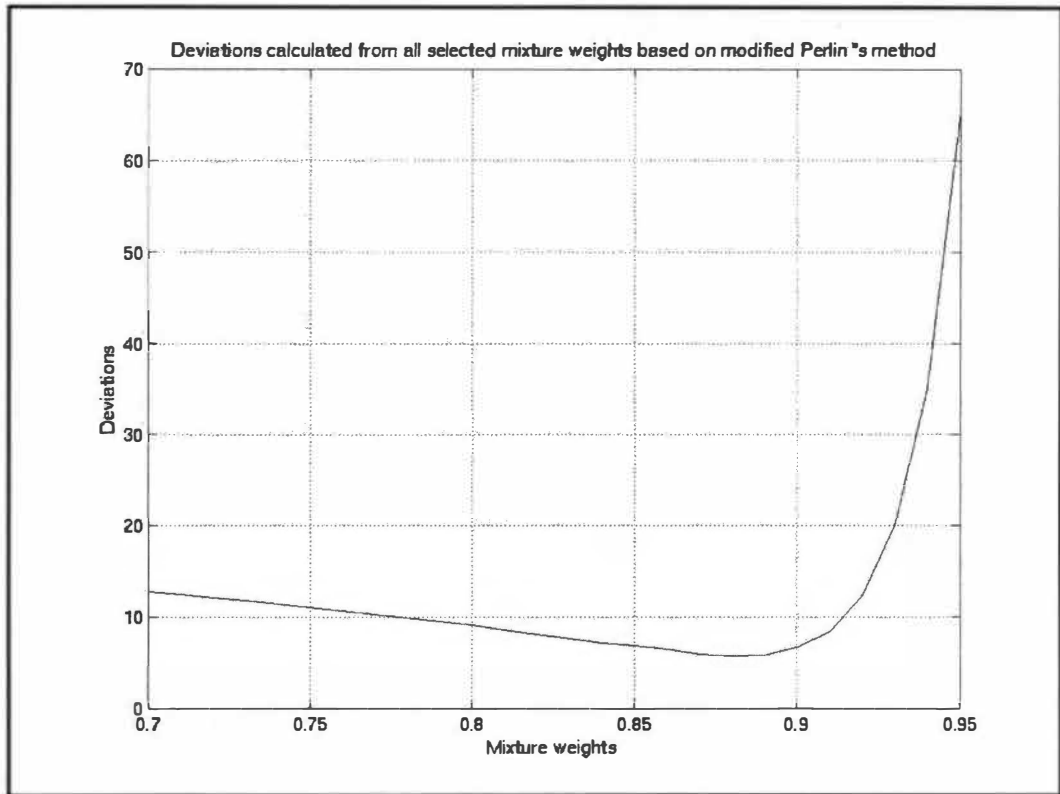


Figure 3.4: The calculated squared deviation $dev(g(w))$ against the mixture weight w_A for a DNA mixture sample with the mass ratio of 9:1. The mass fraction of 88% is identified as the optimal mass proportion of person A. Such result is consistent with the known true value of 90%.

As implied in [18], for resolving DNA mixtures with three contributors, the genotypes of two of the three contributors have to be known, *a priori*, in order to apply LMA to arrive at the genotype of the third contributor. In this case, the search for the optimal weights is over a two-dimensional grid, and the optimal weight vector is $[w_A, w_B, 1 - (w_A + w_B)]$. The genotype of the third person, g_C , is to be calculated by rearranging Equation 3.14:

$$d = w_A \cdot g_A + w_B \cdot g_B + (1 - w_A - w_B) \cdot g_C \quad (3.14)$$

That is:

$$g_C = d - (w_A \cdot g_A + w_B \cdot g_B) / (1 - w_A - w_B).$$

The authors state that LMA can be used to resolve three-people DNA mixtures as mentioned above, but they did not provide a sample calculation to illustrate the algorithm. It appears that requiring the genotype information of all contributors but one of a multi-people mixture in order to apply LMA is highly impractical in forensic cases, thus rendering LMA for resolving multi-people mixture a theoretical treatise at best.

In Chapter 5, results of using LMA and LSD on several selected DNA mixture samples will be presented. It is predicted that the imposition of the same weight fractions to fit all loci is going to present a limitation on that set of weight fractions being optimal for all loci; in the LSD algorithm, a separate optimal weight fraction is allowed to be computed for each locus, independent of all other loci.

CHAPTER 4

LEAST-SQUARE PROBLEM FORMULATION AND SOLUTION APPROACH TO RESOLVE TWO-PEOPLE STR/DNA MIXTURE PROFILES

DNA mixture samples are very common in forensic cases, especially in rape cases. But until now, no direct, systematic, analytic, and quantitative method exists to resolve DNA mixture samples. Some quantitative methods using allele peak data have emerged recently as reviewed in Chapter 3. Their various shortcomings have also been reviewed. In this chapter, the formulation of a new method, the Least-Square Deconvolution (LSD), based on allele peak data, is presented to resolve DNA mixture samples contributed by two individuals. Results of using LSD to resolve real DNA mixtures will be presented in Chapter 5.

4.1. Least-Square Deconvolution (LSD), an Overview

LSD is a novel method applying the least-square modeling approach to find the best-fit genotype combination to resolve 2-people DNA mixture samples. Since the allele peak areas are theoretically proportional to the mass of the corresponding DNA allele in the mixture and such relationship is preserved during PCR amplification, the allele peak data was used in LSD to resolve DNA mixture samples. Throughout this report, the peak area will be used, but peak height can also be used in place of peak area. The objective of LSD is to find the best-fit genotype combination for the two contributors at all loci where peak data is available, using least-square techniques and the measured allele peak data.

The advantage of LSD, compared to other quantitative approaches published in the literature, is its direct calculation of the best-fit genotype combination and the approximate mass ratio, without iterative searching for the optimal mass ratio, as introduced in references [5] and [18]. Some matrix calculations are involved in this method. Figure 4.1 is the flow sheet of LSD. From the flow sheet, one can see that LSD is applied at each locus independently of other loci, thus allowing each locus to be fitted with its own best-fit mass ratio, which can differ from the best-fit mass ratio arrived at other loci. The method of LSD includes four steps, each of which will be presented in the following sections.

4.1.1. Representing All Possible Genotype Combinations at Each Locus with Matrices

The allele peak data of a DNA mixture sample, when available, usually includes the allele peak information from 6 to 18 loci. One locus is to be analyzed at a time. For each locus, the number of alleles can be determined by observing all the true peaks in the electropherogram at this locus [3]. According to the number of true alleles detected, all possible allele assignment patterns for the two contributors are to be first enumerated. For each possible allele combination pattern for the two contributors, a matrix of coefficients with two columns, one for each person, can be formed. This matrix is referred to as the ‘ A ’ matrix, such as those shown in Table 4.1. Each column of A represents the genotype of one contributor at this locus. The element in the matrix A is either 0, 1, or 2, representing the absence or presence of the corresponding allele. A ‘2’ stands for the homozygote allele indicating that the contributor has two copies of this allele at the

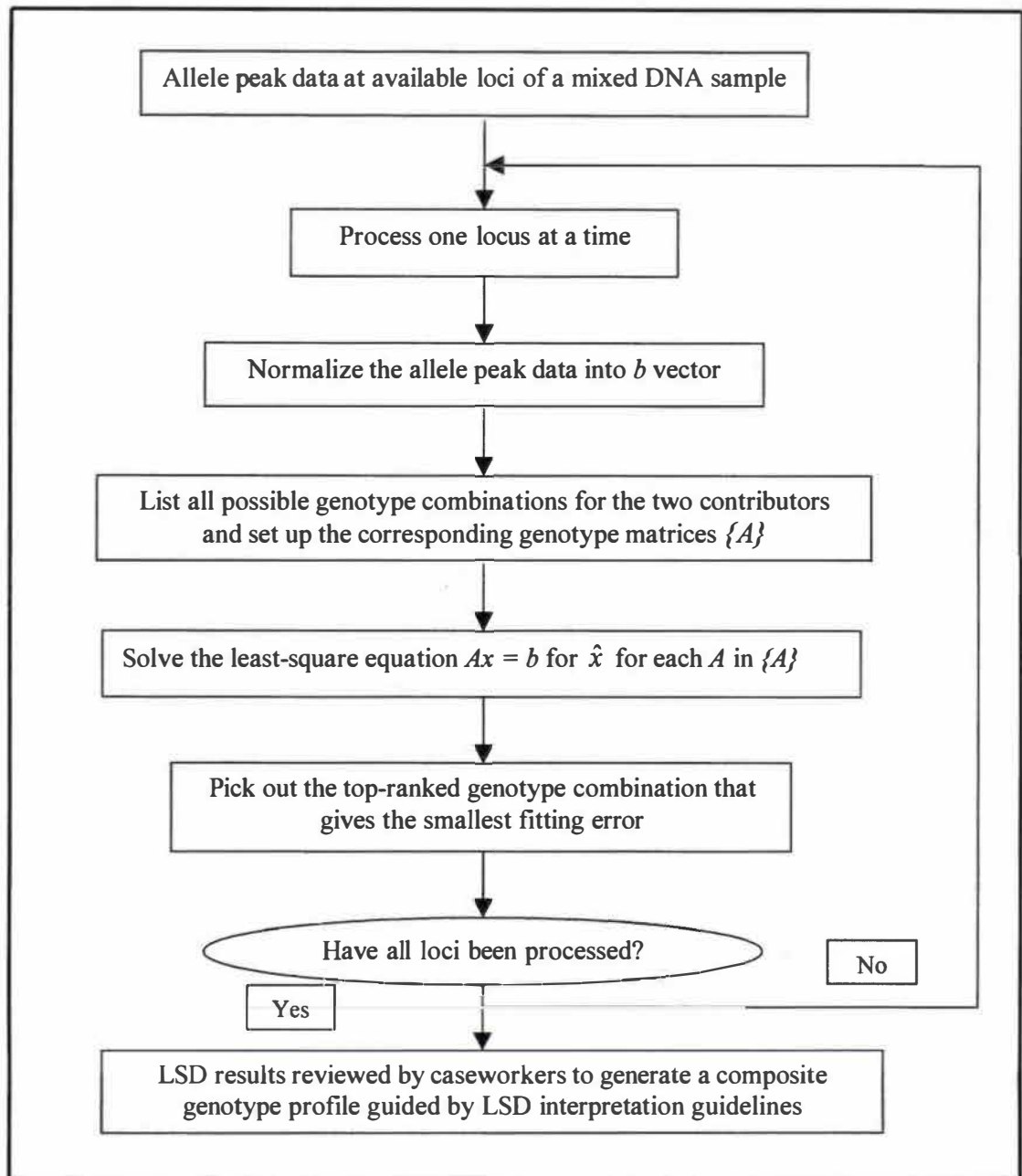


Figure 4.1: Flow sheet of LSD to resolve mixture DNA samples. For each locus, the best fitting genotypes for the two contributors and the approximate mass ratio can be obtained in one step.

Table 4.1: All possible genotype combinations and their corresponding matrices at a three-allele locus for the two contributors. The three alleles are designated as {A, B, C}. The numbers {0, 1, 2} in the matrix A represents 0, 1, or 2 copies of the corresponding allele it represents.

Case	Genotype Combination		Matrix (A)	Pseudoinverse of the Matrix (A^+)
	Person 1	Person 2		
1	A, A	B, C	$\begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$
2	B, B	A, C	$\begin{bmatrix} 0 & 1 \\ 2 & 0 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$
3	C, C	A, B	$\begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 2 & 1 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$
4	A, B	B, C	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{3} \begin{bmatrix} 2 & 1 & -1 \\ -1 & 1 & 2 \end{bmatrix}$
5	A, B	A, C	$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{3} \begin{bmatrix} 1 & 2 & -1 \\ 1 & -1 & 2 \end{bmatrix}$
6	A, C	B, C	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$	$\frac{1}{3} \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \end{bmatrix}$

specified locus. The number of rows of the matrix A corresponds to the number of alleles observed at this locus in the mixture sample.

For a DNA mixture contributed by two people, the number of alleles that can be present at one locus is between one and four. If only one allele is present at a certain locus, then undoubtedly both contributors are homozygous with this allele at this locus. In Tables 4.1, 4.2, and 4.3, all possible genotype combinations for the two source contributors and their corresponding ' A ' matrices are listed for the cases of 2, 3, and 4 alleles present at a locus, respectively. The last column in each of these three tables lists the pseudoinverse matrix of the corresponding genotype matrix ' A '. The pseudoinverse matrix will be used to compute directly the best-fit mass coefficients during the least-square fitting algorithm as in Equation 2.6. This step will be further elaborated in the next section.

A normalized vector ' b ' is to be generated from the measured allele peak area data, such that the smallest allele peak area in the vector is normalized to 1, and those of all other alleles in the same locus are ratioed to it. The number of elements in the vector ' b ' reflects the number of alleles present at this locus. For example, if the peak area measured at one locus is [3454, 4588, 8766], then the corresponding normalized b vector would be:

$$b = \begin{bmatrix} 3454 / 3454 \\ 4588 / 3454 \\ 8766 / 3454 \end{bmatrix} = \begin{bmatrix} 1 \\ 1.33 \\ 2.54 \end{bmatrix}. \quad (4.1)$$

Table 4.2: All possible genotype combinations and their corresponding matrices at a two-allele locus for the two contributors. The two alleles are designated as {A, B}. The numbers {0, 1, 2} in the matrix A represents 0, 1, or 2 copies of the corresponding allele it represents.

Case	Genotype Combination		Matrix (A)	Pseudoinverse of the Matrix ($A^\dagger$)
	Person 1	Person 2		
1	A, A	B, B	$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
2	A, B	A, A	$\begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 0 & 2 \\ 1 & -1 \end{bmatrix}$
3	A, B	B, B	$\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 2 & 0 \\ -1 & 1 \end{bmatrix}$
4	A, B	A, B	$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$	$\frac{1}{4} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

Table 4.3: All possible genotype combinations and their corresponding matrices at a four-allele locus for the two contributors. The four alleles are designated as {A, B, C, D}. The numbers {0, 1} in the matrix A represents 0 or 1 copy of the corresponding allele it represents.

Case	Genotype Combination		Matrix (A)	Pseudoinverse of the Matrix ($A^\dagger$)
	Person 1	Person 2		
1	A, B	C, D	$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$
2	A, C	B, D	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$
3	A, D	B, C	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$

4.1.2. Using the Least-Square Approach To Calculate the Best-Fit Mass Coefficients and the Residuals at Each Locus

The algorithm to calculate the best-fit mass coefficients and the residuals at each locus is presented in this section. The allele peak areas at each locus of a mixed DNA/STR sample are a superposition of both contributors' allele peaks, weighted by their corresponding mass proportions. For example, for a locus with three alleles, the relationship between the genotypes of the two contributors and the allele peak areas measured can be expressed as a linear equation as shown in Equation 4.2:

$$\begin{array}{ccc}
 \begin{array}{c} \text{Assumed} \\ \text{Genotype} \end{array} & & \begin{array}{c} \text{Measured} \end{array} \\
 \\
 \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} * \begin{bmatrix} \text{Mass } 1 \\ \text{Mass } 2 \end{bmatrix} = \text{Mass } 1 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + \text{Mass } 2 \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \alpha \begin{bmatrix} \text{Allele } 1 \text{ peak data} \\ \text{Allele } 2 \text{ peak data} \\ \text{Allele } 3 \text{ peak data} \end{bmatrix} \\
 \\
 A * x = b \quad (4.2)
 \end{array}$$

where 'α' is a proportionality constant. For the locus with three or four alleles, A is a three by two, or four by two matrix respectively, and the multivariable linear equation of $Ax = b$ would generally be inconsistent, with more equations than unknowns and no exact solution would exist, meaning no solution 'x' can be found such that Ax would equal to b exactly. However, the least-square solution of $Ax = b$ can be found, where x_{ls} is a two-element vector composed of the relative mass coefficients for the two contributors to the DNA mixture, such that A times x_{ls} would be 'closest' to the given vector 'b', of all possible x vectors. Computationally, the least-square solution, $\hat{x}$ to the inconsistent $Ax = b$, is given by $\hat{x} = A^+ b$, where A^+ denotes the pseudoinverse of A [16]. The solution $\hat{x}$ is such that $(A \hat{x})$ will give a fitted allele peak vector $\hat{b}$, which is not equal to b exactly, but

would be ‘closest’ to ‘ b ’. The difference between ‘ b ’ and ‘ $\hat{b}$ ’ is called the residual vector. The magnitude of the residual vector, representing how close $\hat{b}$ is to b is

$$E^2 = ||(b - A \hat{x})||^2 \quad (4.3)$$

where $||\cdot||^2$ stands for the norm of the error (or residual) vector. Thus, for each locus, each of the possible genotype matrices is fitted to the given allele peak data vector, b , and the associated residual, E^2 , is calculated and compared against each other. The genotype pair that gives the smallest fitting error is ranked the highest as the most likely genotype profiles for the two contributors, at this locus, provided the relative allele peak data, is consistent with the underneath relative DNA masses in the mixture.

4.1.3. Finding the Most Likely Genotype Combination for Each Locus

If the proportionality between the DNA mass fractions and the associated allele peak areas is preserved, then it is expected that the genotype combination that gives the smallest fitting error residual, if a unique minimum exists, corresponds to the true genotype profiles of the two contributors. Furthermore, by comparing the smallest residual to the next smallest one, one can estimate the confidence with which one can claim the genotype with the smallest fitting residual being the correct one. If the next smallest residual is very much bigger than that of the smallest one, then, one can claim with more confidence that the genotype with the smallest residual is the correct genotype combination pattern at that locus.

For a locus with two alleles, there are four possible genotype combinations for the

two contributors, as shown in Table 4.2. The 2-allele case always needs special consideration, because the matrices A of three of the four possible genotype combination cases are full rank. This means that an exact solution exists in solving for x in $Ax = b$. Therefore, a least-square solution is not involved. Consequently, this also means that the computed mass coefficient vector, x , gives an exact fit to the allele peak area vector and the fitting error is zero. However, this does not necessarily imply that one of these three genotype combination cases is the correct one. It is just that the mathematics involved in solving the problem is such that an exact fit is allowed. Therefore, it is the associated mass ratio that has to be used to determine which genotype combination case is more consistent with the best-fit mass ratios identified from other loci already processed. Two situations exist in the 2-allele case:

1. If the genotypes of the two individuals are different, as in Cases 1, 2, and 3 shown in Table 4.2, then the matrix A is full rank, meaning the two columns of A are independent. Therefore, an exact solution x for the equation $Ax = b$ can be obtained from $x = A^+b$, in this case, A^{-1} exists and is equal to A^+ as described in Chapter 3. As a result, the errors calculated from these three cases are all zero. For this reason, in this case comparing the errors does not help the caseworkers in selecting the best-fit genotype assignment for a two-allele locus. The mass ratio is more indicative of which case is the best-fit case.
2. If the genotypes for both contributors are the same and each individual is heterozygous at this locus, as in Case 4 shown in Table 4.2, then regardless of the mixture DNA's mass ratio, the fitted two allele peak areas at this locus will always be the

same. The matrix A that represents this genotype assignment is $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. The two vectors of this matrix are identical, and the rank is 1. Therefore, an exact solution for the equation $Ax = b$ can be found only when the two elements in b vector are also the same, meaning the two allele peak areas are equal to each other. Since there is always uncertainty associated with the allele peak data measurement, they are seldom equal exactly. As a result, no exact solution would be obtained, and the fitting error for this case would always be greater than 0, thus greater than those for the other genotype cases. In addition, the computed fitted mass ratio would always be 1:1, but in most cases, it would be wrong. The real mass ratio can be anything, but cannot be estimated from the least-square fit solution. Therefore, when the peak areas of the two alleles are indeed close, then the possibility exists that Case 4 is the applicable genotype combination case, regardless of the value of the fitting error and its unestimatable mass ratio.

From the above analysis, one can conclude that for loci with only two alleles, comparing the errors associated with each possible genotype combination choice does not help the caseworkers to decide the correct genotypes for the two contributors. The best-fit genotype combination should be the one with a mass ratio consistent with those estimated for other loci with three or four alleles. However, if the ratio of the two allele peak areas is close to 1:1, the genotype combination Case 4 represented by the matrix $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ should then be considered possible, even though the error for this case is greater than those for the other cases.

In general, the method to identify the most likely genotype combinations at each locus in LSD, is summarized as below:

- For a locus with three or four alleles, the genotype combination with the smallest fitting residual should first be chosen as the most likely one. However, consistency of mass ratios is important too.
- Figure 4.2 shows the flow sheet of using LSD to resolve DNA mixture at a 2-allele locus. For a locus with two alleles, if the ratio of the two allele peak areas is not close to 1:1, the genotype combination with the fitted mass ratio consistent with those for other loci should be selected as the most likely one among the first three cases shown in Table 4.2. If the ratio of the two allele peak areas is close to 1:1 and the approximate mass ratio of the DNA mixture, derived from other loci, is not close to 1:1, the genotype combination in Case 4 shown in Table 4.2 {[A, B][A, B]} should then be considered as the only possibility, even though the error for this case is greater than those for the other cases. If the ratio of the two allele peak areas is close to 1:1 and the approximate mass ratio of the DNA mixture is also close to 1:1, then there are two possible genotype combinations {[A, A][B, B]} and {[A, B][A, B]}, both of which can fit equally well to the measured allele peak areas.

4.1.4. Reviewing the Ranking Result and Forming the Composite Profiles

From pooling together the most likely genotype combinations at all loci, the best-fit genotype profiles can be established for the two source contributors. Since the best-fit mass ratio varies somewhat from locus to locus, an approximate mass ratio of the DNA mixture sample can be obtained by averaging across the individual mass ratios from each

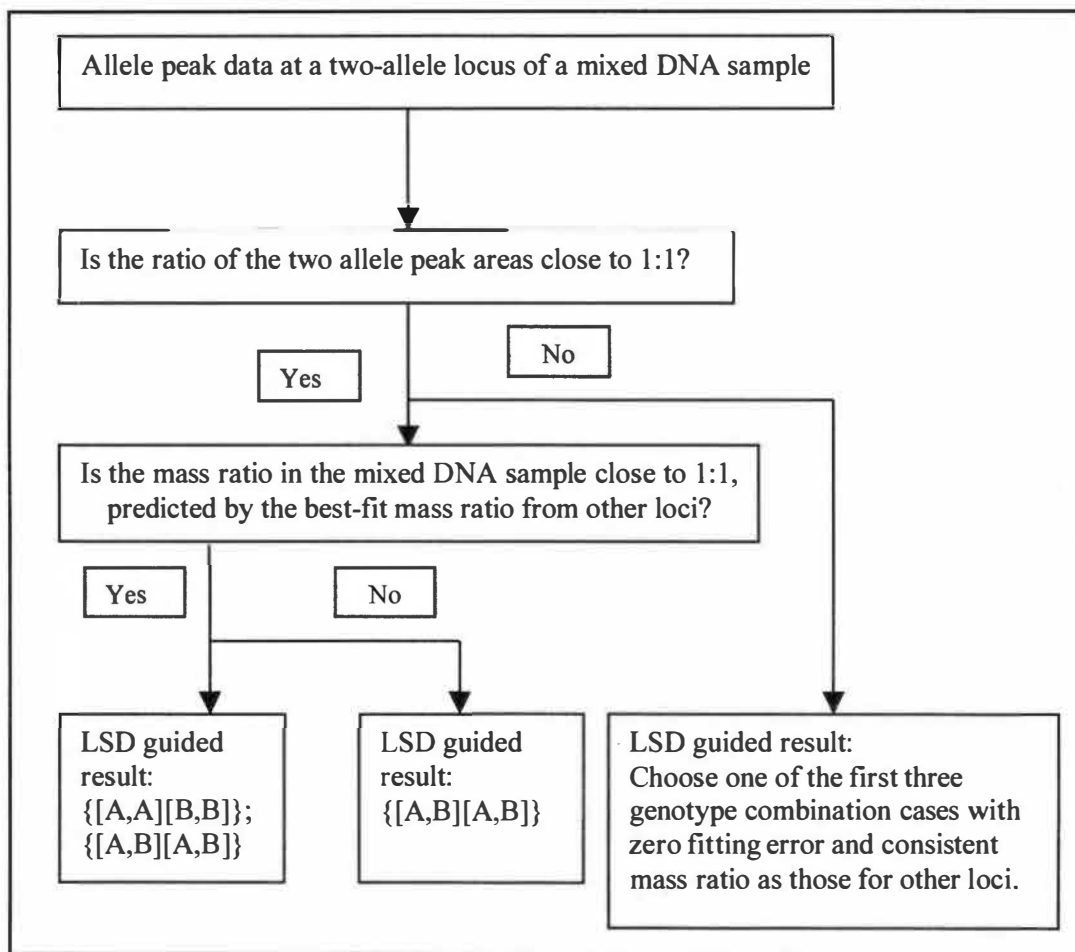


Figure 4.2: Flow sheet of using LSD to resolve mixed DNA at a 2-allele locus.

locus. Examples of LSD results and the associated interpretation for each of 4-allele, 3-allele, and 2-allele locus case will now be presented in the following sections.

4.2. Examples of Using LSD in Resolving Real Forensic Mixture DNA

Ms. Patricia Graham of the Texas Department of Public Safety in Corpus Christi provided five sets of mixture DNA data from rape cases, and LSD was used to resolve them. For each case, the DNA genotype for the victim was known and provided. The DNA typing of a suspect was also included for each of the five cases. Texas DNA mixture data No. 1 will be used to illustrate how LSD algorithm works in resolving mixed DNA. Figure 4.3 is the output of a Genescan software showing allele peaks at nine loci [19]. The peak areas and the genotypes of the victim and the suspect are listed in Table 4.4. Figure 4.4 shows the output of web-LSD, which is based on a previous Visual Basic version and Matlab program of LSD, for this set of DNA mixture data [20]. The fitting of the peak area data for loci VWA (3 alleles), D21S11 (4 alleles), and D7S820 (2 alleles) will be used to illustrate the workings of LSD.

4.2.1. Interpretation of the LSD Result for a Three-Allele Locus

As shown in Table 4.4, locus VWA has three alleles. The allele peak data at this locus are first normalized and represented with a vector ‘ b ’:

$$b = \begin{bmatrix} 616 / 425 \\ 2021 / 425 \\ 425 / 425 \end{bmatrix} = \begin{bmatrix} 1.45 \\ 4.76 \\ 1 \end{bmatrix}. \quad (4.4)$$

All six possible genotype combinations for this 3-allele locus and their corresponding

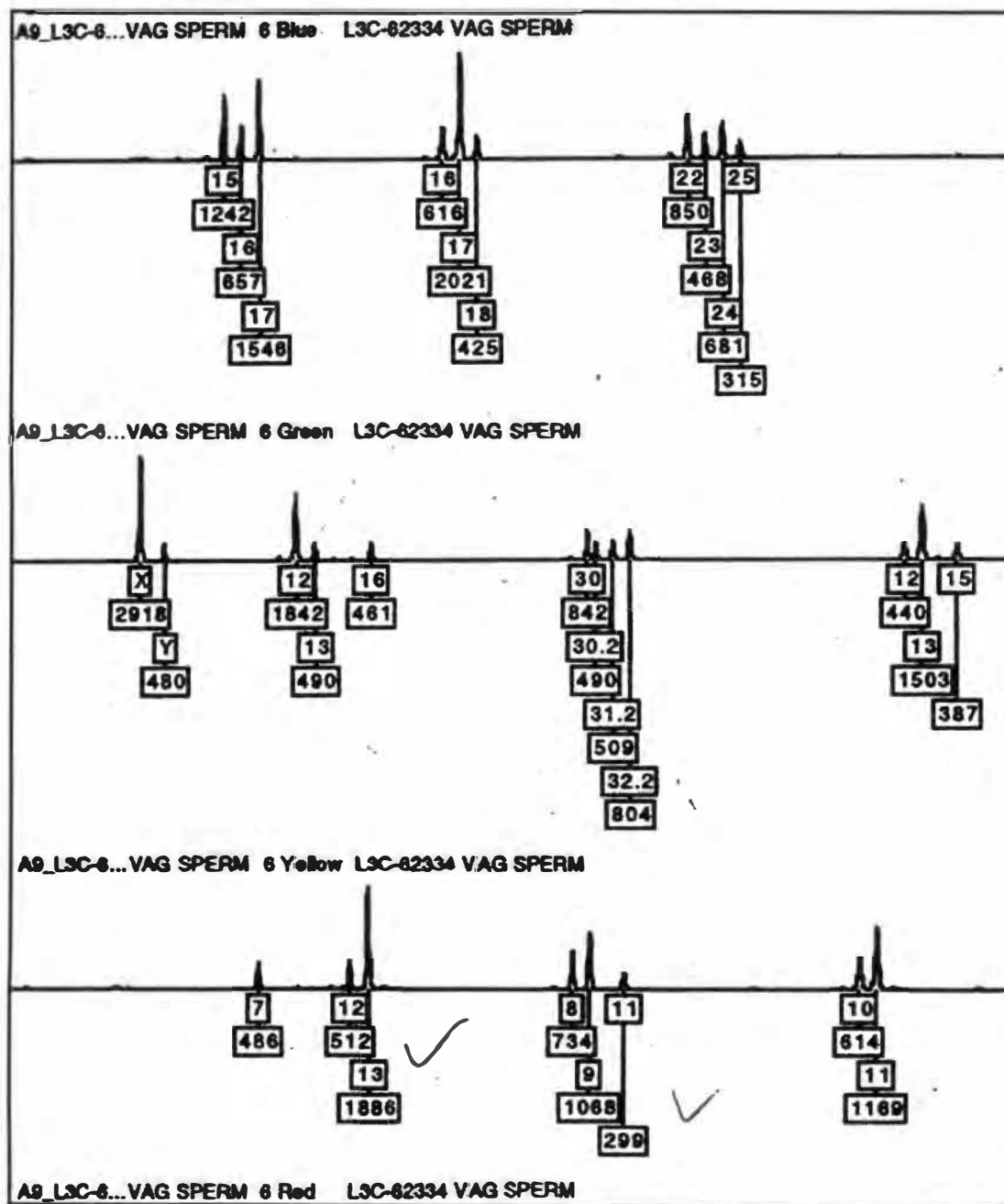


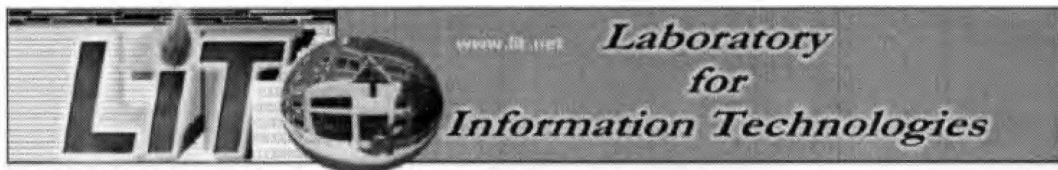
Figure 4.3: The allele peaks at nine loci of Texas DNA mixture data No. 1; both the allele name and the corresponding peak area are indicated.

Source: P. Graham "Allele Peak Data for DNA Mixture samples from real cases," Texas Department of Public Safety, Corpus Christi, 2001, Personal Communication.

Table 4.4: The allele peak area data at nine loci and the genotypes of both contributors for the DNA mixture sample shown in Figure 4.3.

Locus	Alleles in the Mixture	Peak Areas	Genotypes	
			Victim	Suspect
D3S1358	15	1242	15, 17	16, 17
	16	657		
	17	1546		
VWA	16	616	17, 17	16, 18
	17	2021		
	18	425		
FGA	22	850	22, 24	23, 25
	23	468		
	24	681		
	25	315		
D8S1179	12	1842	12, 12	13, 16
	13	490		
	16	461		
D21S11	30	842	30, 32.2	30.2, 31.2
	30.2	490		
	31.2	509		
	32.2	804		
D18S51	12	440	13, 13	12, 15
	13	1503		
	15	387		
D5S818	7	486	13, 13	7, 12
	12	512		
	13	1886		
D13S317	8	734	8, 9	9, 11
	9	1068		
	11	299		
D7S820	10	614	11, 11	10, 10
	11	1169		

Source of data: P. Graham. "Allele Peak Data for DNA Mixture samples from real cases," Texas Department of Public Safety, Corpus Christi, 2001.



The results for your LSD prepared for nxue Case ID: Texas_data_1 from host at Address 192.249.3.1

This report was prepared on Thu 30-May-2002 03:44 PM

LSD Results

Locus	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio	Comments
D3S1358						Three Alleles
1	16,17	15,17	9.6E-02	1.0	1.0:2.1	
2	16,16	15,17	0.11	1.1	1.0:4.2	
3	17,17	15,16	0.40	4.1	1.0:1.2	
4	15,16	15,17	0.71	7.4	1.0:3.6	
5	15,15	16,17	0.92	9.5	1.0:1.8	
6	15,16	16,17	3.51	36	1.0:1.6	
TH01						No Alleles
D21S11						Four Alleles
1	30.2,31.2	30,32.2	3.8E-03	1.0	1.0:1.6	
2	30.2,32.2	30,31.2	0.44	116	1.0:1.0	
3	31.2,32.2	30,30.2	0.44	117	1.0:1.0	
D18S51						Three Alleles
1	12,15	13,13	9.4E-03	1.0	1.0:1.8	
2	13,15	12,13	1.02	108	1.0:1.1	
3	15,15	12,13	3.77	402	1.0:5.0	
4	12,12	13,15	4.16	443	1.0:4.3	
5	12,15	12,13	4.68	499	1.0:-10.6	
6	12,15	13,15	5.39	575	1.0:-12.5	

Figure 4.4: Output of web-LSD for Texas DNA mixture data No. 1. All possible genotype combinations are ranked at each locus according to the magnitudes of the fitting errors.

D5S818 Three Alleles

1	7,12	13,13	1.4E-03	1.0	1.0:1.9	
2	7,13	12,13	1.11	778	1.0:1.0	
3	7,7	12,13	4.00	2793	1.0:4.9	
4	12,12	7,13	4.15	2899	1.0:4.6	
5	7,12	12,13	4.88	3412	1.0:-9.4	
6	7,12	7,13	5.16	3605	1.0:-10.0	

D13S317 Three Alleles

1	9,11	8,9	4.6E-03	1.0	1.0:2.4	
2	11,11	8,9	0.62	137	1.0:6.0	
3	8,11	9,9	1.06	232	1.0:1.0	
4	8,11	8,9	1.49	327	1.0:9.7	
5	8,8	9,11	3.31	724	1.0:1.9	
6	8,11	9,11	8.42	1844	1.0:2.4	

D7S820

Two Alleles
REVIEW the
LSD Results

1	10,10	11,11	0.0E+00		1.0:1.9	
2	10,10	10,11	0.0E+00		1.0:-4.2	
3	11,11	10,11	0.0E+00		1.0:2.2	
4	10,11	10,11	0.41		1.0:1.0	

D16S539

No Alleles

CSF1PO

No Alleles

VWA

Three Alleles

1	16,18	17,17	0.10	1.0	1.0:1.9	
2	17,18	16,17	1.77	18	1.0:1.3	
3	18,18	16,17	5.46	54	1.0:6.2	
4	16,18	16,17	6.18	61	1.0:-7.6	
5	16,16	17,18	7.05	70	1.0:4.0	
6	16,18	17,18	9.03	89	1.0:-10.6	

Figure 4.4: Continued.

D8S1179						Three Alleles
1	13,16	12,12	2.0E-03	1.0	1.0:1.9	
2	12,16	12,13	1.25	629	1.0:1.0	
3	16,16	12,13	4.30	2173	1.0:5.1	
4	13,13	12,16	4.49	2268	1.0:4.7	
5	13,16	12,13	5.16	2606	1.0:-8.6	
6	13,16	12,16	5.49	2775	1.0:-9.1	
TPOX						No Alleles
FGA						Four Alleles
1	23,25	22,24	0.26	1.0	1.0:2.0	
2	24,25	22,23	1.41	5.4	1.0:1.3	
3	23,24	22,25	1.67	6.4	1.0:1.0	
PentaD						No Alleles
PentaE						No Alleles
D19S433						No Alleles
D2S1338						No Alleles
SE33						No Alleles
Thu 30-May-2002 03:44 PM						

Figure 4.4: Continued.

matrix 'A' at this locus are listed in Table 4.5. For each possible genotype combination, the least-square solution $\hat{x}$ for the equation $Ax = b$ is solved by multiplying the corresponding pseudoinverse of A, (i.e. A^+) to the vector 'b'. The error residual was calculated according to Equation 4.3. Case 1 is used as an example to show how the residual error is calculated. For Case 1, the corresponding A matrix is: $\begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$. The linear

equation for this case is:

$$Ax = b \Rightarrow \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} x = \begin{bmatrix} 1.45 \\ 4.76 \\ 1 \end{bmatrix}, \quad (4.5)$$

and the least-square solution $\hat{x}$ is decided as follows:

$$\hat{x} = A^+ b = \frac{1}{2} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1.45 \\ 4.76 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.73 \\ 2.88 \end{bmatrix}, \quad (4.6)$$

where A^+ is obtained from Table 4.1. The error for the fit is calculated as:

$$\begin{aligned} E^2 &= \|(b - A\hat{x})\|^2 = \left\| \begin{bmatrix} 1.45 \\ 4.76 \\ 1 \end{bmatrix} - \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0.73 \\ 2.88 \end{bmatrix} \right\|^2 = \left\| \begin{bmatrix} -0.01 \\ 1.88 \\ -1.88 \end{bmatrix} \right\|^2 \\ &= (-0.01)^2 + 1.88^2 + (-1.88)^2 = 7.0689 \end{aligned} \quad (4.7)$$

The normalized mass ratio vector is: $\begin{bmatrix} 0.73/0.73 \\ 2.88/0.73 \end{bmatrix} = \begin{bmatrix} 1 \\ 3.95 \end{bmatrix}$, which is the normalized

least-square solution $\hat{x}$ reflecting the approximate mass ratio in the mixture sample for that locus. The least-square solution $\hat{x}$ and the associated residual errors for other cases

Table 4.5: All possible genotype combinations for the two contributors and their corresponding matrices for locus VWA with three alleles, 16, 17, 18, for the DNA mixture data shown in Table 4.4.

Case	Genotype Combination		Matrix (A)	Pseudoinverse of the Matrix (A ⁺)
	Person 1	Person 2		
1	16, 16	17, 18	$\begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$
2	17, 17	16, 18	$\begin{bmatrix} 0 & 1 \\ 2 & 0 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$
3	18, 18	16, 17	$\begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 2 & 1 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$
4	16, 17	17, 18	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{3} \begin{bmatrix} 2 & 1 & -1 \\ -1 & 1 & 2 \end{bmatrix}$
5	16, 17	16, 18	$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{3} \begin{bmatrix} 1 & 2 & -1 \\ 1 & -1 & 2 \end{bmatrix}$
6	16, 18	17, 18	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$	$\frac{1}{3} \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \end{bmatrix}$

were also calculated in the same manner as that for Case 1, and results are shown in Table 4.6. The first column of the table is the case number for each genotype combination; all possible genotype combination cases are listed in the second and third columns; the fourth column of the table is the corresponding genotype matrix; the fifth column is the ratio of the calculated errors, with the smallest error normalized to one; the normalized mass ratios calculated from the fitted $\hat{x}$ are listed in the sixth column. Figure 4.5 shows the error residual for each of the possible genotype combination cases. Therefore, all the genotype combinations for locus VWA can be ranked according to the magnitudes of the corresponding residual errors. Case 2 clearly has the smallest fitting error. The next smallest error, that of Case 4, has an error 17 times as large as that for Case 2, indicating genotype combination of Case 2 is clearly the best-fit one if the relative allele peak area proportion is credible. Table 4.7 is the ranking result for locus VWA with three alleles. From the ranking result, one can see that the fitting error of the first ranked genotype combination is clearly much smaller than the next smallest one; therefore, the genotype combination of $\{[16, 18], [17, 17]\}$ is identified as the best-fit for this locus, and the mix is at the approximate mass ratio of 1:1.94. The genotype combination with a negative number in the calculated mass coefficients can be excluded at once because the mass coefficients must be positive.

4.2.2. Interpretation of the LSD Result for a Four-Allele Locus

The allele peak data at locus D21S11 shown in Table 4.4 is used to illustrate how a 4-allele pattern at a locus is resolved into its most likely components using LSD. First, all possible genotype assignments for this locus are listed in Table 4.8. For each possible

Table 4.6: Normalized fitting errors and calculated mass ratios for all possible genotype combination cases for locus VWA with three alleles as shown in Table 4.5.

Cases	Genotype Combinations		Matrix A	Ratio of Errors	Mass Ratio Calculated
	Person 1	Person 2			
1	16, 16	17, 18	$\begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$	69.82	1: 3.97
2	16, 18	17, 17	$\begin{bmatrix} 0 & 1 \\ 2 & 0 \\ 0 & 1 \end{bmatrix}$	1	1: 1.94
3	18, 18	16, 17	$\begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 2 & 1 \end{bmatrix}$	54.11	1: 6.20
4	17, 18	16, 17	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$	17.55	1: 1.25
5	16, 18	16, 17	$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	61.20	1: -7.63
6	16, 18	17, 18	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$	89.42	1: -10.58

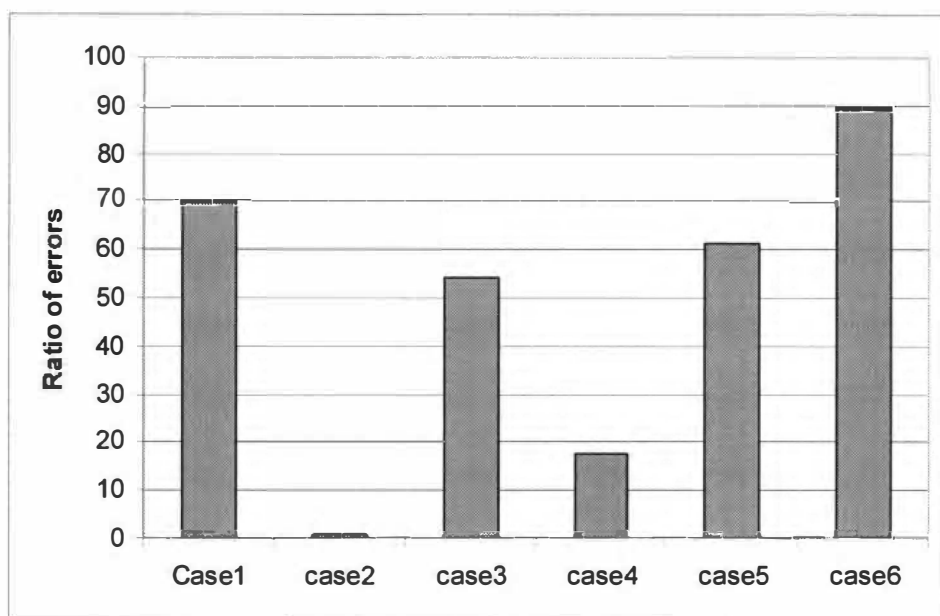


Figure 4.5: Normalized fitting residuals of all possible genotype combinations at the locus VWA in the DNA mixture data shown in Table 4.4.

Table 4.7: Ranking result for locus VWA in the DNA mixture data shown in Table 4.4, according to the magnitudes of fitting errors. The fitting error for the genotype combination of {[16, 18][17, 17]} is much smaller than those for other cases. Therefore, this pattern is identified as the most likely one.

Rank	Genotype Combinations		Matrix A	Ratio of Errors	Mass Ratio Calculated
	Person 1	Person 2			
1	16, 18	17, 17	$\begin{bmatrix} 0 & 1 \\ 2 & 0 \\ 0 & 1 \end{bmatrix}$	1	1: 1.94
2	17, 18	16, 17	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$	17.55	1: 1.25
3	18, 18	16, 17	$\begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 2 & 1 \end{bmatrix}$	54.11	1: 6.20
4	16, 18	16, 17	$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	61.20	1: -7.63
5	16, 16	17, 18	$\begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$	69.82	1: 3.97
6	16, 18	17, 18	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$	89.42	1: -10.58

Table 4.8: All possible genotype combinations and the corresponding matrices for the two contributors at locus D21S11 with four alleles 30, 30.2, 31.2, 32.2 in the DNA mixture as shown in Table 4.4.

Case	Genotype Combination		Matrix (A)	Pseudoinverse of the Matrix (A^+)
	Person 1	Person 2		
1	30, 30.2	31.2, 32.2	$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$
2	30, 31.2	30.2, 32.2	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$
3	30, 32.2	30.2, 31.2	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$

genotype combination, the fitting error and the best-fit mass coefficients are calculated in the same manner as that described in Section 4.2.1. Figure 4.6 shows the relative error of each possible genotype combination when normalized to the smallest error (Case 3). The LSD ranking result for this locus is shown in Table 4.9. The top ranked genotype combination of {[30.2, 31.2][30, 32.2]} is considered to be the most likely one with the approximate mass ratio of 1:1.65. Note that the best-fit mass ratio of 1:1.65 for this locus is consistent with the best-fit mass ratio of 1:1.94 for the 3-allele locus VWA, derived in the previous section.

4.2.3. Interpretation of the LSD Result for a Two-Allele Locus

If only two alleles are present at a locus, then there are four possible genotype combinations for the two contributors, as shown in Table 4.2. The allele peak data at locus D7S820 in Table 4.4 is used to illustrate the LSD method for the 2-allele case.

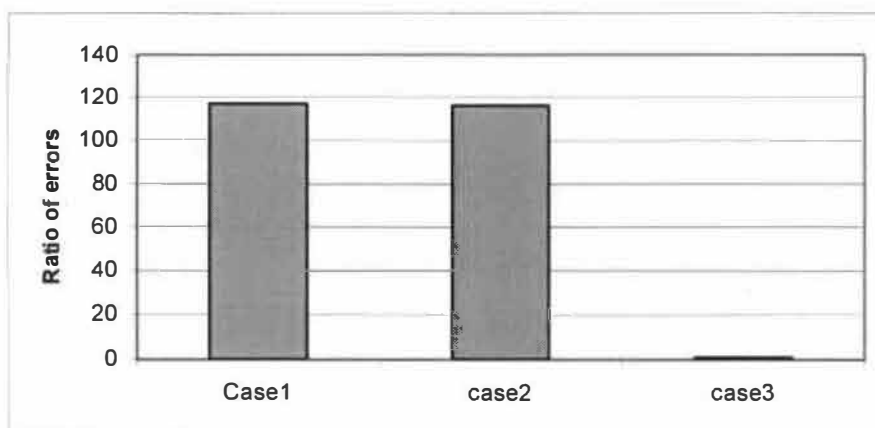


Figure 4.6: Normalized fitting residuals of all possible genotype combinations at the locus D21S11 in the DNA mixture data shown in Table 4.4.

Table 4.9: Ranking result according to the magnitudes of fitting errors for locus D21S11 in the DNA mixture data shown in Table 4.4. The fitting error for the genotype combination of {[30.2, 31.2][30, 32.2]} is much smaller than those for other cases. Therefore, this pattern is identified as the most fitting case at locus D21S11.

Rank	Genotype Combinations		Matrix A	Ratio of Errors	Mass Ratio Calculated
	Person 1	Person 2			
1	30.2, 31.2	30, 32.2	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$	1	1: 1.65
2	30.2, 32.2	30, 31.2	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	116.06	1:1.04
3	31.2, 32.2	30, 30.2	$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$	116.86	1:1.01

All possible genotype assignments for this locus are listed in Table 4.10.

At this locus, the errors of all possible genotype combinations are calculated. The calculated errors and the ranking result for this locus are shown in Table 4.11. Since peak area ratio of the two alleles (614:1169) at locus D7S820 is not close to 1:1, one does not need to consider Case 4 with the genotype combination of {[10, 11][10, 11]} as a possibility. Recall the best-fit mass ratios from loci VWA and D21S11 are 1:1.94 and 1:1.65, respectively. Therefore, either Case 1 or Case 3 is a possibility with the fitted mass ratio of 1:1.9 and 1:2.21, respectively. These two genotype combinations are {[10, 10][11, 11]} and {[11, 11][10, 11]}. The genotype of the victim, if available, can definitely help the caseworkers to select the more appropriate genotype combination case from the two alternative choices. If more than one case yields a mass ratio consistent with those from other loci, then DNA for this locus cannot be resolved unambiguously, which means LSD suggests more than one genotype combination, both of which are equally likely to be the true case. The only way to ascertain the choice is to pick out the genotype combination case, in which the known victim's genotype or the reference genotype is one of the genotype profiles in the combination. It turns out the combination of {[10, 10] [11, 11]} is the pick of the choice because the victim's profile at this locus is [11, 11].

The other six loci of Texas DNA Mixture Data No. 1 are also analyzed with the same LSD interpretation approach. The composite profile for this DNA mixture sample is shown in Table 4.12 by pulling together the top ranked genotype combination cases from 3-allele and 4-allele loci and taking into consideration the post-LSD interpretation

Table 4.10: All possible genotype combinations and the corresponding matrices for the two contributors at locus D7S820 with two alleles 10, 11 in the DNA mixture data shown in Table 4.4.

Case	Genotype Combination		Matrix (A)	Pseudoinverse of the Matrix (A^+)
	Person 1	Person 2		
1	10, 10	11, 11	$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
2	10, 11	10, 10	$\begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 0 & 2 \\ 1 & -1 \end{bmatrix}$
3	10, 11	11, 11	$\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$	$\frac{1}{2} \begin{bmatrix} 2 & 0 \\ -1 & 1 \end{bmatrix}$
4	10, 11	10, 11	$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$	$\frac{1}{4} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

Table 4.11: Ranking result for locus D7S820 in the DNA mixture data shown in Table 4.4 according to the magnitudes of fitting errors. Three genotype combinations have zero fitting errors; two of them (with shaded areas) have consistent mass ratios with other loci. Therefore, both of these two patterns of {[10, 10][11, 11]} and {[11, 11][10, 11]} were selected as the most fitting cases at locus D7S820.

Cases	Rank	Genotype Combinations		Matrix A	Errors	Mass Ratio Calculated
		Person 1	Person 2			
1	1	10, 10	11, 11	$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	0	1: 1.90
2	1	10, 10	10, 11	$\begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}$	0	1: -4.21
3	1	11, 11	10, 11	$\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$	0	1: 2.21
4	2	10, 11	10, 11	$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$	0.41	Indeterminate

Table 4.12: LSD suggested genotype resolution results at all nine loci for the DNA mixture data shown in Table 4.4. There are two possible choices at locus D7S820; the one with shaded area is the true one for the two contributors.

Locus	Alleles	Ratio of Peak Area	LSD guided Result			Comment
			Mass Ratio Calculated	Person 1 (suspect)	Person 2 (victim)	
D3S1385	15, 16, 17	1.89: 1: 2.35	1: 2.08	16, 17	15, 17	Correct
VWA	16, 17, 18	1.45: 4.76: 1	1: 1.94	16, 18	17, 17	Correct
FGA	22, 23, 24, 25	2.70: 1.49: 2.16: 1	1: 1.96	23, 25	22, 24	Correct
D8S1179	12, 13, 16	4.00: 1.06: 1	1: 1.94	13, 16	12, 12	Correct
D21S11	30, 30.2, 31.2, 32.2	1.72: 1: 1.04: 1.64	1: 1.65	30.2, 31.2	30, 32.2	Correct
D18S51	12, 13, 15	1.14: 3.88: 1	1: 1.82	12, 15	13, 13	Correct
D5S818	7, 12, 13	1: 1.06: 3.88	1: 1.89	7, 12	13, 13	Correct
D13S317	8, 9, 11	2.45: 3.57: 1	1: 2.40	9, 11	8, 9	Correct
D7S820	10, 11	1: 1.90	1: 1.90	10, 10	11, 11	Two possible choices
			1: 2.21	11, 11	10, 11	

Note: 'correct' in the table means the top ranked genotype is consistent with the true genotype profiles of the victim and the suspect.

guidelines for the 2-allele loci. Compared with the true genotypes of the two contributors, one can see that the top ranked genotype combinations from LSD give the correct genotype profiles at all 3-allele and 4-allele loci, for this example. For those 2-allele loci, LSD yielded two possible choices for each locus. When the genotype of the victim at this locus is brought in, the better choice can now be picked out, in which person 1's profile also proves to be consistent with the genotype of the suspect. Therefore, in this case, LSD results at all nine loci are consistent with the genotypes of the victim and the suspect. In other words, LSD results strongly suggest that the suspect is indeed the true perpetrator. One can validate this claim by performing a match probability or maximum likelihood calculation. After several months, Ms. Graham notified us that such conclusion is correct that the suspect did turn out to be the true perpetrator in this rape case.

Compared with other approaches published in the literature, LSD is more efficient and simpler, and gives the true genotypes when the allele peak data preserves the relative DNA mass proportionality. The advantage of LSD is its direct calculation of the best-fit genotype combination and the approximate mass ratio, without iteration as reported in Reference [5]. Furthermore, LSD is applied to each locus independently of other loci, thus allowing each locus to be fitted independently with its own best-fit mass ratio. From examining the relative errors of each fit at a locus, the degree of confidence can be separately assigned to the resulting best-fit genotype, locus by locus, allowing a composite profile to be assembled with a high degree of confidence containing only those loci whose LSD results are clear cut. Generally speaking if the allele peak data is consistent with the underlying mass ratio, then at least 8-10 loci can usually be resolved

confidently. However, in rape cases involving close relatives, the mixed DNA evidence may contain a higher number of two-allele loci, caution must be exercised in interpreting LSD results. In this case, the victim's profile becomes even more important in serving as the reference genotype to help ascertain the genotype combination patterns for the two contributors of the mixture.

CHAPTER 5

RESULTS AND DISCUSSION

Up to date, more than 60 sets of DNA mixture data from real criminal cases, provided by different forensic laboratories, have been analyzed using LSD. The true genotypes of both contributors were also known. In these cases, they serve as references to validate LSD guided resolution results. In most cases, the top-ranked genotype combinations selected by LSD are consistent with the true genotype profiles of the two contributors. Such results support LSD's superiority to other existing approaches in interpreting DNA mixtures.

In this chapter, LSD guided deconvolution results of some DNA mixture data, and discussion about the LSD algorithm and guidelines for its use are presented. To test the effectiveness of LSD, the degree of noise that can be tolerated by LSD algorithm before the wrong set of genotypes is chosen to be the top-ranked genotype profiles is first investigated. The results are presented in Section 5.1. Results using LSD to resolve three sets of simulated DNA mixture data with different mass ratios are presented in Section 5.2. This is followed by Section 5.3, which presents the summary of the limitations of LSD algorithm. In Section 5.4, two sets of DNA mixture data obtained from literature were used to test the feasibility of the LSD algorithm to see if LSD arrives at the same genotype profiles as the true cases. In Section 5.5, LSD suggested genotype resolution results of five sets of real forensic DNA mixture data are presented; three of them were provided by Mr. Ray Wickenheiser of the Acadiana Crime Lab in New Iberia, LA; the

other two sets of data were provided by Ms. Patricia Graham of the Texas Department of Public Safety in Corpus Christi. In Section 5.6, comparison of the effectiveness of LSD to that of a new method, Linear Mixture Analysis (LMA), as reported in reference [18], will be made.

5.1. Sensitivity of LSD to Noise in the Allele Peak Data

First, the degree of tolerance of LSD to the presence of random noise in the allele peaks is investigated. Effectiveness of LSD is based on the assumption that the allele peak areas be proportional to the relative mass portions of the corresponding alleles in the mixture. Under most circumstances, such relationships are compromised due to the unequal DNA amplification in the amplification process, or stutter peaks superimposed onto a true allele peak. For a three-allele or four-allele locus being tested, if the relative amount of noise in the measured allele peak signal is not too much, the top ranked genotype combination pair in the LSD ranking result would still correspond to the true genotype profiles of the two contributors. If the noise in the allele peak signal becomes significant, then there would be a severe imbalance between the allele peak areas, thereby masking the true relative DNA mass proportions in the mixture. As a result, the top-ranked genotype combination pair indicated by LSD may not turn out to be the true genotype profiles of the two contributors. In this section, results of a study of the maximum percentage of noise present in the allele peak data tolerable by LSD in resolving simulated DNA mixture data are presented.

Simulated allele peak area data for DNA mixtures with different mass ratios were generated. Allele peak areas at each locus are simply a superposition of the peak areas of two individual allele peaks. Different genotype combinations were used to generate the simulated data. Then, for each round of simulation, a different percentage of Gaussian noise was added to each of the peak area data. The range of mass ratio being tested is between 1:1 and 20:1. Among this range, the following mass ratios were selected 1:1, 1.2:1, 1.5:1, 1.7:1, 2:1, 5:1, 8:1, 10:1, 13:1, 15:1 and 20:1, a total of eleven sets of DNA mixture data. For each set, allele peak areas were simulated at six loci; four of them have three alleles each, the other two loci have four alleles each. At each locus, all possible genotype combinations are to be ranked using LSD. Initially, when the percentage of the noise is 0, the top ranked genotype combination pair should correspond to the correct genotype profiles for the two contributors with zero fitting errors. For each subsequent run, the percentage of noise is increased by 4%, until at a locus, the top ranked genotype combination selected by LSD consistently fails to give the correct genotype combination profiles. Thus, the maximum percentage of the tolerable noise for a given mass ratio is obtained. Applying the same method for other chosen mass ratios, the maximum percentage of the tolerable noise at all the selected mass ratios of the DNA mixture samples could be obtained. The results are summarized in Table 5.1.

Figure 5.1 shows the relationship between the mass ratio in the DNA mixture sample and the maximum tolerable percentage of Gaussian noise. One can see that the maximum tolerable percentage of noise increases as the mass ratio of the DNA mixture

Table 5.1: Maximum Gaussian noise present in the allele peak data tolerable by LSD in resolving DNA mixture samples.

Mass Ratio in the DNA Mixture Sample	Maximum Percentage of Tolerable Noise in the Allele Peak Data
1:1	0
1.2:1	6%
1.5:1	12%
1.7:1	12%
2:1	16%
5:1	16%
8:1	16%
10:1	20%
13:1	20%
15:1	20%
20:1	24%

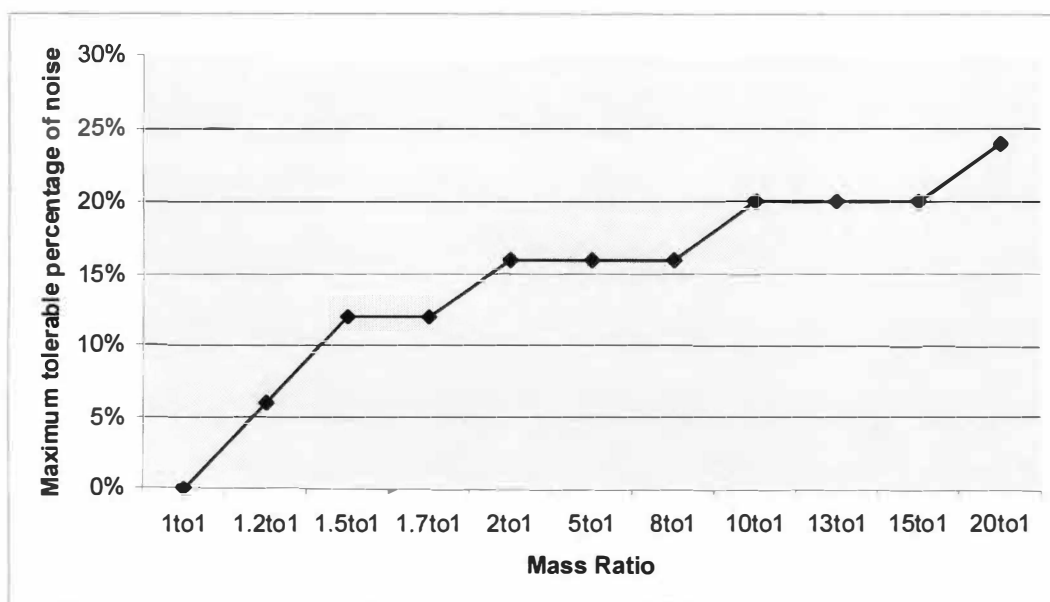


Figure 5.1: Maximum Gaussian noise present in the allele peak data tolerable by LSD in resolving DNA mixture data.

becomes larger. This is within expectation, as the relative peak areas become more and more distinct with increasing difference in the DNA proportions in the mixture, the more noise each peak can accommodate without distorting the overall relative proportions of the DNA mass represented by the allele peak data.

5.2. LSD Interpretation Results Using Simulated DNA Mixture Data

Allele peak data of DNA mixture samples were simulated at six loci, with various number of alleles at each locus, with mass ratios ranging from 1:1 to 1:5. In this thesis, three sets of simulated DNA mixture data were used as presented in Table 5.2. The mass ratios used in the three sets of simulated data are 1:5, 1:2, and 1:1, respectively. First, at each locus, the genotype combination of the two contributors is coded into a matrix ' A '. In accordance to the mass ratio assumed for each set of data, the corresponding theoretical allele peak data vector ' b ' at this locus was then calculated from the equation $Ax = b$, where the ' x ' vector contains the mass ratio coefficients for the two contributors (such as $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$). In real cases, some noise or peak area imbalances are always associated with the measurements of allele peak areas. Therefore, in our simulated data, a 10% Gaussian noise (random Gaussian noise with mean 0 and covariance of 10% of the nominal peak area) is added to the theoretical allele peak values.

LSD was then applied in turn to each of these three sets of simulated DNA mixture data (Chapter 4 includes the details of the algorithm.). The LSD suggested

Table 5.2: Simulated allele peak data and genotypes for the two contributors at six loci for mixture DNA samples with the mass ratios of 1:5, 1:2, and 1:1, respectively. 10% Gaussian noise was added to each set of data.

Locus	Alleles in the Mixture	Genotype Combination		Simulated Allele Peak Data		
		Person 1	Person 2	1: 5	1: 2	1: 1
VWA	15 16 19	15, 15	16, 19	1302 3369 3446	1506 1369 1348	1472 730 769
TH01	6 7 9.3	6, 7	6, 9.3	5620 807 5095	5008 1875 3155	5979 3057 3122
D8S1179	11 13	11, 13	11, 13	5519 4913	5284 5733	4339 4922
FGA	22 25	22, 22	22, 25	9019 6240	4675 2232	8596 2444
D21S11	28 30.2	28, 28	30.2, 30.2	958 5234	1098 1941	1001 873
D18S51	16 17 18 20	16, 20	17, 18	804 5202 4212 864	731 1806 1640 913	975 882 899 873

genotype resolution results are shown in Tables 5.3 through 5.7. If LSD algorithm suggests more than one possible choice of the best-fit genotype combinations at a locus, then both are listed in the tables but the one in shaded cells indicates the true genotype combination profiles of the two contributors.

5.2.1. LSD Interpretation Result Using Simulated DNA Mixtur Data with a Mass Ratio of 1:5

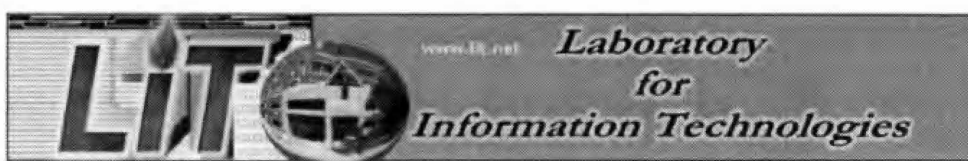
Table 5.3 is the LSD suggested genotype resolution result of the simulated DNA mixture data given in Table 5.2 with the mass ratio of 1:5. Figure 5.2 shows the detailed web-LSD output for this set of data. For loci VWA, TH01, FGA, D21S11, and D18S51, the top-ranked genotype combinations suggested by LSD correspond to the correct genotype profiles of the two contributors. At these five loci, the fitted mass ratios are reviewed to be more or less consistent with each other ranging from 1:4.49 to 1:7.01. The average of the five mass ratios is 1:5.57, which is close to the true mass ratio of 1:5 in this simulated DNA mixture sample.

The remaining locus of D8S1179 has only 2 alleles, and LSD suggested results for a 2-allele locus always needs special review and interpretation. The ranking result for all possible genotype combinations for this locus is presented in Table 5.4. Three genotype combinations all fit the given allele peak data exactly with zero errors. However, their corresponding best-fit mass ratios are 1:1.1, 1: -18.2, and 1:16.2; all are inconsistent with those for other loci. For the genotype combination of {[11, 13] [11, 13]}, if it were true, then regardless of the mass ratio in the DNA mixture, the ratio of the

Table 5.3: LSD suggested genotype resolution result of a set of simulated DNA mixture data shown in Table 5.2, with a mass ratio of 1:5. Five of a total of six loci were successfully resolved by selecting the top-ranked genotype combinations with consistent mass ratios.

Locus	Alleles in the Mixture	Peak Area Ratio Measured	LSD Suggested Genotype Resolution Result			
			Person 1	Person 2	Mass Ratio Calculated	Remarks
VWA	15, 16, 19	1: 2.59: 2.65	15, 15	16, 19	1:5.23	Correct
TH01	6, 7, 9.3	6.96: 1: 6.31	6, 7	6, 9.3	1:7.01	Correct
D8S1179	11, 13	1.12: 1	11, 13	11, 13	Indeterminate	See Table 5.4 for details
FGA	22, 25	1.45: 1	22, 22	22, 25	1:4.49	Correct
D21S11	28, 30.2	1: 5.46	28, 28	30.2, 30.2	1:5.46	Correct
D18S51	16, 17, 18, 20	1: 6.47: 5.24: 1.07	16, 20	17, 18	1:5.64	Correct

Note: 'correct' in the table means the top-ranked genotype combination selected by LSD corresponds to the true genotype profiles of the two contributors.



The results for your LSD prepared for nxue Case ID: Simulated_1to5 from host at Address 192.249.3.1

This report was prepared on Tue 09-Jul-2002 12:45 PM

LSD Results

Locus	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio	Comments
D3S1358						No Alleles
TH01						Three Alleles
1	6,7	6,9.3	4.1E-02	1.0	1.0:7.0	
2	7,7	6,9.3	0.21	5.2	1.0:13.3	
3	7,9.3	6,9.3	0.91	22	1.0:14.3	
4	6,6	7,9.3	14.12	347	1.0:1.1	
5	9.3,9.3	6,7	17.79	437	1.0:1.3	
6	7,9.3	6,7	50.25	1234	1.0:1.3	
D21S11						Two Alleles REVIEW the LSD Results
1	28,28	30.2,30.2	0.0E+00		1.0:5.5	
2	28,28	28,30.2	0.0E+00		1.0:-2.4	
3	28,30.2	30.2,30.2	0.0E+00		1.0:2.2	
4	28,30.2	28,30.2	9.96		1.0:1.0	
D18S51						Four Alleles
1	16,20	17,18	0.76	1.0	1.0:5.6	
2	16,18	17,20	23.54	31	1.0:1.2	
3	18,20	16,17	23.63	31	1.0:1.2	
D5S818						No Alleles
D13S317						No Alleles
D7S820						No Alleles

Figure 5.2: Output of web-LSD for the simulated DNA mixture data with a mass ratio of 1:5. All possible genotype combinations are ranked at each locus according to the magnitudes of the fitting errors.

D16S539

No Alleles

CSF1PO

No Alleles

VWA

Three Alleles

1	15,15	16,19	1.7E-03	1.0	1.0:5.2	
2	15,19	16,19	0.30	169	1.0:3.3	
3	15,16	16,19	0.37	214	1.0:3.5	
4	19,19	15,16	1.26	721	1.0:1.4	
5	16,16	15,19	1.36	775	1.0:1.4	
6	15,16	15,19	5.98	3417	1.0:1.1	

D8S1179

Two Alleles
REVIEW the
LSD Results

1	13,13	11,11	0.0E+00		1.0:1.1	
2	11,11	11,13	0.0E+00		1.0:16.2	
3	13,13	11,13	0.0E+00		1.0:-18.2	
4	11,13	11,13	7.6E-03		1.0:1.0	

TPOX

No Alleles

FGA

Two Alleles
REVIEW the
LSD Results

1	25,25	22,22	0.0E+00		1.0:1.4	
2	22,22	22,25	0.0E+00		1.0:4.5	
3	25,25	22,25	0.0E+00		1.0:-6.5	
4	22,25	22,25	9.9E-02		1.0:1.0	

PentaD

No Alleles

PentaE

No Alleles

D19S433

No Alleles

D2S1338

No Alleles

SE33

No Alleles

Tue 09-Jul-2002 12:45 PM

Figure 5.2: Continued.

Table 5.4: Ranking result at the two-allele locus D8S1179 in the simulated DNA mixture data shown in Table 5.2, with a mass ratio of 1:5. The fourth case in shaded cells is the genotype case of choice, because the mass ratios associated with other three cases are not consistent with those for other loci. It turns out that the fourth case is the correct one.

Case	Rank	Genotype Combinations		Matrix A	Errors	Mass Ratio Calculated
		Person 1	Person 2			
1	1	13, 13	11, 11	$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	0	1: 1.12
2	1	13, 13	11, 13	$\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$	0	1: -18.21
3	1	11, 11	11, 13	$\begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}$	0	1:16.21
4	2	11, 13	11, 13	$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$	0.0076	1:1

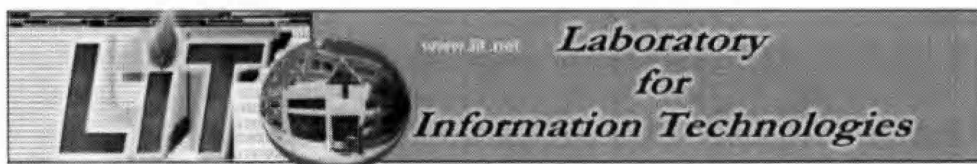
allele peak data would always be close to 1:1, and it is (1.12:1). As a result, the true mass ratio would be unknown at this locus. The best-fit mass ratio, given by LSD usually is not the correct one, due to the ‘ A ’ matrix for this case being underdetermined (meaning not all rows or columns of ‘ A ’ are independent), and mathematically an infinite set of solutions exist and LSD would give the one with the smallest length, the normalized result would always be 1:1. Therefore, for the case of 2-allele loci, when the other three genotype combinations all give a best-fit mass ratio that is inconsistent with those for other loci, the fourth case is the preferred choice, especially when the ratio of the two allele peaks is close to 1:1. As a result, we choose this genotype combination to be the most likely one. It turns out that it is the correct one.

5.2.2. LSD Interpretation Result Using Simulated DNA Mixture Data with a Mass Ratio of 1:2

Table 5.5 is the LSD suggested genotype resolution result for simulated DNA mixture data shown in Table 5.2 with a mass ratio of 1:2 and a 10% Gaussian noise added. Figure 5.3 shows the detailed web-LSD output for this set of data. LSD guided genotype combination pair at each locus of VWA, TH01, and D18S51 can be picked out by first ranking the least-square fitting residuals for all genotype combinations and then selecting the top-ranked one with the smallest fitting errors. At the two-allele locus D8S1179, the genotype combination of {[11, 13] [11, 13]} is selected as the most fitting one, because the ratio of the two allele peak areas is close to 1:1 (1:1.09) and the other patterns’ fitted mass ratios are not consistent with those for other loci. For loci FGA and D21S11, the ranking result for all possible genotype combinations is presented in

Table 5.5: LSD suggested genotype resolution result of a set of simulated DNA mixture data shown in Table 5.2, with a mass ratio of 1:2. Four of the six loci were correctly resolved by picking the top-ranked genotype combinations with consistent mass ratios. At the other two loci, each has two possible choices, because the mass ratio is close to 1:2 and the ratio of the two allele peak areas is also close to 1:2. The shaded cells indicate the true genotypes for the two contributors at these two loci.

Locus	Alleles in the Mixture	Peak Area Ratio Measured	LSD suggested Genotype Resolution Result			
			Person 1	Person 2	Mass Ratio Calculated	Remarks
VWA	15, 16, 19	1.12: 1.02: 1	15, 15	16, 19	1:1.80	Correct
TH01	6, 7, 9.3	2.67: 1: 1.68	6, 7	6, 9.3	1:1.69	Correct
D8S1179	11, 13	1: 1.09	11, 13	11, 13	Indeterminate	Correct
FGA	22, 25	2.09: 1	25, 25	22, 22	1:2.10	With equal confidence
			22, 22	22, 25	1:1.83	With equal confidence
D21S11	28, 30.2	1: 1.77	28, 28	30.2, 30.2	1:1.77	With equal confidence
			30.2, 30.2	28, 30.2	1:2.61	With equal confidence
D18S51	16, 17, 18, 20	1: 2.47: 2.24: 1.25	16, 20	17, 18	1:2.10	Correct



The results for your LSD prepared for nxue Case ID: simulated_1to2 from host at Address 192.249.3.1

This report was prepared on Tue 09-Jul-2002 01:04 PM

LSD Results

Locus	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio	Comments
D3S1358						No Alleles
TH01						Three Alleles
1	6,7	6,9.3	4.6E-05	1.0	1.0:1.7	
2	6,6	7,9.3	0.23	5078	1.0:1.0	
3	7,7	6,9.3	0.49	10641	1.0:4.4	
4	7,9.3	6,9.3	1.32	28715	1.0:6.0	
5	9.3,9.3	6,7	1.40	30420	1.0:2.2	
6	7,9.3	6,7	3.75	81692	1.0:2.7	
D21S11						Two Alleles REVIEW the LSD Results
1	28,28	30.2,30.2	0.0E+00		1.0:1.8	
2	28,28	28,30.2	0.0E+00		1.0:-4.6	
3	30.2,30.2	28,30.2	0.0E+00		1.0:2.6	
4	28,30.2	28,30.2	0.29		1.0:1.0	
D18S51						Four Alleles
1	16,20	17,18	5.7E-02	1.0	1.0:2.1	
2	16,18	17,20	1.52	27	1.0:1.1	
3	16,17	18,20	1.58	28	1.0:1.0	
D5S818						No Alleles
D13S317						No Alleles
D7S820						No Alleles

Figure 5.3: Output of web-LSD for the simulated DNA mixture data with a mass ratio of 1:2. All possible genotype combinations are ranked at each locus according to the magnitudes of the fitting errors.

D16S539							No Alleles
CSF1PO							No Alleles
VWA							Three Alleles
1	15,15	16,19	1.2E-04	1.0	1.0:1.8		
2	19,19	15,16	5.2E-03	43	1.0:2.1		
3	16,16	15,19	6.9E-03	57	1.0:2.1		
4	15,19	15,16	0.27	2217	1.0:1.0		
5	16,19	15,16	0.40	3334	1.0:1.2		
6	16,19	15,19	0.43	3525	1.0:1.2		
D8S1179							Two Alleles REVIEW the LSD Results
1	11,11	13,13	0.0E+00		1.0:1.1		
2	11,11	11,13	0.0E+00		1.0:-25.5		
3	13,13	11,13	0.0E+00		1.0:23.5		
4	11,13	11,13	3.6E-03		1.0:1.0		
TPOX							No Alleles
FGA							Two Alleles REVIEW the LSD Results
1	25,25	22,22	0.0E+00		1.0:2.1		
2	22,22	22,25	0.0E+00		1.0:1.8		
3	25,25	22,25	0.0E+00		1.0:-3.8		
4	22,25	22,25	0.60		1.0:1.0		
PentaD							No Alleles
PentaE							No Alleles
D19S433							No Alleles
D2S1338							No Alleles
SE33							No Alleles
Tue 09-Jul-2002 01:04 PM							

Figure 5.3: Continued.

Table 5.6. For a 2-allele locus, when the mass ratio of the DNA is close to 1:2, and the ratio of the two allele peak areas is also close to 2:1 (or 1:2), two genotype combinations of {[B, B] [A, A]} and {[A, A] [A, B]} will both generate two allele peaks with the ratio of the peak A to peak B area being close to 2:1. Therefore, LSD guided results generate two best-fit genotype combinations for each of the two loci. This is a limitation of LSD imposed by the underlying mathematics, which will be presented in detail in Section 5.3. In this case, one can make the decision as to which one is the best choice, after the genotype of the victim is brought in for reference. In Table 5.5, the true genotypes of the two contributors are indicated as shaded cells.

Table 5.6: Ranking result at the two-allele locus FGA in the simulated DNA mixture data shown in Table 5.2, with a mass ratio of 1:2. Since the ratio of the allele peak areas is also close to 1:2, there are two genotype combinations indicated as shaded cells, both of which can fit the observed allele peak data equally well with the calculated mass ratios consistent with those for other loci.

Case	Rank	Genotype Combinations		Matrix A	Errors	Mass Ratio Calculated
		Person 1	Person 2			
1	1	25, 25	22, 22	$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$	0	1: 2.1
2	1	22, 22	22, 25	$\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$	0	1: 1.8
3	1	25, 25	22, 25	$\begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}$	0	1: -3.8
4	2	22, 25	22, 25	$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$	0.60	1:1

5.2.3. LSD Interpretation Result Using Simulated DNA Mixture Data with a Mass Ratio of 1:1

LSD suggested genotype resolution result using simulated DNA mixture data with the mass ratio of 1:1 is presented in Table 5.7. Figure 5.4 shows the output of web-LSD for this set of data. Mixtures with mass ratios close to 1:1 are very difficult to resolve, because too many genotype combinations would fit equally well to the given peak data, and the slightest peak imbalance would change the best-fit genotype combination pair. For DNA mixture data with the mass ratio close to 1:1, one cannot differentiate the major and minor contributors clearly even at a 4-allele locus. As a result, the LSD algorithm would yield multiple genotype combinations, all with comparable fit. Therefore, the LSD suggested genotype resolution result is not as clear compared with those of other sets of DNA mixture data with mass ratios other than 1:1. From Table 5.7, one can see that LSD generated at least two possible choices for each of the six loci; the true cases are indicated with shaded cells. However, when a reference profile (such as the victim's) is brought in, the right combination pair can usually be selected from the set of possible pairs. It is crucial to have a reference profile available when the underlying mass ratio is close to 1:1.

For loci VWA and TH01, each locus has three detected alleles with the ratio of the allele peak areas being close to 2:1:1. Two genotype combinations of {[A, A] [B, C]} and {[A, B] [A, C]} are possible: both would fit equally well to the peak areas ratio if the mass ratio of the DNA mixture is close to 1:1. Since there is no differentiation between the major and minor contributor, LSD suggests four possible allele assignment patterns

Table 5.7: LSD suggested genotype resolution result of a set of simulated DNA mixture data shown in Table 5.2, with a mass ratio of 1:1. Because of the mass ratio being close to 1:1, all six loci were not able to be resolved with confidence. Each locus has at least two possible choices which will fit the given peak data. The shaded cells indicate the true genotypes for the two contributors.

Locus	Alleles in the Mixture	Peak Area Ratio Measured	LSD Suggested Genotype Resolution Result			
			Person 1	Person 2	Mass Ratio Calculated	Remarks
VWA	15, 16, 19	2.02: 1: 1.05	15, 16	15, 19	1:1.05	With equal confidence
			15, 19	15, 16	1.05:1	With equal confidence
			15, 15	16, 19	1:1.02	With equal confidence
			16, 19	15, 15	1.02:1	With equal confidence
TH01	6, 7, 9.3	1.96: 1: 1.02	6, 6	7, 9.3	1:1.03	With equal confidence
			7, 9.3	6, 6	1.03:1	With equal confidence
			6, 7	6, 9.3	1:1.02	With equal confidence
			6, 9.3	6, 7	1.02:1	With equal confidence
D8S1179	11, 13	1: 1.13	11, 11	13, 13	1:1.13	With equal confidence
			13, 13	11, 11	1.13:1	With equal confidence
			11, 13	11, 13	Indeterminate	With equal confidence
FGA	22, 25	3.52: 1	22, 25	22, 22	1:1.26	With equal confidence
			22, 22	22, 25	1.26:1	With equal confidence
D21S11	28, 30.2	1.15: 1	30.2, 30.2	28, 28	1:1.15	With equal confidence
			28, 28	30.2, 30.2	1.15:1	With equal confidence
			28, 30.2	28, 30.2	Indeterminate	With equal confidence
D18S51	16, 17, 18, 20	1.12: 1.01: 1.03: 1	17, 20	16, 18	1:1.07	With equal confidence
			16, 18	17, 20	1.07:1	With equal confidence
			18, 20	16, 17	1:1.05	With equal confidence
			16, 17	18, 20	1.05:1	With equal confidence
			17, 18	16, 20	1:1.04	With equal confidence
			16, 20	17, 18	1.04:1	With equal confidence

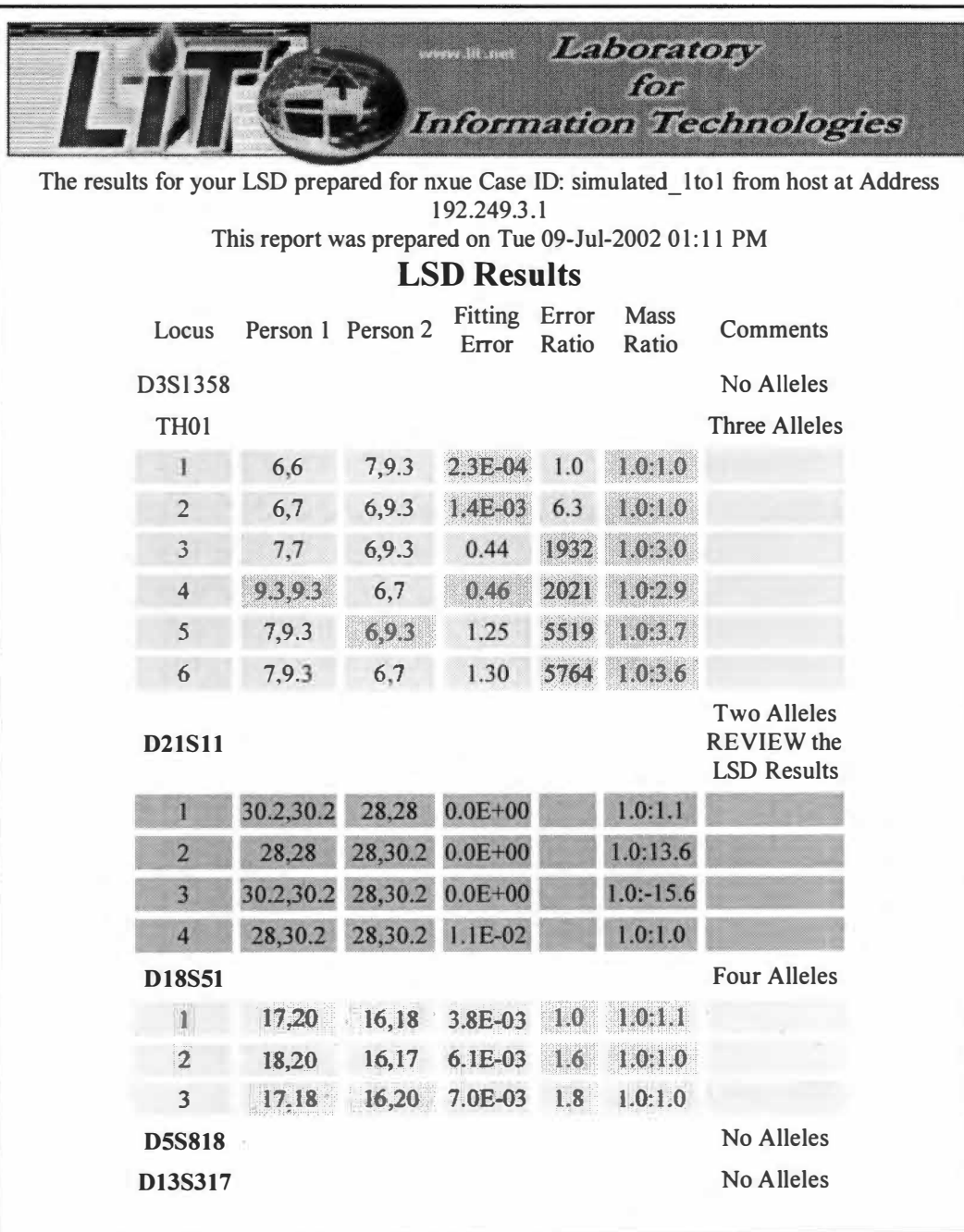


Figure 5.4: Output of web-LSD for the simulated DNA mixture data with a mass ratio of 1:1. All possible genotype combinations are ranked at each locus according to the magnitudes of the fitting errors.

D7S820

No Alleles

D16S539

No Alleles

CSF1PO

No Alleles

VWA

Three Alleles

1	15,16	15,19	4.6E-04	1.0	1.0:1.1	
2	15,15	16,19	1.4E-03	3.1	1.0:1.0	
3	16,16	15,19	0.46	1017	1.0:3.1	
4	19,19	15,16	0.52	1133	1.0:2.9	
5	16,19	15,19	1.28	2817	1.0:3.9	
6	16,19	15,16	1.43	3132	1.0:3.6	

D8S1179

Two Alleles
REVIEW the
LSD Results

1	11,11	13,13	0.0E+00		1.0:1.1	
2	11,11	11,13	0.0E+00		1.0:-16.9	
3	13,13	11,13	0.0E+00		1.0:14.9	
4	11,13	11,13	9.0E-03		1.0:1.0	

TPOX

No Alleles

FGA

Two Alleles
REVIEW the
LSD Results

1	25,25	22,22	0.0E+00		1.0:3.5	
2	22,25	22,22	0.0E+00		1.0:1.3	
3	25,25	22,25	0.0E+00		1.0:-2.8	
4	22,25	22,25	3.17		1.0:1.0	

PentaD

No Alleles

PentaE

No Alleles

D19S433

No Alleles

D2S1338

No Alleles

SE33

No Alleles

Tue 09-Jul-2002 01:11 PM

Figure 5.4: Continued.

for each of the two loci of VWA and TH01, as shown in Table 5.7, not being able to consistently distinguish between person 1 and person 2's profiles.

For loci D8S1179 and D21S11, each locus has two alleles with the ratio of the allele peak data close to 1:1. Since the mass ratio in the mixed DNA sample is also close to 1:1, there are three possible genotype combinations, {[A, B] [A, B]}, {[A, A] [B, B]}, and {[B, B] [A, A]}, which all give comparable allele peak signals. As a result, all three possibilities need to be considered. If the victim's profile is known, then the correct choice can usually be identified.

Since the peak area of the two alleles present at locus FGA are not comparable, the genotype combination of {[22, 25] [22, 25]} can be excluded directly. By comparing the mass ratios calculated from all possible genotype combinations at this locus with those estimated from other loci, the allele assignment pattern of {[22, 25] [22, 22]} was selected as the best-fit one.

For locus D18S51 with four alleles all with comparable peak areas, all six possible genotype combinations are ranked equally because the mass ratio in this DNA mixture sample is around 1:1. The true one is shown in the shaded cells. Since there is always some noise or imbalance associated with the allele peak quantitation, in this case, LSD guided result would be very sensitive to the relative amount of noise present in the respective allele signals. Therefore, a reference genotype becomes indispensable in assigning a decisive genotype combination pair to the two contributors.

5.3. Some Limitations of LSD Algorithm

Compared with other methodologies in resolving mixed DNA samples LSD has many advantages. Although LSD is direct and elegant, it does have limitations for some special cases, due to the nature of the mathematics behind it. From the LSD suggested genotype resolution results of the simulated DNA mixture data presented in Section 5.2, the limitations of LSD can be summarized as follows:

1. For all loci regardless of the number of detected alleles:

If the mass ratio in a DNA mixture is close to 1:1 and the genotype of both contributors are unknown, then one cannot consistently assign the genotype of each locus that should be assigned to person 1 and all that should be assigned to person 2 across all loci. Therefore, resulting in possible reversed assignment at some loci. An example of this case is shown in Table 5.8. Both of these two genotype assignments are equally possible; one cannot choose one over the other with a greater confidence.

2. For loci with four observed alleles:

If the mass ratio is close to 1:1, no decisive assignment can be made unless the genotype of one contributor is known with certainty.

Table 5.8: Example of a possible reversed genotype assignment for the two contributors at a four-allele locus with mass ratio close to 1:1.

Genotype Assignment Pattern 1			Genotype Assignment Pattern 2		
Person 1	Person 2	Mass Ratio	Person 1	Person 2	Mass Ratio
A, B	A, C	1:1.03	A, C	A, B	1.03:1

Note: If the genotype combination of {[A, B] [A, C]} is selected as the best fit by using LSD algorithm and the estimated mass ratio of this DNA mixture sample is close to 1:1, then it is relatively difficult to differentiate the major contributor from the minor contributor. Therefore, both of these two allele assignment patterns are possible.

3. For loci with three observed alleles:

- If the mass ratio is close to 1:1 predicted from other loci and the ratio of the allele peak areas is close to 2:1:1, there are two possible genotype combinations that would both fit the measured relative allele peak areas, as presented in Table 5.9. If the genotype of the victim is known, it will be helpful to decide which one of these two to choose.

- If the mass ratio, predicted from other loci, is close to 1:2 and the ratio of the three allele peak areas is close to 1:1:1, then there are three possible genotype combinations, all of which would fit the observed allele peak data equally well, as shown in Table 5.10. One cannot make a decision as to which one to choose unless the genotype of the victim is clearly known at this locus.

4. For loci with two observed alleles, two cases need special consideration and interpretation:

a. If the ratio of the two allele peak areas is close to 1:1

- If the approximate mass ratio estimated from the other loci with more than two alleles is also close to 1:1, then there are two possible allele assignment patterns, which would fit equally well to the peak area ratio of 1:1, as shown in Table 5.11. A clear determination of the genotype assignment pattern cannot be made unless the genotype of one contributor is clearly known.

- If the approximate mass ratio estimated from other loci with more than two alleles is not close to 1:1, then there is only one plausible combination of {[A, B] [A, B]}, which would fit the peak area ratio of 1:1. Therefore, the two contributors must have the same genotype [A, B] at this locus.

Table 5.9: Two possible genotype combinations at a three-allele locus with the ratio of the allele peak areas of 2:1:1, and the mass ratio of the DNA mixture is close to 1:1, as predicted from other loci.

Possible Genotype Combination 1		Possible Genotype Combination 2	
Person 1	Person 2	Person 1	Person 2
A, A	B, C	A, B	A, C
Ratio of the Two Allele Peak Areas		2:1:1	
Approximate Mass Ratio predicted from Other Loci		1:1	

Table 5.10: Three possible genotype combinations at a three-allele locus with the ratio of the allele peak areas of 1:1:1, and the mass ratio of the DNA mixture is close to 1:2, as predicted from other loci.

Possible Genotype Combination 1		Possible Genotype Combination 2		Possible Genotype Combination 3	
Person 1	Person 2	Person 1	Person 2	Person 1	Person 2
A, A	B, C	B, B	A, C	C, C	A, B
Ratio of the Two Allele Peak Areas				1:1:1	
Approximate Mass Ratio predicted from Other Loci				1:2	

Table 5.11: Two possible genotype combinations at a two-allele locus with the peak area ratio of 1:1, and the mass ratio of the DNA mixture is also close to 1:1, as predicted from other loci.

Genotype Combination 1		Genotype Combination 2	
Person 1	Person 2	Person 1	Person 2
A, A	B, B	A, B	A, B
Ratio of the Two Allele Peak Areas		1:1	
Approximate Mass Ratio Predicted from Other Loci		1:1	

b. If the ratio of the two allele peak areas is close to 2:1(or 2:1) and the approximate mass ratio decided from other loci is also close to 2:1(or 2:1), then there are two possible genotype assignment patterns, as presented in Table 5.12. Both of these will rank first with zero errors, and both are equally likely to be the real case. As a result, no clear selection can be made, if the genotype profiles of both contributors are unknown.

In the absence of the special cases presented above, as long as there is no peak area inversion (Peak area inversion means that the relative allele peak areas are not in accordance with the corresponding mass portions), LSD should most likely yield the correct genotypes for the two contributors.

5.4. LSD Interpretation Results of DNA Mixture Data from Literature Reports

After LSD methodology was formulated and its limitations were understood, two sets of DNA mixture data obtained from literature reports were analyzed using LSD. Since both contributors' genotypes are known, we can evaluate the workings of LSD by

Table 5.12: Special interpretation rule for a 2-allele locus, when both the mass ratio and the peak area ratio are close to 1:2. Two possible genotype combination profiles are equally likely in fitting the peak area data.

Genotype Combination 1		Genotype Combination 2	
Person 1	Person 2	Person 1	Person 2
B, B	A, A	A, A	A, B
Ratio of the Two Allele Peak Areas		1:2	
Approximate Mass Ratio Predicted from Other Loci		2:1	

comparing the LSD suggested genotype resolution results with the true genotypes of both contributors.

One set of data is from Reference [3]: “Taking Account of Peak Areas When Interpreting Mixed DNA Profiles,” by T. Clayton, J. Whitaker, R. Sparkes, and P. Gill. A set of DNA mixture data at six loci was presented in that paper, as summarized in Table 5.13. LSD suggested genotype resolution result is shown in Table 5.14. Figure 5.5 shows the output of web-LSD for this set of data. By comparing the LSD suggested result with the true genotypes of both contributors, one can see that among the six loci, the top-ranked genotype combinations selected are consistent with that of the correct one at five loci (all but TH01). From the result at these five loci, one can see that the approximate mass ratio of the DNA mixture is around 1:2. Locus TH01 has only two alleles and post-LSD review is always required. LSD result yields two possible allele assignment patterns (both of which have zero fitting residuals) with the calculated mass ratios consistent with those estimated from the other five loci. This result is consistent with LSD limitation 4(b), as discussed in Section 5.3. If the DNA analyst does not know the genotype of either contributor, then she cannot make the decision as to which one to choose.

The other set of DNA mixture sample is obtained from Reference [21]: “Taking Account of Peak Areas When Interpreting Mixed DNA Profiles,” by I. Evett, P. Gill, and J. Lambert. The genotype of both contributors and the mass ratio of this DNA mixture sample, prior to PCR amplification, were known, making it possible to evaluate the LSD suggested genotype resolution result. All available information about this set of DNA

Table 5.13: Allele peak data and the genotypes for the two contributors at six loci from a mixed DNA sample presented in Reference [3].

Locus	Alleles in the Mixture	Genotype Combination		Allele Peak Areas
		Person 1	Person 2	
VWA	15 16 17 19	17, 19	15, 16	1247 1193 2279 2000
TH01	5 7	5, 7	7, 7	5735 10769
D8S1179	13 14 15	13, 13	14, 15	3234 752 894
FGA	22 23	23, 23	22, 23	534 2792
D21S11	61 65 70 77	65, 70	61, 77	373 590 615 356
D18S51	14 15 16 18	16, 18	14, 15	1339 1465 2895 2288

Table 5.14: LSD suggested genotype resolution result of a mixed DNA sample presented in Reference [3]. All six loci were correctly resolved by selecting the top-ranked genotype combinations with consistent mass ratios. There are two possible choices at locus TH01; the shaded cells indicate the true genotype combination for the two contributors.

Locus	Alleles in the Mixture	Peak Area Ratio Measured	LSD Suggested Genotype Resolution Result			
			Person 1	Person 2	Mass Ratio Calculated	Remarks
VWA	15 16 17 19	1.05: 1: 1.91: 1.68	15, 16	17, 19	1: 1.75	Correct
TH01	5 7	1: 1.88	7, 7	5, 7	1: 2.28	With equal confidence
			5, 5	7, 7	1: 1.88	With equal confidence
D8S1179	13 14 15	4.30: 1: 1.19	14, 15	13, 13	1: 1.96	Correct
FGA	22 23	1: 5.23	22, 23	23, 23	1: 2.11	Correct
D21S11	61 65 70 77	1.05: 1.66: 1.73: 1	61, 77	65, 70	1: 1.65	Correct
D18S51	14 15 16 18	1: 1.09: 2.16: 1.71	14, 15	16, 18	1: 1.85	Correct

The results for your LSD prepared for nxue Case ID: data_from_reference3 from host at Address
192.249.3.1

This report was prepared on Fri 12-Jul-2002 09:51 AM

LSD Results

Locus	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio	Comments
D3S1358						No Alleles
TH01						Two Alleles REVIEW the LSD Results
1	5,5	7,7	0.0E+00		1.0:1.9	
2	5,5	5,7	0.0E+00		1.0:-4.3	
3	7,7	5,7	0.0E+00		1.0:2.3	
4	5,7	5,7	0.39		1.0:1.0	
D21S11						Four Alleles
1	61,77	65,70	3.6E-03	1.0	1.0:1.7	
2	65,77	61,70	0.45	124	1.0:1.0	
3	61,65	70,77	0.45	125	1.0:1.0	
D18S51						Four Alleles
1	14,15	16,18	0.11	1.0	1.0:1.8	
2	14,18	15,16	0.82	7.7	1.0:1.2	
3	15,18	14,16	0.86	8.1	1.0:1.1	
D5S818						No Alleles
D13S317						No Alleles
D7S820						No Alleles
D16S539						No Alleles

Figure 5.5: Output of web-LSD for the DNA mixture data presented in Reference [3]. All possible genotype combinations are ranked at each locus according to the magnitudes of the fitting errors.

CSF1PO						No Alleles
VWA						Four Alleles
1	15,16	17,19	2.8E-02	1.0	1.0:1.8	
2	16,19	15,17	0.60	21	1.0:1.1	
3	15,19	16,17	0.61	22	1.0:1.1	
D8S1179						Three Alleles
1	14,15	13,13	1.8E-02	1.0	1.0:2.0	
2	13,14	13,15	1.49	83	1.0:1.1	
3	14,14	13,15	4.84	272	1.0:5.5	
4	15,15	13,14	5.45	306	1.0:4.5	
5	14,15	13,15	5.64	316	1.0:-7.9	
6	14,15	13,14	6.72	377	1.0:-9.1	
TPOX						No Alleles
FGA						Two Alleles REVIEW the LSD Results
1	22,22	23,23	0.0E+00		1.0:5.2	
2	22,22	22,23	0.0E+00		1.0:-2.5	
3	22,23	23,23	0.0E+00		1.0:2.1	
4	22,23	22,23	8.94		1.0:1.0	
PentaD						No Alleles
PentaE						No Alleles
D19S433						No Alleles
D2S1338						No Alleles
SE33						No Alleles
Fri 12-Jul-2002 09:51 AM						

Figure 5.5: Continued.

mixture data is presented in Table 5.15; the top-ranked genotype combination at each locus from LSD ranking result is shown in Table 5.16. Figure 5.6 shows the output of web-LSD for this set of DNA mixture data. Although the six best-fit mass ratios calculated from the six loci are not exactly consistent, ranging from 1: 4.8 (FGA) to 1:11.4 (D8S1179), this mixed DNA sample can still be resolved confidently. At each of the six loci, the genotypes of Person 1 and Person 2 correspond to the true genotypes of victim and the suspect, respectively. As a result of the noise and inconsistent PCR amplification among the six loci, there is some difference between the calculated average mass ratio of 1: 8.04 and the known one prior to amplification in this DNA mixture of 1:10. However, such a difference does not affect the resolution result of the genotypes for the two contributors.

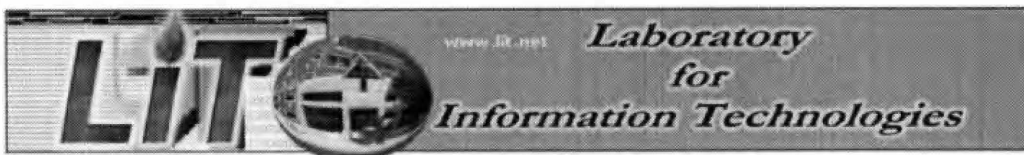
From the above interpretation, one can see that both of these two sets of DNA mixture data from literature reports were correctly resolved using LSD by selecting the top-ranked genotype combination with consistent mass ratio at each locus. No information about the genotype of any contributor and the mass ratio in the mixture data is necessary for LSD to resolve them. Such information was only used to verify the LSD suggested genotype resolution results. In Reference [3] and [21], where these two sets of data come from, at least one contributor's genotype is necessary to resolve the DNA mixture. Moreover, the method presented in those two references to resolve DNA mixture is not as direct as LSD.

Table 5.15: Allele peak data and the genotypes of the two contributors at six loci of a mixed DNA sample with the mass ratio of 1:10 prior to PCR amplification, as presented in Reference [21].

Locus	Alleles in the Mixture	Genotype Combination		Allele Peak Areas
		Suspect	Victim	
TH01	8 9.3	8, 9.3	9.3, 9.3	17441 22368
D21S11	59 65 67 70	67, 70	59, 65	1226 1434 8816 8894
D18S51	13 16 17	13, 13	16, 17	38985 1914 1991
D8S1179	10 11 14	10, 14	10, 11	6416 383 5659
VWA	16 17 18 19	16, 18	17, 19	4669 931 4724 188
FGA	21 22 23	21, 22	21, 23	16099 10538 1014

Table 5.16: LSD suggested genotype resolution result of a mixed DNA sample obtained from Reference [21]. The top-ranked genotype combinations at all loci are consistent with the true genotypes of the two contributors, which are known and used as references.

Locus	Alleles in the Mixture	Peak Area Ratio Measured	LSD Suggested Genotype Resolution Result			
			Person 1	Person 2	Mass Ratio Calculated	Remarks
TH01	8 9.3	1: 1.28	9.3, 9.3	8, 9.3	1: 7.08	With confidence
D21S11	59 65 67 70	1: 1.17: 7.19: 7.25	59, 65	67, 70	1: 6.66	With confidence
D18S51	13 16 17	20.37: 1: 1.04	16, 17	13, 13	1: 9.98	With confidence
D8S1179	10 11 14	16.75: 1: 1.476	10, 11	10, 14	1: 11.39	With confidence
VWA	16 17 18 19	24.84: 4.95: 25.13: 1	17, 19	16, 18	1: 8.39	With confidence
FGA	21 22 23	15.88: 10.39: 1	21, 23	21, 22	1: 4.76	With confidence



The results for your LSD prepared for nxue Case ID: data_from_reference21 from host at Address 192.249.3.1

This report was prepared on Fri 12-Jul-2002 09:58 AM

LSD Results

Locus	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio	Comments
D3S1358						No Alleles
TH01						Two Alleles REVIEW the LSD Results
1	8,8	9,3,9,3	0.0E+00		1.0:1.3	
2	8,8	8,9,3	0.0E+00		1.0:-9.1	
3	9,3,9,3	8,9,3	0.0E+00		1.0:7.1	
4	8,9,3	8,9,3	4.0E-02		1.0:1.0	
D21S11						Four Alleles
1	59,65	67,70	1.6E-02	1.0	1.0:6.7	
2	59,67	65,70	37.68	2295	1.0:1.0	
3	59,70	65,67	37.69	2296	1.0:1.0	
D18S51						Three Alleles
1	16,17	13,13	8.1E-04	1.0	1.0:10.0	
2	13,16	13,17	111.97	138371	1.0:1.0	
3	16,17	13,17	137.74	170218	1.0:-2.4	
4	16,17	13,16	138.84	171568	1.0:-2.4	
5	16,16	13,17	186.79	230824	1.0:21.4	
6	17,17	13,16	187.57	231786	1.0:20.5	
D5S818						No Alleles

Figure 5.6: Output of web-LSD for the DNA mixture data presented in Reference [21]. All possible genotype combinations are ranked at each locus according to the magnitudes of the fitting errors.

D13S317						No Alleles
D7S820						No Alleles
D16S539						No Alleles
CSF1PO						No Alleles
VWA						Four Alleles
1	17,19	16,18	7.85	1.0	1.0:8.4	
2	16,19	17,18	487.58	62	1.0:1.2	
3	18,19	16,17	488.74	62	1.0:1.1	
D8S1179						Three Alleles
1	10,11	10,14	0.32	1.0	1.0:11.4	
2	11,11	10,14	1.95	6.1	1.0:31.5	
3	11,14	10,14	2.95	9.3	1.0:2012.0	
4	11,14	10,10	94.88	299	1.0:1.1	
5	14,14	10,11	124.06	390	1.0:1.2	
6	11,14	10,11	310.64	977	1.0:1.4	
TPOX						No Alleles
FGA						Three Alleles
1	21,23	21,22	6.70	1.0	1.0:4.8	
2	22,23	21,22	14.02	2.1	1.0:-11.8	
3	23,23	21,22	15.04	2.2	1.0:26.3	
4	22,23	21,21	44.11	6.6	1.0:1.4	
5	22,22	21,23	110.66	17	1.0:1.6	
6	22,23	21,23	212.84	32	1.0:3.8	
PentaD						No Alleles
PentaE						No Alleles
D19S433						No Alleles
D2S1338						No Alleles
SE33						No Alleles
Fri 12-Jul-2002 09:58 AM						

Figure 5.6: Continued.

5.5. LSD Interpretation Results of Real DNA Mixture Samples

To further test the effectiveness of LSD, DNA mixture samples from real forensic cases were obtained and analyzed. These data were kindly provided by Mr. Wickenheiser, director of the Acadiana Crime Lab in New Iberia, Louisiana, and by Ms. Patricia Graham in the Texas Department of Public Safety in Corpus Christi. To date, these mixed DNA profiles were mostly successfully resolved using LSD. LSD suggested genotype resolution results for six sets of real DNA mixture data were selected and presented in this thesis. One of the six sets has already been presented and discussed in Chapter 4; the other five will be presented now in this chapter to illustrate the variety of sample data that can be encountered.

5.5.1. LSD Interpretation Results of DNA Mixture Samples from the Acadiana Crime Lab

Five sets of real DNA mixture data were provided by the Acadiana Crime Lab. The allele peaks of each mixed DNA sample were scanned and measured by an ABI Prism 310 Capillary Electrophoresis Gene Sequencer using Genescan and Genotyper software. Mr. Wickenheiser suggested that if the peak area measured is larger than 4000 units, the linear relationship between the peak area and DNA quantity is most probably compromised due to DNA loading saturation. Each set of data contains the allele peak areas measured at all 13 CODIS core loci. DNA analysts in the Acadiana Crime Lab know the genotypes of both contributors in each case. The objective for us is to see if LSD would effectively resolve the mixture samples.

After applying LSD to each of the five sets of data, a composite profile for each data sample was obtained by identifying the top-ranked genotype combination for each locus and choosing those with consistent mass ratios, when a choice is to be made. Among the five DNA mixture sample profiles, LSD guided results for four data sets are consistent with the true genotype profiles of the two contributors across all loci. For the fifth set of data, due to the peak area saturation, the top-ranked genotype combinations in LSD ranking result did not correspond to the correct genotype combinations at several loci. This was due to the non-preservation of the proportionality between the measured allele peak signals and the relative DNA mass amount. For some loci among the samples, LSD guided results suggest more than one possible genotype combinations for the two contributors, all of which are equally likely to be the true case. A choice is made only after the reference profiles at those loci were brought in to ascertain the correct one. In this chapter, three sets of allele peak data of DNA mixture samples provided by the Acadiana Crime Lab and the corresponding LSD interpretation results are presented. Each set of data is selected for a different purpose:

- Acadiana DNA mixture data No. 5 shows that LSD algorithm can yield correct genotypes for both contributors at all thirteen loci, in the absence of DNA loading saturation and mass ratio away from 1:1;
- Acadiana DNA mixture data No. 1 shows the effects of peak area saturation;
- Acadiana DNA mixture data No. 2 illustrates the limitations of LSD when the approximate mass ratio in the DNA mixture is close to 1:1.

5.5.1.1. LSD Interpretation Result of Acadiana DNA Mixture Data No. 5

Acadiana DNA mixture data No. 5 is presented in Table 5.17. Figure 5.7 shows the output of the web-LSD for this set of DNA mixture data. The corresponding LSD suggested genotype resolution result at all thirteen loci is summarized in Table 5.18. Among the total thirteen loci, twelve of them can be resolved unambiguously, which means LSD suggests only one possible genotype combination profile for the two contributors at each locus. For each of these twelve loci, the only possible genotype combination in LSD guided result corresponds to the true genotype profiles of the victim and the suspect. For the locus D3S1358, D21S11, D18S51, D5S818, D13S317, D7S820, D16S539, and CSF1PO, because the top-ranked genotype combination in each case has a fitting error so much smaller than that of the second ranked one, and the associated mass ratios are, more or less, consistent with each other at these twelve loci with the average being 1: 6.26. For locus FGA, since only one allele is observed, it can be concluded that both the victim and the perpetrator has the genotype of [22, 22] at this locus. At the locus VWA, D8S1179, TH01, and TPOX, each locus has two observed alleles with the ratio of the two peak areas close to 1:1, since the mass ratio of the DNA mixture sample estimated from other loci is not close to 1:1, according to the limitation of LSD 4(a) (Section 5.3), there is only one possible genotype combination for each of these four loci as shown in Table 5.18. For locus D21S11, LSD algorithm alludes to two possible choices, because the associated fitting errors are comparable. Detailed LSD ranking result for this locus is shown in Table 5.19. From the ranking result, one can see that both of the top two ranked genotype combinations can fit the measured allele peak data

Table 5.17: Allele peak data and the genotypes of both contributors at thirteen loci in Acadiana DNA Mixture Data No. 5.

Locus	Alleles in the Mixture	Allele Peak Heights Measured	Contributors' Genotypes	
			Suspect	Victim
D3S1358	14 16 17 19	616 688 3592 3479	14, 16	17, 19
VWA	14 16	3483 3534	14, 16	14, 16
FGA	22	5503	22, 22	22, 22
D8S1179	13 14	3111 3175	13, 14	13, 14
D21S11	28 29 30	3166 489 2964	29, 29	28, 30
D18S51	13 14 15 18	2383 380 2041 271	14, 18	13, 15
D5S818	11 12 13	538 2448 2240	11, 11	12, 13
D13S317	11 12 13 14	309 1892 366 1734	11, 13	12, 14
D7S820	8 10 12	1685 1409 144	8, 12	8, 10
D16S539	11 12 13 14	3047 460 562 2329	12, 13	11, 14
TH01	9 9.3	2276 2666	9, 9.3	9, 9.3
TPOX	8 11	2556 2447	8, 11	8, 11
CSF1PO	10 11 12	1693 339 2164	11, 12	10, 12



The results for your LSD prepared for nxue Case ID: Acadiana_data_Page6 from host at
Address 192.249.3.1

This report was prepared on Thu 30-May-2002 04:31 PM

LSD Results

Locus	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio	Comments
D3S1358						Four Alleles
1	14,16	17,19	2.4E-02	1.0	1.0:5.4	
2	14,19	16,17	21.91	926	1.0:1.0	
3	16,19	14,17	21.93	927	1.0:1.0	
TH01						Two Alleles REVIEW the LSD Results
1	9,9	9.3,9.3	0.0E+00		1.0:1.2	
2	9,9	9,9.3	0.0E+00		1.0:-13.7	
3	9.3,9.3	9,9.3	0.0E+00		1.0:11.7	
4	9,9.3	9,9.3	1.5E-02		1.0:1.0	
D21S11						Three Alleles
1	29,29	28,30	8.5E-02	1.0	1.0:12.5	
2	28,29	28,30	0.11	1.3	1.0:7.3	
3	29,30	28,30	0.67	7.8	1.0:11.3	
4	28,28	29,30	12.81	150	1.0:1.1	
5	30,30	28,29	14.98	176	1.0:1.2	
6	29,30	28,29	44.36	520	1.0:1.2	

Figure 5.7: Output of web-LSD for Acadiana DNA Mixture data No. 5. All possible genotype combinations are ranked at each locus according to the magnitudes of the fitting errors.

D18S51						Four Alleles
1	14,18	13,15	0.88	1.0	1.0:6.8	
2	15,18	13,14	48.64	55	1.0:1.2	
3	14,15	13,18	49.15	56	1.0:1.1	
D5S818						Three Alleles
1	11,11	12,13	7.5E-02	1.0	1.0:8.7	
2	11,12	12,13	0.13	1.7	1.0:5.0	
3	11,13	12,13	0.64	8.6	1.0:7.6	
4	12,12	11,13	5.00	67	1.0:1.1	
5	13,13	11,12	6.30	84	1.0:1.3	
6	11,13	11,12	19.83	265	1.0:1.2	
D13S317						Four Alleles
1	11,13	12,14	0.15	1.0	1.0:5.4	
2	11,14	12,13	22.83	155	1.0:1.1	
3	13,14	11,12	22.92	155	1.0:1.0	
D7S820						Three Alleles
1	8,12	8,10	0.28	1.0	1.0:7.7	
2	12,12	8,10	1.84	6.6	1.0:21.5	
3	10,12	8,10	2.84	10	1.0:386.2	
4	10,12	8,8	38.59	138	1.0:1.1	
5	10,10	8,12	57.26	204	1.0:1.3	
6	10,12	8,12	139.89	499	1.0:1.6	

Figure 5.7: Continued.

D16S539

Four Alleles

1	12,13	11,14	1.24	1.0	1.0:5.3	
2	12,14	11,13	22.85	18	1.0:1.3	
3	13,14	11,12	23.19	19	1.0:1.2	

CSF1PO

Three Alleles

1	11,12	10,12	5.1E-02	1.0	1.0:4.5	
2	11,11	10,12	0.97	19	1.0:11.4	
3	10,11	10,12	1.90	38	1.0:27.4	
4	10,11	12,12	7.98	158	1.0:1.1	
5	10,10	11,12	14.49	287	1.0:1.5	
6	10,11	11,12	35.90	710	1.0:1.9	

VWA

Two Alleles
REVIEW the
LSD Results

1	14,14	16,16	0.0E+00		1.0:1.0	
2	14,14	14,16	0.0E+00		1.0:-138.6	
3	16,16	14,16	0.0E+00		1.0:136.6	
4	14,16	14,16	1.1E-04		1.0:1.0	

Figure 5.7: Continued.

D8S1179						Two Alleles REVIEW the LSD Results
1	13,13	14,14	0.0E+00		1.0:1.0	
2	13,13	13,14	0.0E+00		1.0:- 99.2	
3	14,14	13,14	0.0E+00		1.0:97.2	
4	13,14	13,14	2.1E-04		1.0:1.0	
TPOX						Two Alleles REVIEW the LSD Results
1	11,11	8,8	0.0E+00		1.0:1.0	
2	8,8	8,11	0.0E+00		1.0:44.9	
3	11,11	8,11	0.0E+00		1.0:- 46.9	
4	8,11	8,11	9.9E-04		1.0:1.0	
FGA						One Allele Only
1	22,22	22,22			1:1	

Thu 30-May-2002 04:31 PM

Figure 5.7: Continued.

Table 5.18: LSD suggested genotype resolution result for Acadiana DNA Mixture Data No. 5. The top-ranked LSD choices correspond to the true genotype combinations of the two contributors at all but the D21S11 locus. LSD suggests two possible choices at locus D21S11; the true one is indicated in the shaded cells.

Locus	Alleles	Measured Peak Area Ratio	LSD Suggested Genotype Resolution Result			
			Person 1 (Suspect)	Person 2 (Victim)	Mass Ratio Calculated	Remarks
D3S1385	14 16 17 19	1: 1.12: 5.83: 5.64	14, 16	17, 19	1: 5.42	Correct
VWA	14 16	1: 1.01	14, 16	14, 16	Indeterminate	Limitation of LSD 4(a)
FGA	22	-	22, 22	22, 22	Indeterminate	One-allele case
D8S1179	13 14	1: 1.02	13, 14	13, 14	Indeterminate	Limitation of LSD 4(a)
D21S11	28 29 30	6.47: 1: 6.06	29, 29	28, 30	1: 12.54	Correct
			28, 29	28, 30	1: 7.29	
D18S51	13 141 5 18	8.79: 1.40: 7.53: 1	14, 18	13, 15	1: 6.80	Correct
D5S818	11 12 13	1: 4.55: 4.16	11, 11	12, 13	1: 8.71	Correct
D13S317	11 12 13 14	1: 6.12: 1.18: 5.61	11, 13	12, 14	1: 5.37	Correct
D7S820	8 10 12	11.70: 9.78: 1	8, 12	8, 10	1: 7.73	Correct
D16S539	11 12 13 14	6.61: 1: 1.22: 5.06	12, 13	11, 14	1: 5.26	Correct
TH01	9 9.3	1: 1.17	9, 9.3	9, 9.3	Indeterminate	Limitation of LSD 4(a)
TPOX	8 11	1.04: 1	8, 11	8, 11	Indeterminate	Limitation of LSD 4(a)
CSF1PO	10 11 12	4.99: 1: 6.38	11, 12	10, 12	1: 4.54	Correct

Note: 'correct' in the table means that the top-ranked genotype choice for each locus in LSD ranking result corresponds to the true genotype profiles of the two contributors.

Table 5.19: Detailed LSD ranking result at locus D21S11 in Acadiana DNA Mixture Data No. 5, in which the fitting errors for the top two ranked combinations are comparable. Therefore, a clearly preferred choice cannot be made. Shaded cells indicate the true genotype combination for the two contributors.

Rank	Genotype Combinations		Ratio of the Errors	Mass Ratio Calculated
	Person 1 (Suspect)	Person 2 (Victim)		
1	29, 29	28, 30	1	1: 12.54
2	28, 29	28, 30	1.35	1: 7.29
3	29, 30	28, 30	7.80	1: 11.35
4	28, 28	29, 30	150.12	1: 1.09
5	30, 30	28, 29	175.63	1: 1.23
6	29, 30	28, 29	519.90	1: 1.19

well, and the mass ratio calculated from the second ranked allele assignment (1:7.29) is more consistent with the average (1:6.26) than that calculated from the top ranked one (1:12.54). Therefore, both genotype combinations have to be considered to be the possible one. After this point, the victim's profile is brought in to compare to that of the LSD suggested result. It turned out that the perpetrator has also been identified and typed, and the first choice with person 1 being [29, 29] is the correct one. The victim is found to be LSD's Person 2. Even with the known victim's genotype of [28, 30] at this locus, one still cannot make the correct choice, because both top-ranked genotype combinations have [28, 30] as the genotype for Person 2. In actual application, we suggest that this locus be deleted from the final composite profile for the perpetrator for any follow-up consideration. The cause that the mass ratio associated with the top-ranked genotype combination is much larger than those from other loci is surely due to the imbalance in the peak areas resulting from imbalanced PCR amplification.

5.5.1.2. LSD Interpretation Result of Acadiana DNA Mixture Data No. 1

Acadiana DNA mixture data set No. 1 and the true genotype profiles of the two contributors are presented in Table 5.20. The output of web-LSD for this set of data is presented in Figure 5.8. The summary of LSD guided result is shown in Table 5.21. Mr. Wickenheiser notified us that peak areas over 4,000 units compromise the linear relationship with DNA quantity, due to peak area saturation. Since many allele peak areas measured for this set of data are larger than 4,000 units, they might not represent the relative mass of such alleles accurately. As a result, only seven out of a total of thirteen loci have the top-ranked genotype combinations in LSD suggested result

Table 5.20: Allele peak data and the true genotypes for both contributors at thirteen loci in Acadiana DNA Mixture Data No. 1.

Locus	Alleles in the Mixture	Allele Peak Heights Measured	Contributors' Genotypes	
			Person 1 (Victim)	Person 2 (Suspect)
D3S1358	14 15 17	5786 1426 5578	15, 17	14, 17
VWA	15 16 19	5899 5890 1169	16, 19	15, 16
FGA	22 24 25	4636 4765 866	25, 25	22, 24
D8S1179	11 13 14 15	6069 1959 4769 1425	13, 15	11, 14
D21S11	28 30.2 32.2	6479 1239 5717	28, 30.2	28, 32.2
D18S51	16 17 20	1323 3076 2748	16, 16	17, 20
D5S818	11 12	5354 1153	11, 12	11, 11
D13S317	8 11 12	5227 1136 5208	11, 12	8, 12
D7S820	10 12	5913 385	10, 12	10, 10
D16S539	8 10 11	7082 6995 1300	10, 11	8, 10
TH01	6 7 9.3	6388 6370 1287	7, 9.3	6, 7
TPOX	8 9 12	1393 7149 1219	8, 12	9, 9
CSF1PO	8 10 11	729 4251 4599	8, 11	10, 11



The results for your LSD prepared for nxue Case ID: Acadiana_data_page2 from host at Address
192.249.3.1

This report was prepared on Thu 30-May-2002 04:42 PM

LSD Results

Locus	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio	Comments
D3S1358						Three Alleles
1	15,15	14,17	1.1E-02	1.0	1.0:8.0	
2	14,15	14,17	0.24	23	1.0:5.1	
3	15,17	14,17	0.44	41	1.0:5.9	
4	14,14	15,17	4.24	398	1.0:1.2	
5	17,17	14,15	4.67	439	1.0:1.3	
6	15,17	14,15	16.19	1522	1.0:1.1	
TH01						Three Alleles
1	9,3,9,3	6,7	9.8E-05	1.0	1.0:9.9	
2	6,9,3	6,7	0.32	3314	1.0:6.9	
3	7,9,3	6,7	0.34	3504	1.0:7.0	
4	6,6	7,9,3	7.80	79744	1.0:1.2	
5	7,7	6,9,3	7.85	80309	1.0:1.2	
6	7,9,3	6,9,3	26.48	270749	1.0:1.0	

Figure 5.8: Output of web-LSD for Acadiana DNA Mixture data No. 1. All possible genotype combinations are ranked at each locus according to the magnitudes of the fitting errors.

D21S11						Three Alleles
1	28,30.2	28,32.2	4.9E-02	1.0	1.0:5.1	
2	30.2,30.2	28,32.2	0.19	3.8	1.0:9.8	
3	30.2,32.2	28,32.2	0.87	18	1.0:10.2	
4	28,28	30.2,32.2	6.53	132	1.0:1.1	
5	32.2,32.2	28,30.2	8.94	181	1.0:1.4	
6	30.2,32.2	28,30.2	26.07	528	1.0:1.4	
D18S51						Three Alleles
1	16,16	17,20	3.1E-02	1.0	1.0:4.4	
2	16,17	17,20	0.19	6.1	1.0:2.4	
3	16,20	17,20	0.52	17	1.0:3.3	
4	17,17	16,20	0.58	19	1.0:1.3	
5	20,20	16,17	0.88	29	1.0:1.6	
6	16,20	16,17	3.86	126	1.0:1.3	
D5S818						Two Alleles REVIEW the LSD Results
1	12,12	11,11	0.0E+00		1.0:4.6	
2	11,12	11,11	0.0E+00		1.0:1.8	
3	12,12	11,12	0.0E+00		1.0:-2.5	
4	11,12	11,12	6.64		1.0:1.0	

Figure 5.8: Continued.

D13S317 Three Alleles

1	11,11	8,12	1.4E-04	1.0	1.0:9.2	
2	8,11	8,12	0.32	2304	1.0:6.3	
3	11,12	8,12	0.34	2464	1.0:6.4	
4	8,8	11,12	6.42	45931	1.0:1.2	
5	12,12	8,11	6.48	46361	1.0:1.2	
6	11,12	8,11	22.34	159689	1.0:1.0	

D7S820 Two Alleles
REVIEW the
LSD Results

1	12,12	10,10	0.0E+00		1.0:15.4	
2	10,12	10,10	0.0E+00		1.0:7.2	
3	12,12	10,12	0.0E+00		1.0:-2.1	
4	10,12	10,12	103.08		1.0:1.0	

D16S539 Three Alleles

1	11,11	8,10	2.2E-03	1.0	1.0:10.8	
2	8,11	8,10	0.29	130	1.0:7.4	
3	10,11	8,10	0.38	169	1.0:7.9	
4	8,8	10,11	9.60	4285	1.0:1.2	
5	10,10	8,11	9.89	4417	1.0:1.2	
6	10,11	8,11	32.20	14379	1.0:1.0	

CSF1PO Three Alleles

1	8,11	10,11	9.1E-02	1.0	1.0:6.9	
2	8,8	10,11	0.11	1.3	1.0:12.1	
3	8,10	10,11	0.73	8.0	1.0:11.5	
4	11,11	8,10	11.67	128	1.0:1.1	
5	10,10	8,11	14.09	155	1.0:1.3	
6	8,10	8,11	41.37	454	1.0:1.2	

Figure 5.8: Continued.

VWA						Three Alleles
1	19,19	15,16	3.0E-05	1.0	1.0:10.1	
2	15,19	15,16	0.33	11075	1.0:7.0	
3	16,19	15,16	0.34	11421	1.0:7.1	
4	15,15	16,19	8.15	275159	1.0:1.2	
5	16,16	15,19	8.19	276209	1.0:1.2	
6	16,19	15,19	27.51	928267	1.0:1.0	
D8S1179						Four Alleles
1	13,15	11,14	0.49	1.0	1.0:3.2	
2	14,15	11,13	6.91	14	1.0:1.3	
3	13,14	11,15	7.25	15	1.0:1.1	
TPOX						Three Alleles
1	8,12	9,9	1.0E-02	1.0	1.0:2.7	
2	9,12	8,9	4.62	453	1.0:1.1	
3	8,12	8,9	10.91	1071	1.0:-4.4	
4	12,12	8,9	11.15	1094	1.0:7.0	
5	8,8	9,12	11.83	1161	1.0:6.0	
6	8,12	9,12	12.03	1181	1.0:-4.5	
FGA						Three Alleles
1	25,25	22,24	1.1E-02	1.0	1.0:10.9	
2	24,25	22,24	0.24	22	1.0:7.1	
3	22,25	22,24	0.44	40	1.0:8.3	
4	24,24	22,25	9.48	854	1.0:1.2	
5	22,22	24,25	10.14	914	1.0:1.2	
6	22,25	24,25	32.38	2918	1.0:1.1	

Thu 30-May-2002 04:42 PM

Figure 5.8: Continued.

Table 5.21: LSD suggested genotype resolution result for Acadiana DNA Mixture Data No. 1. 7 loci out of the total 13 were correctly resolved, which means the top-ranked genotype combinations selected by LSD at seven loci correspond to the true genotype profiles of the two contributors.

Locus	Alleles in the Mixture	Peak Area Ratio Measured	LSD Suggested Genotype Resolution Result			
			Person 1 (Victim)	Person 2 (Suspect)	Mass Ratio Calculated	Remarks
D3S1385	14 15 17	4.06: 1: 3.91	15, 15	14, 17	1: 7.97	Wrong
VWA	15 16 19	5.05: 5.04: 1	19, 19	15, 16	1: 10.09	Wrong
FGA	22 24 25	5.35: 5.50: 1	25, 25	22, 24	1: 10.86	Correct
D8S1179	11 13 14 15	4.26: 1.37: 3.35: 1	13, 15	11, 14	1: 3.20	Correct
D21S11	28 30.2 32.2	5.23: 1: 4.61	28, 30.2	28, 32.2	1: 5.15	Correct
D18S51	16 17 20	1: 2.33: 2.08	16, 16	17, 20	1: 4.40	Correct
D5S818	11 12	4.64: 1	12, 12	11, 11	1: 4.64	Wrong
D13S317	8 11 12	4.60: 1: 4.58	11, 11	8, 12	1: 9.19	Wrong
D7S820	10 12	15.36: 1	10, 12	10, 10	1: 7.18	Correct
D16S539	8 10 11	5.45: 5.38: 1	11, 11	8, 10	1: 10.83	Wrong
TH01	6 7 9.3	4.96: 4.95: 1	9.3, 9.3	6, 7	1: 9.91	Wrong
TPOX	8 9 12	1.14: 5.86: 1	8, 12	9, 9	1: 2.74	Correct
CSF1PO	8 10 11	1: 5.83: 6.31	8, 11	10, 11	1: 6.85	Correct

Note: 'correct' in the table means that the top-ranked genotype choice at that locus in LSD ranking result corresponds to the true genotype profiles of the two contributors. 'Wrong' means that the top-ranked genotype combination at that locus does not correspond to the true genotype profiles.

consistent with the true genotype profiles of the two contributors. At the other six loci, LSD guided results failed to yield the correct genotype profiles. For example: at locus VWA, the true genotypes for the two contributors are [16, 19] and [15, 16], where allele 16 is shared by both contributors and should have a larger peak area than either of the other two alleles. However, the peak areas measured at alleles 15, 16, and 19 are 5899, 5890, and 1169 respectively, where the peak area of allele 16 (5890) is not bigger than that of allele 15 (5899) at this locus. This peak area inversion is the reason that LSD's top choice does not give the correct result. For the same reason, LSD failed to yield the correct allele assignments at the other five loci: D3S1358, D5S818, D13S317, D16S539, and TH01. As for the other seven loci, in spite of allele peak-area saturation, leading to inconsistent best-fit mass ratios, the correct genotype combinations can still be obtained from the top-ranked profiles from the LSD ranking result.

The next mixture DNA sample chosen to apply LSD to is one in which the mass ratio is very close to 1:1; thus no clear resolution is possible.

5.5.1.3. LSD Interpretation Result of Acadiana DNA Mixture Data No. 2

Acadiana DNA mixture data No. 2 and the true genotype profiles of the two contributors are presented in Table 5.22. The output of web-LSD for this set of data is shown in Figure 5.9. The corresponding LSD suggested genotype resolution result is summarized in Table 5.23. LSD guided results for this DNA mixture data suggest at least two genotype combinations for each of the thirteen loci. The calculated mass ratios at several loci are very close to 1:1. Therefore, there is no major and minor contributor to

Table 5.22: Allele peak data and the true genotypes for the two contributors at thirteen loci in Acadiana DNA Mixture Data No. 2.

Locus	Alleles in the Mixture	Allele Peak Heights Measured	Contributors' True Genotypes	
			Person 1 (Victim)	Person 2 (Suspect)
D3S1358	15 16 18	754 1253 551	16, 18	15, 16
VWA	14 17	489 1178	14, 17	17, 17
FGA	22 23	757 626	22, 23	22, 23
D8S1179	11 12 15	513 1013 612	11, 12	12, 15
D21S11	30 31.2	1108 332	30, 30	30, 31.2
D18S51	12 17 18	150 415 190	17, 18	12, 17
D5S818	10 11	1164 948	10, 11	10, 11
D13S317	9 12	339 957	9, 12	12, 12
D7S820	8 11	279 463	8, 11	11, 11
D16S539	11 12 13	506 765 304	11, 12	12, 13
TH01	7 9.3	1143 374	7, 7	7, 9.3
TPOX	9 10	809 645	9, 10	9, 10
CSF1PO	10 11	267 527	10, 11	11, 11



The results for your LSD prepared for nxue Case ID: Acadiana_data_Page3 from host at
Address 192.249.3.1

This report was prepared on Thu 30-May-2002 04:26 PM

LSD Results

Locus	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio	Comments
D3S1358						Three Alleles
1	16,18	15,16	3.0E-03	1.0	1.0:1.4	
2	16,16	15,18	6.8E-02	23	1.0:1.0	
3	18,18	15,16	0.41	138	1.0:3.6	
4	15,15	16,18	0.81	273	1.0:2.4	
5	15,18	15,16	1.21	408	1.0:4.5	
6	15,18	16,18	2.33	784	1.0:2.9	
TH01						Two Alleles REVIEW the LSD Results
1	9,3,9,3	7,7	0.0E+00		1.0:3.1	
2	7,9,3	7,7	0.0E+00		1.0:1.0	
3	9,3,9,3	7,9,3	0.0E+00		1.0:-3.0	
4	7,9,3	7,9,3	2.11		1.0:1.0	
D21S11						Two Alleles REVIEW the LSD Results
1	31,2,31,2	30,30	0.0E+00		1.0:3.3	
2	30,31,2	30,30	0.0E+00		1.0:1.2	
3	31,2,31,2	30,31,2	0.0E+00		1.0:-2.9	
4	30,31,2	30,31,2	2.73		1.0:1.0	

Figure 5.9: Output of web-LSD for Acadiana DNA Mixture data No. 2. All possible genotype combinations are ranked at each locus according to the magnitudes of the fitting errors.

D18S51						Three Alleles
1	12,18	17,17	3.6E-02	1.0	1.0:1.2	
2	12,17	17,18	8.3E-02	2.3	1.0:1.2	
3	12,12	17,18	1.12	32	1.0:4.0	
4	18,18	12,17	1.56	44	1.0:3.0	
5	12,18	17,18	2.08	59	1.0:11.6	
6	12,18	12,17	3.07	86	1.0:6.9	

D5S818						Two Alleles REVIEW the LSD Results
1	11,11	10,10	0.0E+00		1.0:1.2	
2	10,10	10,11	0.0E+00		1.0:8.8	
3	11,11	10,11	0.0E+00		1.0:-10.8	
4	10,11	10,11	2.6E-02		1.0:1.0	

D13S317						Two Alleles REVIEW the LSD Results
1	9,9	12,12	0.0E+00		1.0:2.8	
2	9,9	9,12	0.0E+00		1.0:-3.1	
3	12,12	9,12	0.0E+00		1.0:1.1	
4	9,12	9,12	1.66		1.0:1.0	

D7S820						Two Alleles REVIEW the LSD Results
1	8,8	11,11	0.0E+00		1.0:1.7	
2	8,8	8,11	0.0E+00		1.0:-5.0	
3	11,11	8,11	0.0E+0		1.0:3.0	
4	8,11	8,11	0.22		1.0:1.0	

D16S539						Three Alleles
1	12,13	11,12	7.3E-03	1.0	1.0:1.7	
2	12,12	11,13	0.22	30	1.0:1.1	
3	13,13	11,12	0.36	50	1.0:4.2	
4	11,13	11,12	1.14	157	1.0:5.0	
5	11,11	12,13	1.15	157	1.0:2.1	
6	11,13	12,13	3.37	462	1.0:2.4	

Figure 5.9: Continued.

CSF1PO						Two Alleles REVIEW the LSD Results
1	10,10	11,11	0.0E+00		1.0:2.0	
2	10,10	10,11	0.0E+00		1.0:-4.1	
3	11,11	10,11	0.0E+00		1.0:2.1	
4	10,11	10,11	0.47		1.0:1.0	
VWA						Two Alleles REVIEW the LSD Results
1	14,14	17,17	0.0E+00		1.0:2.4	
2	14,14	14,17	0.0E+00		1.0:-3.4	
3	17,17	14,17	0.0E+00		1.0:1.4	
4	14,17	14,17	0.99		1.0:1.0	
D8S1179						Three Alleles
1	11,12	12,15	1.6E-02	1.0	1.0:1.2	
2	12,12	11,15	1.9E-02	1.2	1.0:1.1	
3	11,11	12,15	0.31	19	1.0:3.2	
4	15,15	11,12	0.47	30	1.0:2.5	
5	11,15	12,15	1.06	67	1.0:3.4	
6	11,15	11,12	1.57	99	1.0:2.7	
TPOX						Two Alleles REVIEW the LSD Results
1	10,10	9,9	0.0E+00		1.0:1.3	
2	9,9	9,10	0.0E+00		1.0:7.9	
3	10,10	9,10	0.0E+00		1.0:-9.9	
4	9,10	9,10	3.2E-02		1.0:1.0	
FGA						Two Alleles REVIEW the LSD Results
1	23,23	22,22	0.0E+00		1.0:1.2	
2	22,22	22,23	0.0E+00		1.0:9.6	
3	23,23	22,23	0.0E+00		1.0:-11.6	
4	22,23	22,23	2.2E-02		1.0:1.0	

Thu 30-May-2002 04:26 PM

Figure 5.9: Continued.

Table 5.23: LSD suggested genotype resolution result for Acadiana DNA Mixture Data No. 2. All thirteen loci were resolved ambiguously, which means each locus has at least two possible choices. The true genotype profiles, as indicated in shaded cells, were obtained at 12 loci in the LSD guided result. None of the reasonable choices at locus D7S820 is correct.

Locus	Alleles in the Mixture	Peak Area Ratio Calculated	LSD Suggested Genotype Resolution Result			
			Person 1 (Victim)	Person 2 (Suspect)	Mass Ratio	Remarks
D3S1385	15 16 18	1.37: 2.27: 1	16, 18	15, 16	1: 1.38	True one
			15, 16	16, 18	1.38: 1	
VWA	14 17	1: 2.41	17, 17	14, 17	1: 1.42	
			14, 17	17, 17	1.42: 1	True one
FGA	22 23	1.21: 1	23, 23	22, 22	1: 1.21	
			22, 22	23, 23	1.21: 1	
			22, 23	22, 23	Indeterminate	True one
D8S1179	11 12 15	1: 1.97: 1.19	11, 12	12, 15	1: 1.21	True one
			12, 15	11, 12	1.21: 1	
			12, 12	11, 15	1: 1.11	
			11, 15	12, 12	1.11: 1	
D21S11	30 31.2	3.34: 1	30, 31.2	30, 30	1: 1.17	
			30, 30	30, 31.2	1.17: 1	True one
D18S51	12 17 18	1: 2.77: 1.27	12, 18	17, 17	1: 1.22	
			17, 17	12, 18	1.22: 1	
			12, 17	17, 18	1: 1.23	
			17, 18	12, 17	1.23: 1	True one
D5S818	10 11	1.23: 1	11, 11	10, 10	1: 1.23	
			10, 10	11, 11	1.23: 1	
			10, 11	10, 11	Indeterminate	True one
D13S317	9 12	1: 2.82	12, 12	9, 12	1: 1.09	
			9, 12	12, 12	1.09: 1	True one
D7S820	8 11	1: 1.66	8, 8	11, 11	1: 1.66	Wrong
			11, 11	8, 8	1.66: 1	
D16S539	11 12 13	1.66: 2.52: 1	12, 13	11, 12	1: 1.70	
			11, 12	12, 13	1.70: 1	True one
TH01	7 9.3	3.06: 1	7, 9.3	7, 7	1: 1.03	
			7, 7	7, 9.3	1.03: 1	True one
			10, 10	9, 9	1: 1.25	
TPOX	9 10	1.25: 1	9, 9	10, 10	1.25: 1	
			9, 10	9, 10	Indeterminate	True one
CSF1PO	10 11	1: 1.97	10, 10	11, 11	1: 1.97	
			11, 11	10, 10	1.97: 1	
			11, 11	10, 11	1: 2.05	
			10, 11	11, 11	2.05: 1	True one

speak of and it becomes difficult to assign which genotype to which person at each locus. This case is an illustration of one of the limitations of LSD (Section 5.3).

At six loci, D3S1358, VWA, D21S11, D13S317, D16S539, and TH01, the best-fit genotype combination at each locus can be selected by comparing the fitting errors and reviewing the consistent mass ratios. But according to LSD limitation 1 (Section 5.3), when the approximate mass ratio in the DNA mixture estimated from several loci is close to 1:1, each genotype combination will give rise to two possible genotype assignment patterns with those for person 1 and person 2 permuted.

For locus D7S820, the true genotypes of the contributors {[8, 11] [11, 11]} gave a best fitted mass ratio of 1:3, far from the best-fit mass ratio obtained at the other loci which are all less than or around 1:2. Therefore, this combination was not selected as a plausible choice. Table 5.24 shows the detailed LSD fitting for this locus. Theoretically,

Table 5.24: LSD ranking result at a two-allele locus D7S820 in Acadiana DNA Mixture Data No.2. Three genotype combinations have zero errors; the shaded cells indicate the true case for the two contributors, which is not selected as a plausible choice because of the inconsistent mass ratio.

Rank	Genotype Combinations	Errors	Mass Ratio Calculated
1	[11, 11] [8, 8]	0	1: 1.66
2	[11, 11] [8, 11]	0	1: 3.03
3	[8, 8] [8, 11]	0	1: -5.03
4	[8, 11] [8, 11]	0.22	1: 1

when the mass ratio is close to 1:1, for the genotype combination of {[8, 11] [11, 11]}, the ratio of the allele peak data should be more or less equal to 1:3. But the ratio of the allele peak data measured at this locus is 1:1.66, so the genotype combination pair of {[8, 8][11,11]} is selected with the mass ratio of 1:1.66, which is more consistent with the estimated mass ratio of 1:1.

At loci D8S1179 and D18S51, LSD fitting produced four (out of a total six) possible allele assignment patterns for each locus. The reason is that the ratio of the allele peak data measured at each locus is close to 1:2:1, and the mass ratio in the DNA mixture estimated from several loci is close to 1:1. According to LSD limitation 3 (Section 5.3), two possible genotype combinations can fit equally well to the allele peak data and each genotype combination can have two allele assignment patterns with person 1 and person 2 permuted. Therefore, each locus has four genotype combination patterns that need to be taken into consideration. The detailed ranking result at locus D8S1179 is shown in Table 5.25. The errors as well as the mass ratios of the top two ranked genotype combinations are very close to each other. Therefore, it is impossible for one to make a clear decision as to which is overwhelmingly the correct choice, unless the genotype of one contributor is known.

For loci FGA, D5S818, and TPOX, each locus has only two alleles, and the ratio of the two allele peak data happens to be close to 1:1. According to LSD limitation 4(a) (Section 5.3). Three allele assignment patterns can fit equally well to the allele peak data and all of them were selected as possible choices for this locus (two of which are

Table 5.25: LSD ranking result at D8S1179, a three-allele locus in Acadiana DNA Mixture Data No.2, according to the fitting error associated with each case. The top-ranked genotype combination in shaded cells turns out to be the true one for the two contributors.

Rank	Genotype Combinations	Ratio of the Errors	Mass Ratio Calculated
1	[11, 12] [12, 15]	1	1: 1.21
2	[12, 12] [11, 15]	1.17	1: 1.11
3	[11, 11] [12, 15]	19.23	1: 3.17
4	[15, 15] [11, 12]	29.89	1: 2.49
5	[11, 15] [12, 15]	66.60	1: 3.40
6	[11, 15] [11, 12]	98.58	1: 2.66

permuted arrangement of each other). The caseworkers would not be able to decide which one to choose, unless the genotype of one contributor is available.

For the 2-allele locus CSF1PO, since the ratio of the two allele peak areas is very close to 1:2, according to LSD limitation 4(b) (Section 5.3), there are two possible allele assignment patterns at this locus: {[A, A][B, B]} and {[B, B][A, B]}, both of which would fit equally well to the measured allele peak data with comparable mass ratios. The approximate mass ratio estimated from other loci in the DNA mixture is around 1:1, which makes it difficult to differentiate the major and minor contributors. Therefore, the permuted allele assignment patterns for person 1 and person 2 of these two cases should also be taken into consideration. As a result, LSD suggests a total of four genotype combination pairs for this locus.

For this set of DNA mixture data, although LSD does not suggest a very clear-cut resolution due to the close mass ratio of 1:1. However, as long as the genotype of one contributor is known, LSD is still useful in yielding the most fitting genotype profile for the other contributor, because it is the special combination of the genotypes of the two contributors, as stipulated by the LSD genotype combination pair that minimizes the fitting error, when fitting to the allele peak data.

From the above discussion, it can be concluded that in the absence of DNA loading saturation and with mass ratio away from 1:1; LSD algorithm would yield correct genotypes for both contributors by picking out the top-ranked genotype combinations with consistent mass ratios. Sometimes, LSD suggests more than one possible genotype combinations at one locus, due to the imbalance associated with the measurement of the allele peak data and the limitations of LSD as presented in Section 5.3. In this case, if the genotype of one contributor is known, then it should be brought in after LSD, to help the DNA analysts to decide among the choices which one is applicable. The LSD guided results of the Acadiana DNA mixture data suggest that the LSD algorithm is a very feasible methodology in resolving mixed DNA samples under most circumstances, bearing in mind the limitations discussed in Section 5.3.

5.5.2. LSD Interpretation Result of DNA Mixture Samples from the Texas Department of Public Safety in Corpus Christi

To further test the functionality and effectiveness of the LSD algorithm, five additional sets of allele peak data from forensic mixed DNA samples were run using

LSD. These data were provided by Ms. Patricia Graham in the Texas Department of Public Safety in Corpus Christi [20]. These 2-people DNA mixture data were obtained from swabs in rape cases. The genotype of the female victim is known precisely in each case. Sperm fraction in the mixture sample was separated from the mixture evidence by differential extraction. Therefore, genotype profiles of the male perpetrators are also obtained for all five samples. LSD was applied to the mixture data and the LSD suggested results were compared to the known genotype profiles of the two contributors. Among the five sets of data, four have peak area information at nine loci, and the fifth has information at twelve loci. In this report, three of the five sets of the DNA mixture data from Corpus Christi and the corresponding LSD suggested genotype resolution results were selected and presented for discussion for different objectives:

- Texas DNA mixture data No. 1 serves as a typical example when the LSD guided results are clear cut, where the proportionality between the DNA masses and the corresponding allele peak areas are well preserved and none of the LSD limitations applies. LSD suggested results for it were already presented and discussed as an example of LSD usage in Chapter 4, and will not be discussed further here.
- Texas DNA mixture data No. 3 is selected to show the evoking of limitations of LSD at two loci when the approximate mass ratio in the DNA mixture is close to 1:2, as predicted from other loci.
- Texas DNA mixture data No. 4 is another example where the LSD guided genotype resolution results are clear-cut. The mass ratios for the top-ranked profiles in the LSD ranking results at all loci are all consistent with each other, and the fitting errors for the top two ranked cases are all fairly far apart.

Figure 5.10 is the electropherogram showing the allele peaks at nine loci for Texas DNA mixture data No. 3. The peak area data and the true genotypes of the female victim and the male suspect are summarized in Table 5.26. Figure 5.11 shows the output of web-LSD for this set of DNA mixture data. The summary of LSD suggested genotype resolution result is presented in Table 5.27. Among the nine loci, seven have 3 or 4 alleles per locus and the top-ranked genotype profiles for the two contributors do correspond to the true genotypes (Person 2 corresponds to the female victim). At each of the remaining 2-allele loci, D21S11 and D7S820, two possible choices are selected, because both choices in each locus give a mass ratio consistent with those for other loci. The best-fit mass ratios across the nine loci are consistent with each other, ranging from 1:1.20 (d5S818) to 1:1.76 (D3S1358). Locus VWA has only one allele, so there is no doubt that each contributor has the genotype [16, 16] at this locus. For loci D21S11 and D7S820, since the ratio of the allele peak areas measured at each locus is indeed close to 1:2 and the approximate mass ratio in this mixed DNA sample can also be close to 1:2 as predicted from the other loci, according to LSD limitation 4(b) (Section 5.3). Two genotype combinations are possible with consistent mass ratios across both loci, and fit equally well to the allele peak data measured, as shown in Table 5.27. Only one of these two is consistent with the true genotypes of the female victim and the male perpetrator. Therefore, when the victim's genotype's is brought in to resolve the choices, one would choose the one where person 2's profile agrees with that of the known victim's profile. Those are given as the shaded cells in Table 5.27. This set of mixture sample was successfully resolved at all loci. The victim's profile was brought in at two 2-allele loci to help select between two alternative genotype combination choices.

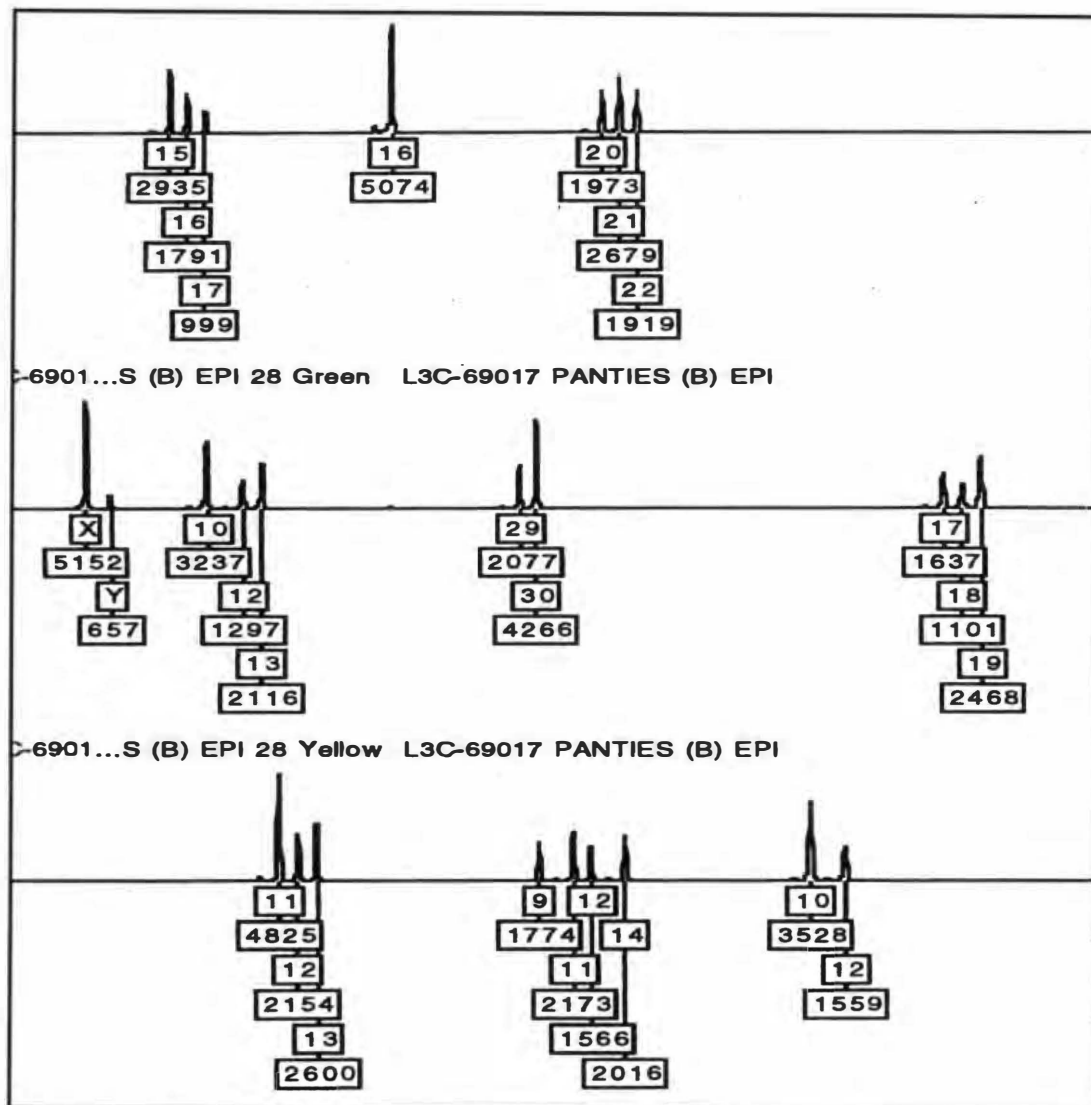
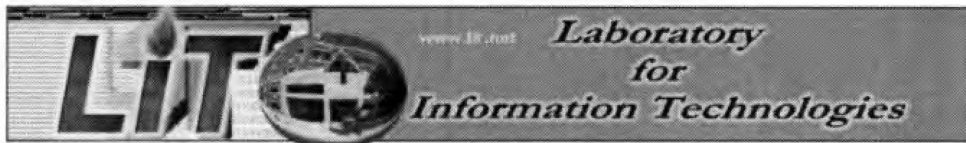


Figure 5.10: The allele peaks at nine loci of the Texas DNA mixture data No. 3; both the allele name and the corresponding peak areas are indicated.

Source: P. Graham. "Allele Peak Data for DNA Mixture Samples from Real Cases," Texas Department of Public Safety, Corpus Christi, Personal Communication, 2001.

Table 5.26: Allele peak data and the genotypes of both contributors at nine loci in Texas DNA Mixture Data No. 3.

Locus	Alleles in the Mixture	Allele Peak Areas Measured	Contributors' Genotypes	
			Female Victim	Male Suspect
D3S1358	15 16 17	2935 1791 999	15, 16	15, 17
VWA	16	5074	16, 16	16, 16
FGA	20 21 22	1973 2679 1919	20, 22	21, 21
D8S1179	10 12 13	3237 1297 2116	10, 13	10, 12
D21S11	29 30	2077 4266	29, 30	30, 30
D18S51	17 18 19	1637 1101 2468	17, 19	18, 19
D5S818	11 12 13	4825 2154 2600	11, 13	11, 12
D13S317	9 11 12 14	1774 2173 1566 2016	11, 14	9, 12
D7S820	10 12	3528 1559	10, 12	10, 10



The results for your LSD prepared for nxue Case ID: Texas_data_3 from host at Address 192.249.3.1

This report was prepared on Thu 30-May-2002 04:56 PM

LSD Results

Locus	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio	Comments
D3S1358						Three Alleles
1	15,17	15,16	7.0E-03	1.0	1.0:1.8	
2	16,17	15,15	0.31	45	1.0:1.1	
3	17,17	15,16	0.66	93	1.0:4.7	
4	16,17	15,16	1.53	218	1.0:7.8	
5	16,16	15,17	1.88	267	1.0:2.2	
6	16,17	15,17	4.64	661	1.0:3.1	
TH01						No Alleles
D21S11						Two Alleles REVIEW the LSD Results
1	29,29	30,30	0.0E+00		1.0:2.1	
2	29,29	29,30	0.0E+00		1.0:-3.9	
3	30,30	29,30	0.0E+00		1.0:1.9	
4	29,30	29,30	0.56		1.0:1.0	

Figure 5.11: Output of web-LSD for Texas DNA Mixture data No. 3. All possible genotype combinations are ranked at each locus according to the magnitudes of the fitting errors.

D18S51						Three Alleles
1	18,19	17,19	2.0E-02	1.0	1.0:1.5	
2	19,19	17,18	0.12	5.9	1.0:1.1	
3	18,18	17,19	0.28	14	1.0:3.7	
4	17,17	18,19	0.77	38	1.0:2.2	
5	17,18	17,19	1.03	51	1.0:4.0	
6	17,18	18,19	2.48	124	1.0:2.3	
D5S818						Three Alleles
1	11,12	11,13	3.6E-04	1.0	1.0:1.2	
2	12,13	11,11	2.1E-02	59	1.0:1.0	
3	12,12	11,13	0.53	1473	1.0:3.4	
4	13,13	11,12	0.77	2123	1.0:2.7	
5	12,13	11,13	1.38	3804	1.0:4.8	
6	12,13	11,12	2.00	5511	1.0:3.6	
D13S317						Four Alleles
1	9,12	11,14	1.4E-02	1.0	1.0:1.3	
2	12,14	9,11	7.4E-02	5.3	1.0:1.1	
3	11,12	9,14	8.7E-02	6.3	1.0:1.0	
D7S820						Two Alleles REVIEW the LSD Results
1	12,12	10,10	0.0E+00		1.0:2.3	
2	10,10	10,12	0.0E+00		1.0:1.6	
3	12,12	10,12	0.0E+00		1.0:-3.6	
4	10,12	10,12	0.80		1.0:1.0	

Figure 5.11: Continued.

D16S539						No Alleles
CSF1PO						No Alleles
VWA						One Allele Only
1	16,16	16,16			1:1	
D8S1179						Three Alleles
1	10,12	10,13	6.1E-03	1.0	1.0:1.7	
2	10,10	12,13	0.20	32	1.0:1.1	
3	12,12	10,13	0.37	61	1.0:4.1	
4	13,13	10,12	1.12	182	1.0:2.1	
5	12,13	10,13	1.16	189	1.0:5.0	
6	12,13	10,12	3.26	531	1.0:2.5	
TPOX						No Alleles
FGA						Three Alleles
1	21,21	20,22	4.0E-04	1.0	1.0:1.5	
2	22,22	20,21	6.8E-02	171	1.0:2.4	
3	20,20	21,22	7.8E-02	198	1.0:2.3	
4	21,22	20,21	0.13	336	1.0:1.0	
5	20,22	20,21	0.62	1575	1.0:1.7	
6	20,22	21,22	0.68	1708	1.0:1.7	

Thu 30-May-2002 04:56 PM

Figure 5.11: Continued.

Table 5.27: LSD suggested genotype resolution result for Texas DNA mixture data No. 3. The top-ranked genotype combinations yielded by LSD correspond to the true genotype profiles of the two contributors at seven loci. The LSD results provide more than one possible choice at two loci; the one with shaded cells indicates the true genotypes for the two contributors, after the victim's profile is brought in for resolution.

Locus	Alleles in the Mixture	Peak Area Ratio Measured	LSD Suggested Genotype Resolution Result			
			Person 1 (Suspect)	Person 2 (Victim)	Mass Ratio Calculated	Remarks
D3S1385	15 16 17	2.94: 1.79: 1	15, 17	15, 16	1: 1.76	True one
VWA	16	--	16, 16	16, 16	Indeterminate	True one
FGA	20 21 22	1.02: 1.40: 1	21, 21	20, 22	1: 1.45	True one
D8S1179	10 12 13	2.50: 1: 1.63	10, 12	10, 13	1: 1.66	True one
D21S11	29 30	1; 2.05	30, 30	29, 30	1: 1.89	True one
			29, 29	30, 30	1: 2.05	
D18S51	17 18 19	1.49: 1: 2.24	18, 19	17, 19	1: 1.53	True one
D5S818	11 12 13	2.24: 1: 1.21	11, 12	11, 13	1: 1.20	True one
D13S317	9 11 12 14	1.13: 1.39: 1: 1.29	9, 12	11, 14	1: 1.25	True one
D7S820	10 12	2.263: 1	10, 10	10, 12	1: 1.58	True one
			12, 12	10, 10	1: 2.26	

Analysis of the third sample is now presented. Figure 5.12 shows the allele peaks at nine loci for Texas DNA mixture data set No. 4. This is a relatively clear set of data, resulting in a clear-cut interpretation of the LSD guided results. The allele peak area data and the true genotypes of the female victim and male perpetrator are presented in Table 5.28. The output of web-LSD is presented in Figure 5.13. From the summary of LSD suggested genotype resolution result shown in Table 5.29, it is clear that all nine loci are resolved correctly and unambiguously, where the top-ranked genotype combinations selected by LSD are indeed consistent with the true genotype profiles of the two contributors at all nine loci. The mass ratios calculated from the nine loci are very consistent with each other ranging from 1: 2.54 (D3S1358) to 1: 4.45 (VWA). Deconvolution at locus D18S51 needs some discussion. D18S51 has only two alleles and as always, three of the four possible genotype combination profiles have zero fitting errors. The best fitting mass ratios calculated from these three possible combinations are 1:1.2, 1:-14.6, and 1:12.6, all of which are not consistent with those calculated from the other eight loci. Therefore, none of them is selected as a possible choice for this locus. Since the ratio of the allele peak areas measured at this locus is 1.16:1, which is very close to 1:1, then according to LSD limitation 4(a) (Section 5.3), the genotype combination of {[12, 14][12, 14]} should always be considered as a possible choice, and we do choose this genotype combination pair as the preferred choice, which turns out to be the correct one.

From the LSD suggested genotype resolution results of the DNA mixture data provided by the Acadiana Crime Lab and the Texas Department of Public Safety in

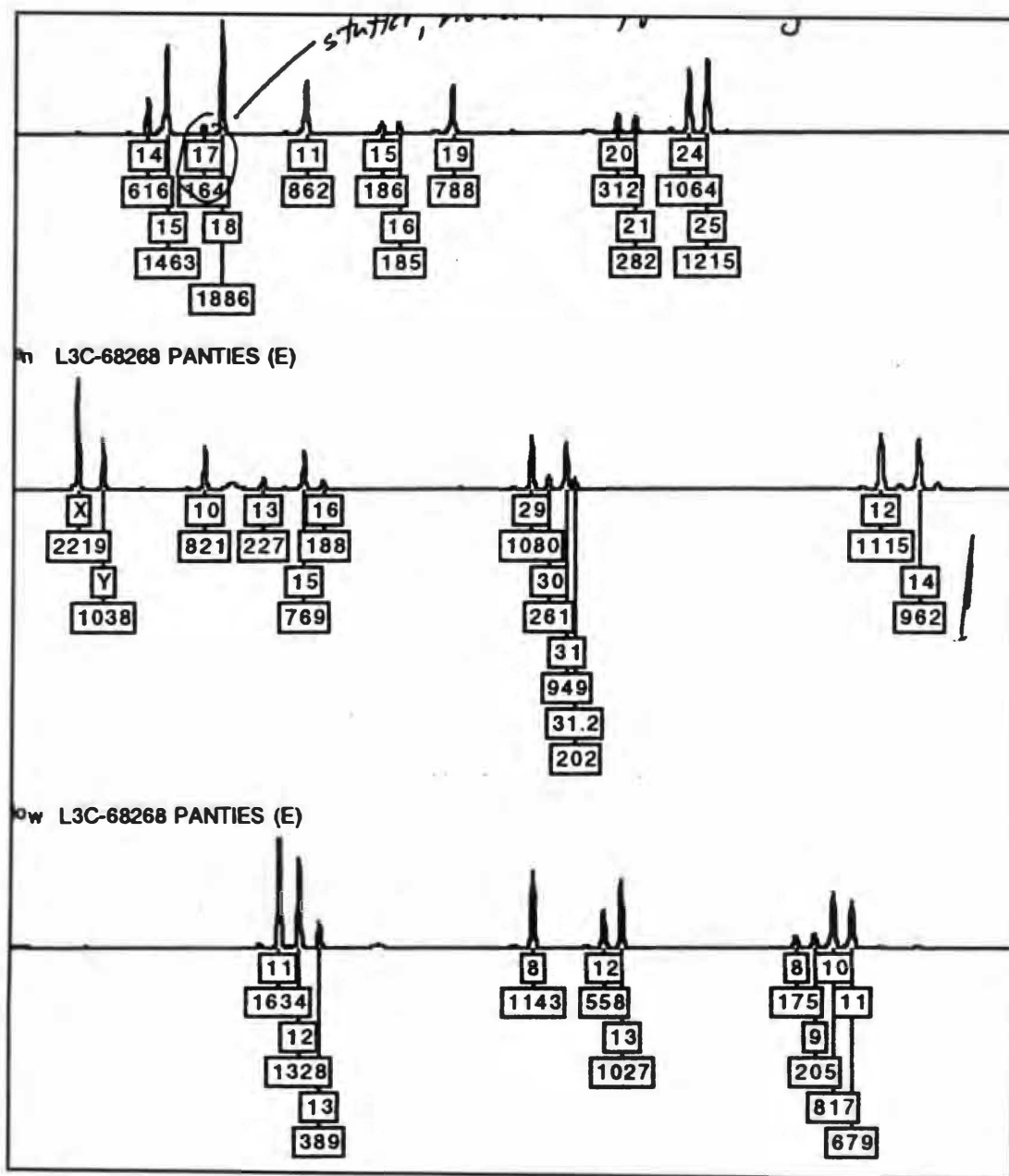


Figure 5.12: The allele peaks at nine loci of the Texas DNA mixture data No. 4; both the allele name and the corresponding peak area are indicated.

Source: P. Graham. "Allele Peak Data for DNA Mixture Samples from Real Cases," Texas Department of Public Safety, Corpus Christi, Personal Communication, 2001.

Table 5.28: Allele peak data and the true genotypes of both contributors at nine loci in Texas DNA Mixture Data No. 4.

Locus	Alleles in the Mixture	Allele Peak Areas Measured	Contributors' Genotypes	
			Female Victim	Male Suspect
D3S1358	14 15 18	616 1463 1886	14, 18	15, 18
VWA	11 15 16 19	862 186 185 788	15, 16	11, 19
FGA	20 21 24 25	312 282 1064 1215	20, 21	24, 25
D8S1179	10 13 15 16	821 227 769 188	13, 16	10, 15
D21S11	29 30 31 31.2	1080 261 949 202	30, 31.2	29, 31
D18S51	12 14	1115 962	12, 14	12, 14
D5S818	11 12 13	634 328 389	11, 13	11, 12
D13S317	8 12 13	1143 558 1027	12, 12	8, 13
D7S820	8 9 10 11	175 205 817 679	8, 9	10, 11



The results for your LSD prepared for nxue Case ID: Texas_data_4 from host at Address
192.249.3.1

This report was prepared on Thu 30-May-2002 04:43 PM

LSD Results

Locus	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio	Comments
D3S1358						Three Alleles
1	14,18	15,18	3.3E-02	1.0	1.0:2.5	
2	14,14	15,18	0.24	7.2	1.0:5.4	
3	18,18	14,15	0.95	29	1.0:1.1	
4	14,15	15,18	0.95	29	1.0:5.7	
5	15,15	14,18	2.13	65	1.0:1.7	
6	14,15	14,18	6.56	201	1.0:1.8	
TH01						No Alleles
D21S11						Four Alleles
1	30,31.2	29,31	0.25	1.0	1.0:4.4	
2	31,31.2	29,30	15.06	60	1.0:1.2	
3	30,31	29,31.2	15.25	60	1.0:1.1	

Figure 5.13: Output of web-LSD for Texas DNA Mixture data No. 4. All possible genotype combinations are ranked at each locus according to the magnitudes of the fitting errors.

D18S51						Two Alleles REVIEW the LSD Results
1	14,14	12,12	0.0E+00		1.0:1.2	
2	12,12	12,14	0.0E+00		1.0:12.6	
3	14,14	12,14	0.0E+00		1.0:-14.6	
4	12,14	12,14	1.3E-02		1.0:1.0	
D5S818						Three Alleles
1	11,13	11,12	1.5E-02	1.0	1.0:3.6	
2	13,13	11,12	0.31	20	1.0:7.6	
3	12,13	11,12	1.06	70	1.0:8.9	
4	11,11	12,13	2.91	192	1.0:1.1	
5	12,12	11,13	5.12	338	1.0:1.5	
6	12,13	11,13	14.58	961	1.0:1.7	
D13S317						Three Alleles
1	12,12	8,13	2.2E-02	1.0	1.0:3.9	
2	8,12	8,13	0.21	9.7	1.0:2.1	
3	8,8	12,13	0.35	16	1.0:1.4	
4	12,13	8,13	0.49	23	1.0:2.8	
5	13,13	8,12	0.55	25	1.0:1.7	
6	12,13	8,12	2.78	129	1.0:1.2	
D7S820						Four Alleles
1	8,9	10,11	0.33	1.0	1.0:3.9	
2	8,11	9,10	10.26	32	1.0:1.2	
3	9,11	8,10	10.40	32	1.0:1.1	

Figure 5.13: Continued.

D16S539

No Alleles

CSF1PO

No Alleles

VWA

Four Alleles

1	15,16	11,19	8.0E-02	1.0	1.0:4.4	
2	16,19	11,15	11.99	150	1.0:1.1	
3	15,19	11,16	11.99	150	1.0:1.1	

D8S1179

Four Alleles

1	13,16	10,15	6.0E-02	1.0	1.0:3.8	
2	15,16	10,13	9.77	163	1.0:1.1	
3	13,15	10,16	9.82	164	1.0:1.0	

TPOX

No Alleles

FGA

Four Alleles

1	20,21	24,25	0.15	1.0	1.0:3.8	
2	21,24	20,25	8.97	60	1.0:1.1	
3	20,24	21,25	9.03	61	1.0:1.1	

Thu 30-May-2002 04:43 PM

Figure 5.13: Continued.

Table 5.29: LSD suggested genotype resolution result for Texas DNA mixture data No. 4. All 9 loci were correctly resolved using LSD, which means the top-ranked genotype combinations selected by LSD are consistent with the true genotype profiles of the two contributors at all nine loci.

Locus	Alleles in the Mixture	Peak Area Ratio Measured	LSD Suggested Genotype Resolution Result			
			Person 1 (Victim)	Person 2 (Suspect)	Mass Ratio Calculated	Remarks
D3S1385	14 15 18	1: 2.38: 3.06	14, 18	15, 18	1:2.54	True one
VWA	11 15 16 19	4.66: 1.01: 1: 4.26	15, 16	11, 19	1:4.45	True one
FGA	20 21 24 25	1.11: 1: 3.77: 4.31	20, 21	24, 25	1:3.84	True one
D8S1179	10 13 15 16	4.37: 1.21	13, 16	10, 15	1:3.83	True one
D21S11	29 30 31 31.2	5.35: 1.29: 4.70: 1	30, 31.2	29, 31	1:4.38	True one
D18S51	12 14	1.16: 1	12, 14	12, 14	Indeterminate	True one
D5S818	11 12 13	4.20: 3.41: 1	11, 13	11, 12	1:3.60	True one
D13S317	8 12 13	2.05: 1: 1.84	12, 12	8, 13	1:3.89	True one
D7S820	8 9 10 11	1: 1.17: 4.67: 3.88	8, 9	10, 11	1:3.94	True one

Corpus Christi, and other data not reported here, it is concluded that as long as the allele peak data reflect the relative mass proportion of the alleles in the mixture, LSD will always yield the true genotype profiles for the two contributors as its top-ranked profiles, keeping in mind the special consideration accorded to all 2-allele loci LSD results. Therefore, LSD algorithm appears to be a feasible and powerful method in resolving DNA mixture samples of two people.

5.6. Comparison Between LSD and LMA

Linear Mixture Analysis (LMA) is a recently published algorithm to resolve STR mixture DNA [18] (also in Section 3.2.2.). Both LMA and LSD methods use quantitative allele peak data and linear algebra principals to solve the DNA mixture problem. After extensive comparison study we conclude that LSD is more flexible in arriving at the most fitting genotype profiles for the individual contributors of the mixture. LSD operates locus by locus and fits each locus separately, followed by pulling together only those loci at which resolution is clear and consistent with each other, to form a final composite profile for the two contributors. Since the fit is locus by locus, a different best-fit mass ratio maybe found for each locus, and they may differ across the loci by as much as two fold. This flexibility of independent fitting allows the fit to be optimum at each locus. A degree of confidence can be separately assigned to each locus in accordance to its relative degree of resolution (Chapter 4). Whereas, in LMA, all available loci are processed as one entity and a single mass ratio is sought to fit the given allele peak data simultaneously at all loci. As a result, one cannot pick and choose those loci with high degree of resolution in the formation of a composite profile (Section 3.2.2.). LMA and LSD differ greatly in three

other aspects, as presented below:

When LMA is used to resolve a two-people DNA mixture, the genotype of one of the two contributors has to be known, and entered into the LMA algorithm. Such information is necessary to derive the other contributor's genotype. When using LSD to resolve such mixed DNA profiles, no *a priori* genotype information is necessary. The best-fit genotype combination pair for both contributors is to be obtained simultaneously in one step. The reference profile, when known, is used in LSD to validate the LSD result for the person to which the reference profile corresponds. For those loci at which LSD results do not agree with the reference profile, they can be dropped from the final composite profile for the other unknown person.

A second characteristic of LMA different from that of LSD, is in its need of iterative computation in arriving at the optimal mass fraction to fit the given allele peak area (Section 3.2.2.). This in itself is more laborious to set up and takes time to search through all the possible mass fractions. The most serious difference of LMA in comparison to LSD is that the relative DNA mass fractions are formulated to be the **same** for **all** loci. In other words, the authors attempt to use the same set of DNA mixture mass fractions to uniformly fit the measured allele peak areas at all loci. Theoretically, this assumption is reasonable, since the allele peak areas are supposed to be proportional to the relative DNA masses in the mixture, and which is supposedly to be preserved across all loci during the PCR amplification process. However, our experience with examining real DNA mixtures through LSD has revealed that peak imbalances are always associated

with the allele peak areas (or heights) within a locus and uneven PCR amplification is also evident across all loci. Therefore, the best-fit mass ratio calculated for one locus would vary from those for other loci, by as much as three fold. For this reason, LMA's use of the same DNA mixture mass fractions for all loci may not result in the optimal fit at some loci, thus increasing the odds of selecting an incorrect genotype combination as the most fitting one at those loci for the two contributors.

In contrast, when LSD is used to resolve a DNA/STR mixture sample, each locus is analyzed separately to find the best-fit genotypes for the two contributors. When all loci have been analyzed, results from all loci are pooled and checked for cross-loci mass ratio consistency in order to form a composite profile. All loci are reviewed separately to decide if the LSD suggested genotype resolution results at those loci are consistent and confident enough to be included in the final presentation of the resolution results. In LMA, the authors do mention that the degree of confidence of the resolution result can be checked by evaluating the size of the fitting residuals calculated for each locus. However, at some loci, the large fitting residuals may be caused by trying to force a prior-specified but non-optimal mass ratio to fit the observed allele peak areas and not necessarily from the improper fit of genotype combination pair to the measured allele peak areas. As a result, a wrong conclusion about the genotype choice may be made. This uncertainty may result in a lack of fit compared to our LSD algorithm.

In the original Reference [18], where LMA was introduced, an example of using LMA to resolve a DNA mixture sample using the Matlab[®] software platform was

presented to illustrate the effectiveness of LMA. The DNA mixture data used and the genotypes of both contributors are reproduced in Table 5.30. The detailed Matlab program implementing LMA to resolve it was not included in [18]. Before we compare the effectiveness of LMA algorithm to that of LSD, a control experiment was run to ensure that our Matlab implementation of the LMA algorithm (modified from that of the original paper because there is a misstatement in the stated algorithm) would yield the same result as in Reference [18]. The DNA mixture data shown in Table 5.30 was used to conduct the control experiment. Figure 5.14 shows the corresponding LMA genotype resolution result as presented in Reference [18]. In Figure 5.14, the first row shows the normalized DNA mixture data at each of the ten loci; the second row shows the genotype of Person A, which is required to calculate the genotype of Person B in the LMA algorithm, the third row shows the genotype of Person B yielded by LMA; the fourth row is the actual genotype of person B. All 10 loci were correctly resolved with the optimal mass fraction being 29.18% for Person A from iterative search, which is very close to the known value of 30% prior to PCR amplification. Figure 5.15 shows the genotype resolution result yielded by our implementation of LMA, also carried out using the Matlab® platform. The optimal mass fraction of Person A is identified as 29% from the iterative search shown in Figure 5.16. The calculated vector g_B and the final genotype result in our implementation of LMA are exactly the same as those presented in Reference [18] shown in Table 5.31. This control experiment verifies that the LMA algorithm implemented by us, based on our understanding of LMA, is correct, as it gives the same result as that published in [18].

Table 5.30: Allele peak data and the genotypes of the two contributors at ten loci of a mixed DNA sample with the mass ratio of 3:7 prior to PCR, as presented in Reference [18].

Locus	Alleles in the Mixture	Genotype Combination		Normalized Allele Peak Areas
		Person A	Person B	
D3S1358	14 15	14 15	14 15	1.04 0.96
VWA	17 18	18 18	17 17	1.48 0.52
D16S539	11 13 14	13 14	11 11	1.45 0.29 0.27
D2S1338	16 18 20 21	16 20	18 21	0.32 0.64 0.37 0.68
D8S1179	9 12 13 14	12 14	9 13	0.73 0.27 0.68 0.32
D21S11	27 29 30	27 30	29 30	0.28 0.79 0.93
D18S51	12 13 14 17	12 17	13 14	0.34 0.70 0.68 0.29
D19S433	12.2 14 15	14 14	12.2 15	0.70 0.61 0.69
TH01	6 7 9	6 7	7 9	0.32 1.01 0.67
FGA	19 24 25.2	19 24	19 25.2	1.06 0.28 0.66

Source of data: M. Perlin and B. Szabady, “ Linear Mixture Analysis: A Mathematical Approach to Resolving Mixed DNA Samples,” *Journal of Forensic Science*, 46-6(2001): 1372-1378.

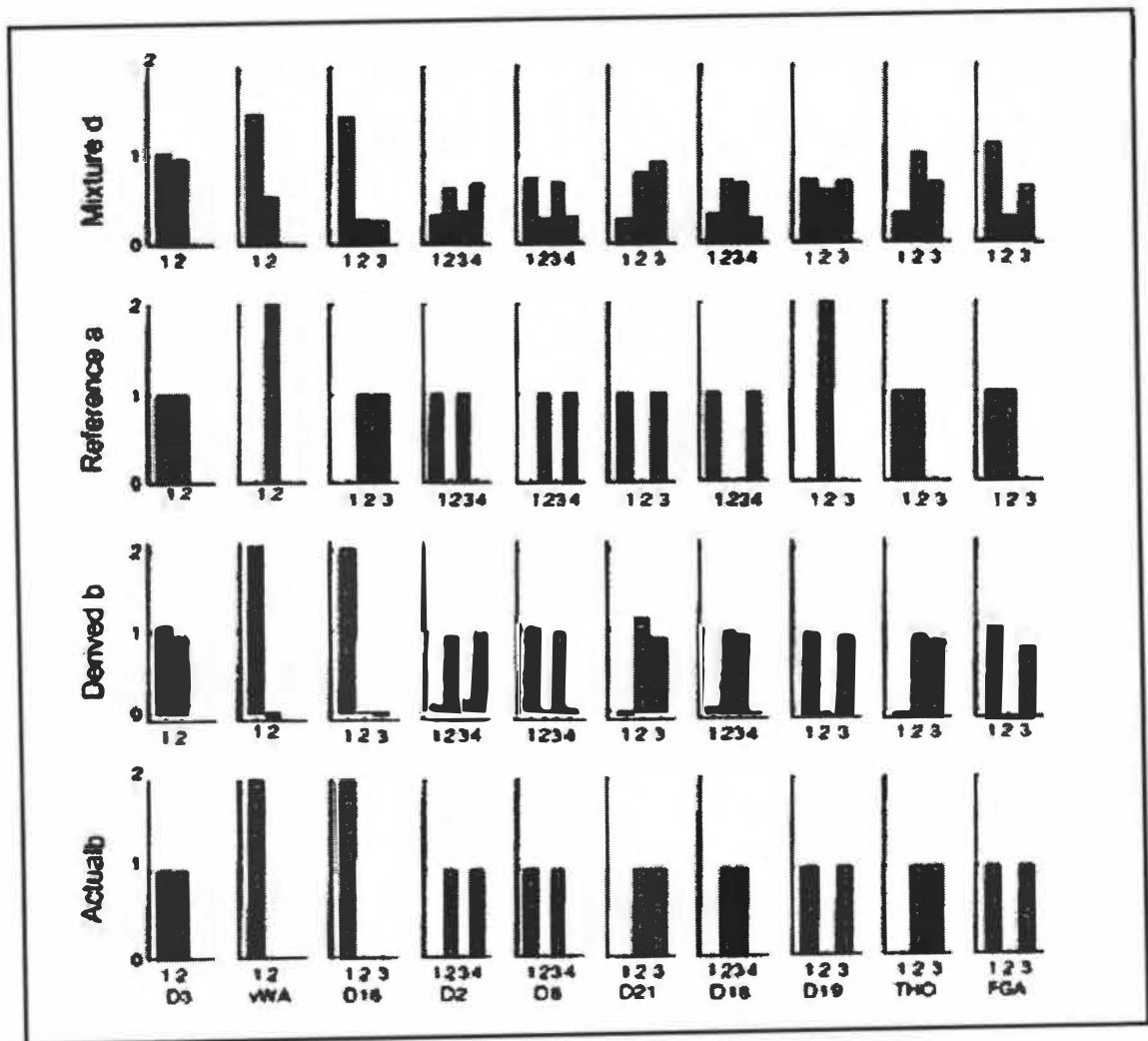


Figure 5.14: Reproduced from [18]: LMA genotype resolution result of the DNA mixture data shown in Table 5.30. First row shows the quantitative allele peak data at each locus; the second row shows the known reference profile of Person A; the third row shows the LMA suggested genotype of Person B, which is consistent with the actual genotype profile of Person B, shown in the Fourth row [18].

Note: Reprinted, with permission from the Journal of Forensic Sciences, Vol. 46, No. 6, copyright ASTM International.

Source: M. Perlin and B. Szabady, "Linear Mixture Analysis: A Mathematical Approach to Resolving Mixed DNA Samples," *Journal of Forensic Science*, 46-6(2001): 1372-1378.

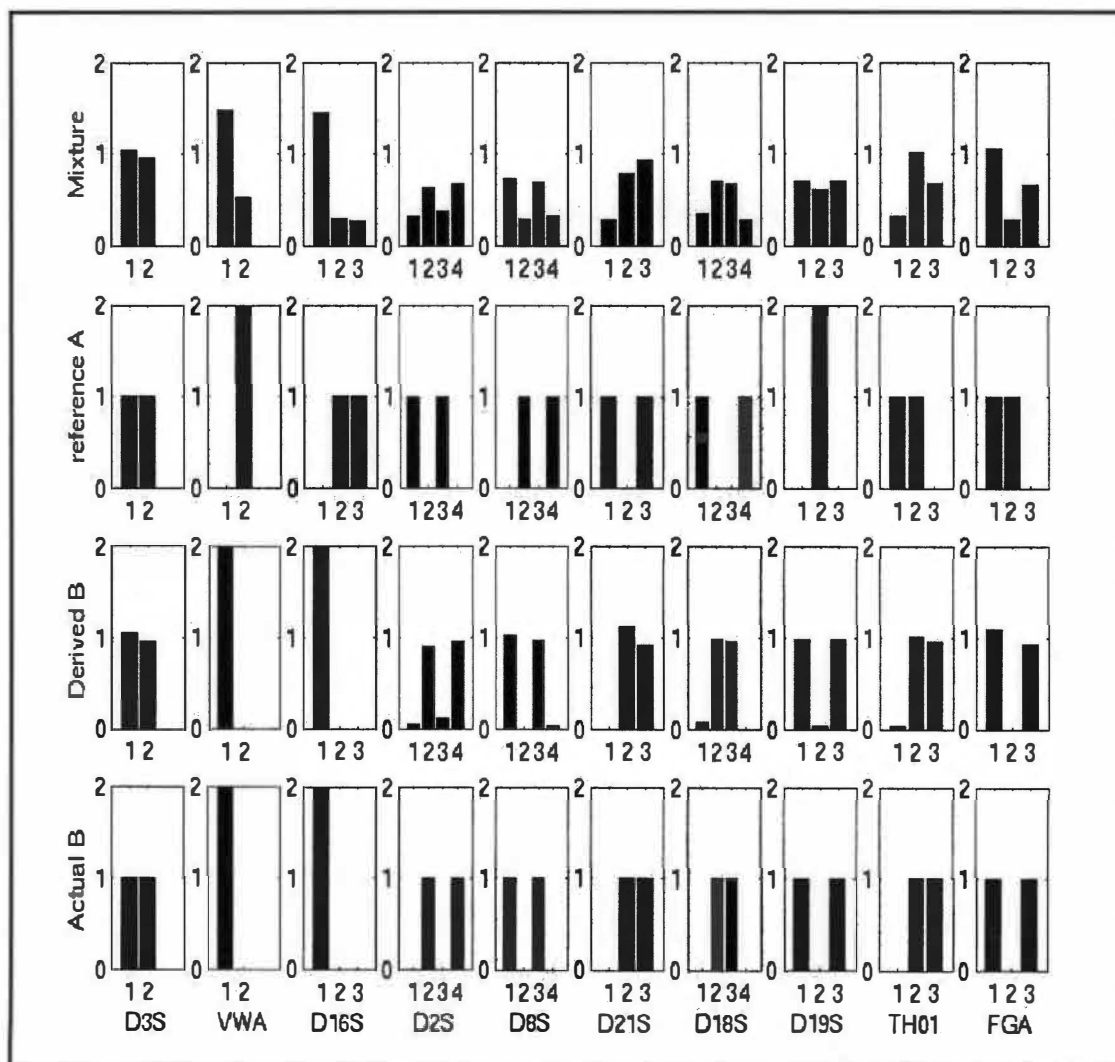


Figure 5.15: Genotype resolution result of the DNA mixture data in Table 5.30, using LMA algorithm modified from the original paper implemented by the author. First row shows the quantitative allele peak data at each locus; the second row shows the known reference profile of Person A; the third row shows the LMA suggested genotype of Person B, which is consistent with the actual genotype profile of Person B, shown in the fourth row. The result is almost the same as that presented in Figure 5.14.

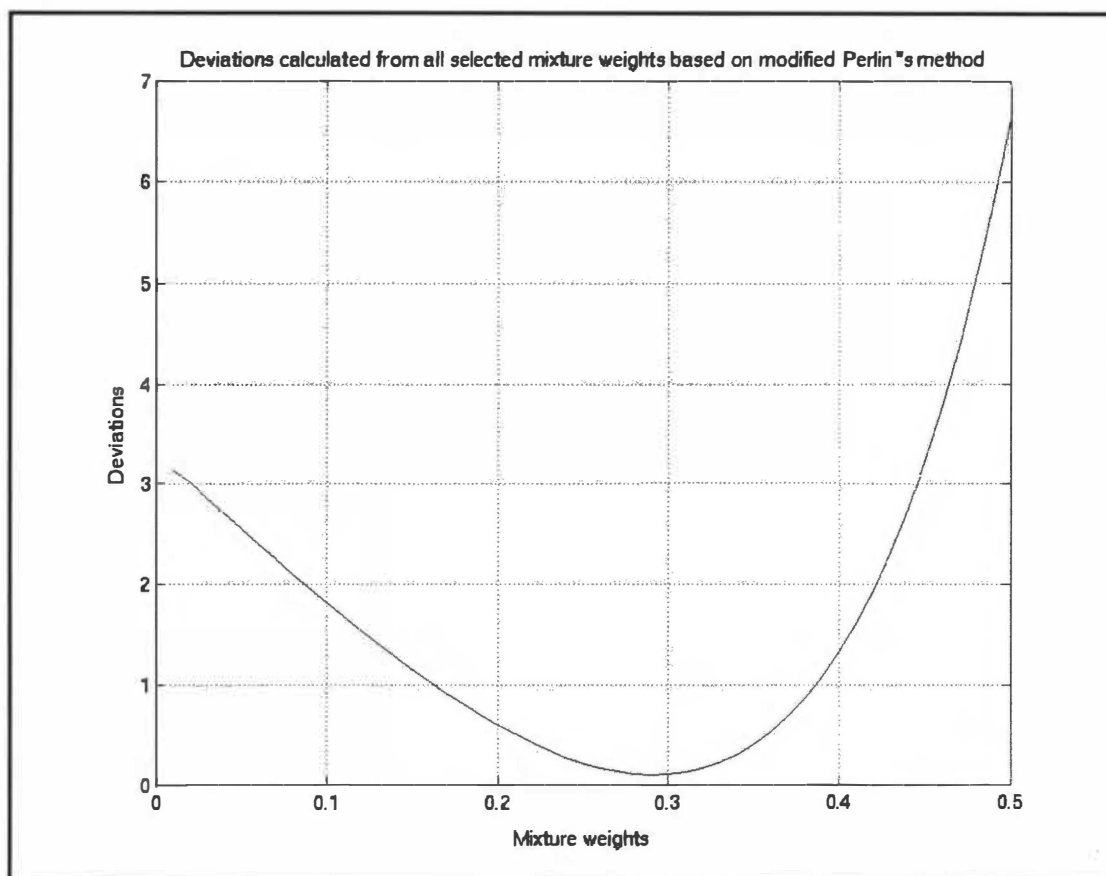


Figure 5.16: Sum of squared deviation as a function of the set of assumed mass fractions of LMA on mixture data presented in Table 5.30. The mass fraction of 29% is identified as the optimal mass proportion of person A. Such result is consistent with the known value of 30%.

Table 5.31: Validation of our implementation of LMA [18]. Exactly the same results are generated.

Locus	Alleles in the Mixture	LMA Result for Person B Presented in [18]			Genotype of Person B Yielded from Our Implementation of LMA			True Genotype of Person B
		Vector g_B		Alleles	Vector g_B		Alleles	
D3S1358	14	1.0516	1	14	1.05	1	14	14
	15	0.9484	1	15	0.95	1	15	15
VWA	17	2.0835	2	17	2.08	2	17	17
	18	-0.0835	0	17	-0.08	0	17	17
D16S539	11	2.0406	2	11	2.04	2	11	11
	13	-0.0041	0	11	0	0	11	11
D2S1338	16	-0.0365	0	18	-0.03	0	18	18
	20	0.0384	0	21	0.04	0	21	21
D8S1179	9	0.8951	1	9	0.89	1	9	9
	12	0.1122	0	13	0.11	0	13	13
D21S11	27	0.9543	1	29	0.95	1	29	29
	29	1.0278	1	30	1.03	1	30	30
D18S51	12	-0.0239	0	13	-0.02	0	13	13
	13	0.9620	1	14	0.96	1	14	14
D19S433	14	0.0341	0	12.2	0.04	0	12.2	12.2
	15	-0.0185	0	15	-0.01	0	15	15
TH01	6	1.1121	1	7	1.11	1	7	7
	7	0.9064	1	9	0.91	1	9	9
FGA	19	0.0741	0	19	0.08	0	19	19
	24	0.9816	1	25.2	0.98	1	25.2	25.2
	25.2	0.9538	1		0.95	1		
		-0.0096	0		-0.01	0		
		0.9872	1		0.98	1		
		0.0361	0		0.04	0		
		0.9813	1		0.98	1		
		0.0366	0		0.04	0		
		1.0104	1		1.01	1		
		0.9530	1		0.95	1		
		1.0819	1		1.08	1		
		-0.0124	0		-0.01	0		
		0.9304	1		0.93	1		

Now, we are ready to compare the workings of LMA and LSD to see if they give the same results. We present two cases of mixture DNA samples, for which the genotypes of both contributors are known. LSD gives all correct result, while LMA gives wrong result at several loci in both cases. The first mixture DNA is selected from Reference [21] shown in Table 5.32. Again, in using LMA, the genotype of one contributor is a necessary input to calculate the other contributor's genotype; whereas in LSD, only the allele peak information is necessary to resolve the DNA mixture sample. The reference genotype is only used after LSD is run for those loci where resolution is not clear enough to allow for an unambiguous choice of the best-fit profiles.

The LSD suggested genotype resolution result, presented in Table 5.33, shows that all six loci are resolved correctly by pulling out the top-ranked genotype combination pair at each locus. Note that the best-fit mass ratio calculated for the six loci varies from 1: 4.76 (FGA) to 1: 11.39 (D8S1179). Note that such variation does not affect the genotype resolution result.

In contrast, results from using the LMA algorithm, as implemented by us, are shown in Table 5.34. LMA resolution results show that only two out of the given six loci are correctly resolved. The shaded boxes in the table indicate the incorrectly resolved loci. The first three columns of the table are the allele peak data. Note that in the column labeled 'Normalized allele peak areas', the areas are normalized such that the sum is 2, as stipulated by the LMA algorithm. The fourth column is the known genotype of person A at each locus; the fifth column is the vector g_A that represents the genotype of person

Table 5.32: Allele peak data and the genotypes of the two contributors at six loci of a mixed DNA sample with the mass ratio of 1:10 prior to PCR amplification, as presented in Reference [21].

Locus	Alleles in the Mixture	Genotype Combination		Allele Peak Areas
		Suspect	Victim	
TH01	8 9.3	8, 9.3	9.3, 9.3	17441 22368
D21S11	59 65 67 70	67, 70	59, 65	1226 1434 8816 8894
D18S51	13 16 17	13, 13	16, 17	38985 1914 1991
D8S1179	10 11 14	10, 14	10, 11	6416 383 5659
VWA	16 17 18 19	16, 18	17, 19	4669 931 4724 188
FGA	21 22 23	21, 22	21, 23	16099 10538 1014

Table 5.33: LSD suggested genotype resolution result of a mixed DNA sample presented in Reference [21]. The top-ranked genotype combinations at all loci are consistent with the true genotypes of the two contributors, which are known and used as references.

Locus	Alleles in the Mixture	Peak Area Ratio Measured	LSD Suggested Genotype Resolution Result			
			Person 1	Person 2	Mass Ratio Calculated	Remarks
TH01	8 9.3	1: 1.28	9.3, 9.3	8, 9.3	1: 7.08	With confidence
D21S11	59 65 67 70	1: 1.17: 7.19: 7.25	59, 65	67, 70	1: 6.66	With confidence
D18S51	13 16 17	20.37: 1: 1.04	16, 17	13, 13	1: 9.98	With confidence
D8S1179	10 11 14	16.75: 1: 1.476	10, 11	10, 14	1: 11.39	With confidence
VWA	16 17 18 19	24.84: 4.95: 25.13: 1	17, 19	16, 18	1: 8.39	With confidence
FGA	21 22 23	15.88: 10.39: 1	21, 23	21, 22	1: 4.76	With confidence

Table 5.34: LMA suggested genotype resolution result of the DNA mixture data given in Table 5.32. Only two out of six loci were resolved correctly. The genotype of one contributor has to be known in advance. Shaded areas indicate that the LMA results at these loci do not correspond to the true genotype profile. From iterative search, 86% is identified as the optimal mass fraction for person A in the DNA mixture.

Locus	Alleles in the Mixture	Normalized Allele Peak Areas	Genotype of Person A Assumed Known		LMA Suggested Result for Person B		
			Alleles	Vector g_A	Vector g_B		Alleles
Th01	8	0.88	8	1	0.14	0	9.3
	9.3	1.12	9.3	1	1.86	2	
D21S11	59	0.12	67 70	0	0.86	1	59 65
	65	0.14		0	1.00	1	
	67	0.87		1	0.07	0	
	70	0.87		1	0.07	0	
D18S51	13	1.82	13	2	0.71	1	13 16 17
	16	0.09		0	0.64	1	
	17	0.09		0	0.64	1	
D8S1179	10	1.03	10 14	1	1.21	1	10
	11	0.06		0	0.43	0	
	14	0.91		1	0.36	0	
VWA	16	0.89	16 18	1	0.21	0	17
	17	0.18		0	1.29	1	
	18	0.90		1	0.29	0	
	19	0.03		0	0.21	0	
FGA	21	1.16	21 22	1	2.14	2	21 21 23
	22	0.76		1	-0.71	-1	
	23	0.07		0	0.50	1	

A at all six loci. The optimal mass proportion of Person A in the mixture sample, searched from 0.01 to 0.99, was identified as 86% from Figure 5.17. Our implementation of LMA then calculated a vector g_B according to Equation 3.12. The result is shown in the sixth column. According to the LMA algorithm, the final genotype of person B is to come from the rounding of the elements of the calculated vector g_B to the nearest integers. If we ignore those negative numbers in vector g_B , the genotype of person B for the six loci after rounding of the elements of the vector g_B is shown in the seventh column. As indicated, the LMA genotype resolution result at four loci D18S51, D8S1179, VWA, and FGA are not correct.

A second set of DNA mixture sample data was used to compare the effectiveness of LMA and LSD in resolving 2-people mixtures and are shown in Table 5.35, provided by the Promega Corporation with the known mass ratio of 9:1 prior to PCR. The true genotypes of both contributors are shown and will be used to verify the resolution results.

The LSD suggested genotype resolution results, presented in Table 5.36, show that all nine loci are correctly resolved by pulling out the top-ranked genotype combination pair at each locus. Note that the optimal mass ratio calculated for each locus varies somewhat from locus to locus, ranging from 5.8:1 to 18:4:1, and the average being 9.7:1, which is very close to the known mass ratio prior to amplification. The resolution at all nine loci is strong in that a clear separation in fitting errors exists between the top and the next ranked genotype pairs.

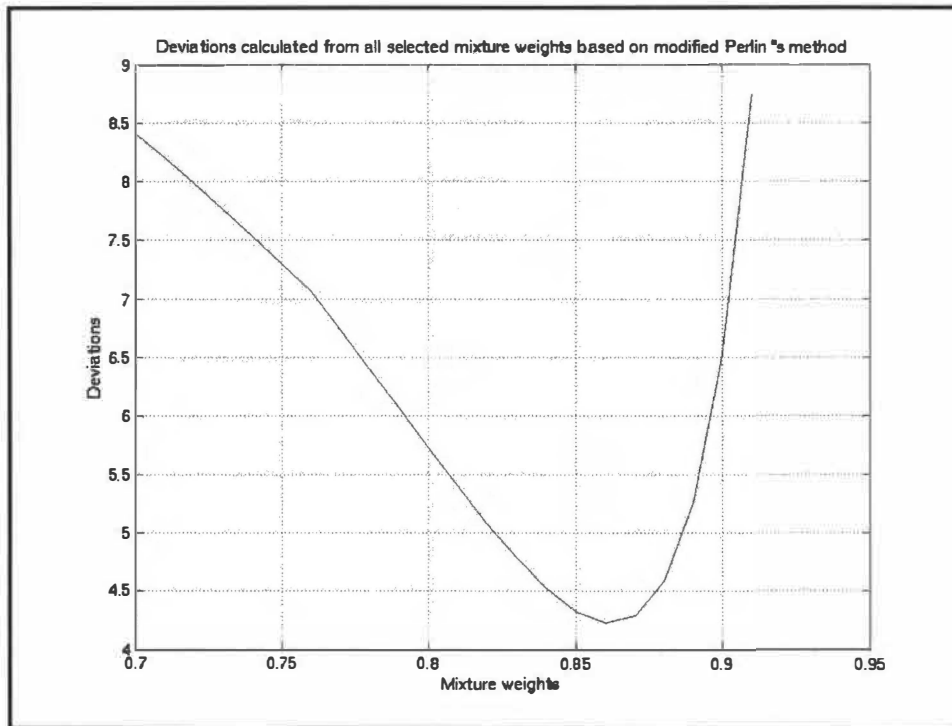


Figure 5.17: Sum of squared deviation as a function of the set of assumed mass fractions of LMA on mixture data presented in Table 5.32. The mass fraction of 86% is identified as the optimal mass proportion of person A.

Table 5.35: Allele peak data and the true genotypes of both contributors of a DNA mixture sample, provided by the Promega corporation, with the known mass ratio of 9:1, prior to amplification. This data set is used to compare the applicability of LSD and LMA in yielding the correct genotype resolution.

Locus	Alleles in the Mixture	Ratio of Allele Peak Data	True Genotypes	
			Person 1	Person 2
TH01	6, 9, 9.3	19.79:1:1.19	6, 6	9, 9.3
D18S51	12, 13, 15, 17	14.23:1.40:11.34: 1	12, 15	13, 17
D5S818	10, 11, 12	3.08:1:2.97	10, 12	11, 11
D13S17	9, 10, 11	1:5.65:5.10	10, 11	9, 10
D7S820	9, 11, 12	54.10:1.94:1	9, 9	11, 12
CSF1PO	10, 11, 12	1:11.74:12.42	11, 12	10, 12
VWA	15, 17, 18	3.89:1:4.28	15, 18	17, 17
D8S1179	12, 13, 15	4.40:1:3.99	12, 15	13, 13
FGA	21, 24, 25	10.58:12.76:1	21, 24	24, 25

Source of mixture data: P. Newman, "Allele Peak Data for DNA Mixture samples," Promega Corporation, Personal Communication, September 2001.

Table 5.36: LSD suggested genotype resolution result of the DNA mixture data given in Table 5.35. All 9 loci were unambiguously resolved correctly by selecting the top-ranked genotype combination pair at each locus. The average mass ratio is 9.7:1.

Locus	Alleles in the Mixture	LSD Suggested Genotype Resolution Result		
		Person 1	Person 2	Mass Ratio Calculated
TH01	6, 9, 9.3	6, 6	9, 9.3	9.0:1
D18S51	12, 13, 15, 17	12, 15	13, 17	10.7:1
D5S818	10, 11, 12	10, 12	11, 11	6.0:1
D13S17	9, 10, 11	10, 11	9, 10	5.8:1
D7S820	9, 11, 12	9, 9	11, 12	18.4:1
CSF1PO	10, 11, 12	11, 12	10, 12	13.0:1
VWA	15, 17, 18	15, 18	17, 17	8.1:1
D8S1179	12, 13, 15	12, 15	13, 13	8.4:1
FGA	21, 24, 25	21, 24	24, 25	7.9:1

In contrast, results from using the LMA algorithm as implemented by us, shown in Table 5.37, indicate that only seven out of nine loci were correctly resolved. The optimal mass proportion of person A in the mixture is searched from 0.01 to 0.99. Figure 5.18 Indicates that 88% should be identified as the optimal one, which is very close to the known value of 90%. The LMA genotype resolution results at loci D18S51 and D7S820 are not correct. Note that these two loci correspond to those loci, from LSD, to have the best-fit mass ratio the most removed from the average ratio of 9.7:1, indicating the imposed fit for these loci with the mass fractions of 88% or a mass ratio of 7.3:1 (0.88:0.12) is a poor fit, although it appear to be a good fit for other loci. However, it cannot be known, *a priori*, that these loci would give a ‘poorer’ fit than that of others; therefore, would be judged as less trustworthy and would not be included for final consideration.

From these two examples, it appears that LMA is not as feasible and accurate as LSD in resolving mixed DNA samples, not to mention that LMA is more cumbersome to implement as well as requiring much more additional input information. In summary, for resolving 2-people mixture, four main differences between LMA and LSD exist as shown in Table 5.38.

Table 5.37: LMA suggested genotype resolution result of the DNA mixture data given in Table 5.35. Seven out of nine loci were resolved correctly. The genotype of one contributor has to be known in advance. Shaded areas indicate that the LMA results at these loci do not correspond to the true genotype profile. The mass fraction of 88% is identified as the optimal mass proportion of Person A in the mixture.

Locus	Alleles in the Mixture	Normalized Allele Peak Areas	Genotype of Person A Assumed Known		LMA Suggested Result for Person B		True Genotype of Person B
			Alleles	Vector g_A	Vector g_B	Alleles	
TH01	6 9 9.3	1.800: 0.091: 0.108	6	$\begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0.33 \\ 0.76 \\ 0.90 \end{bmatrix}$	9 9.3	9 9.3
D18S51	12 13 15 17	1.017: 0.100: 0.811: 0.072	12 15	$\begin{bmatrix} - \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1.14 \\ 0.83 \\ -0.58 \\ 0.60 \end{bmatrix}$	12 13 17	13 17
D5S818	10 11 12	0.876: 0.284: 0.840	10 12	$\begin{bmatrix} - \\ 1 \\ 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -0.03 \\ 2.37 \\ -0.33 \end{bmatrix}$	11 11	11 11
D13S17	9 10 11	0.170: 0.962: 0.868	10 11	$\begin{bmatrix} - \\ 0 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1.42 \\ 0.68 \\ -0.10 \end{bmatrix}$	9 10	9 10
D7S820	9 11 12	1.897: 0.068: 0.035	9	$\begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1.14 \\ 0.57 \\ 0.29 \end{bmatrix}$	9 11	11 12
CSF1PO	10 11 12	0.080: 0.933: 0.987	11 12	$\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.66 \\ 0.44 \\ 0.89 \end{bmatrix}$	10 12	10 12
VWA	15 17 18	0.858: 0.220: 0.922	15 18	$\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} -0.18 \\ 1.83 \\ 0.35 \end{bmatrix}$	17 17	17 17
D8S1179	12 13 15	0.937: 0.213: 0.850	12 15	$\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.48 \\ 1.78 \\ -0.25 \end{bmatrix}$	13 13	13 13
FGA	21 24 25	0.869: 1.048: 0.082	21 24	$\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} -0.09 \\ 1.40 \\ 0.68 \end{bmatrix}$	24 25	24 25

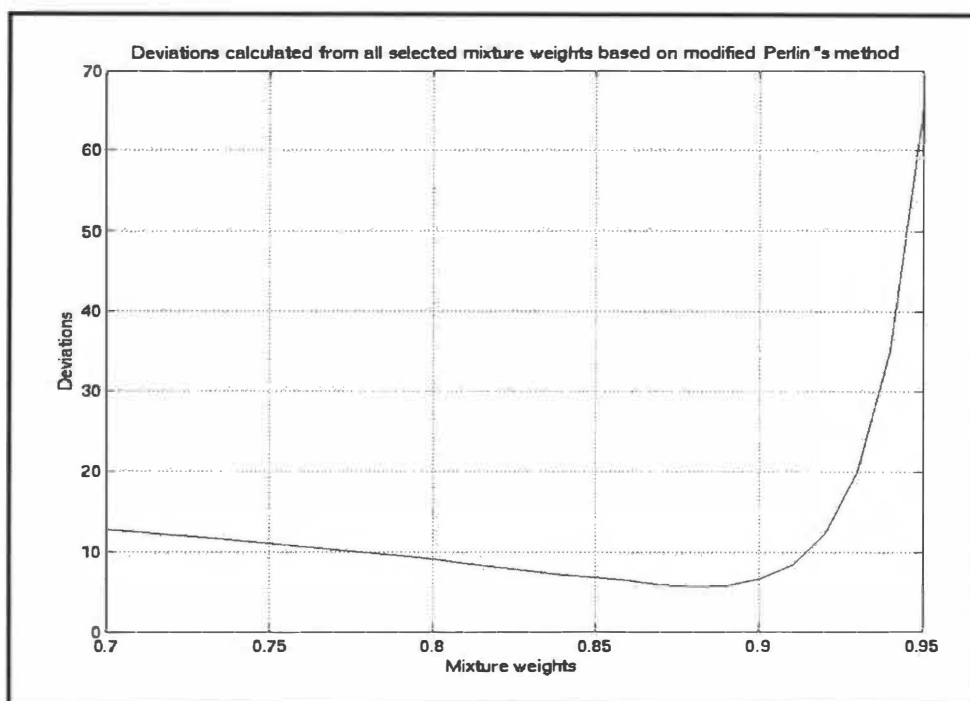


Figure 5.18: Sum of squared deviation as a function of the set of assumed mass fractions of LMA on mixture data presented in Table 5.35. The mass fraction of 88% is identified as the optimal mass proportion of person A. Such result is consistent with the known true value of 90%.

Table 5.38: Summary of comparison between LMA and LSD in resolving 2-people DNA mixtures.

	LMA	LSD
Necessary input information	Allele peak data and one of the two contributors' genotypes	Only the allele peak data
Computation approach	Iterative searching through a large number of assumed mass fractions	Direct computation in one step for the best-fit mass ratio
Loci processing	Treat all loci as one entity	Locus by locus, independently
Optimal mass fraction	Formulated to be the same for all loci	Different for each locus; optimal with respect to each locus

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this document, the least square methodology was presented as a novel application to resolve two-people mixture STR-DNA samples using only the allele peak quantitative information. At each locus, each of possible genotype combination pairs is, in turn, fitted, in the least square sense, to the observed allele peak areas (or heights). The one with the smallest fitting error is tentatively identified as the most likely one. The corresponding best-fit relative DNA mass ratio is also noted, to be used in comparison to those derived for other loci. The degree of confidence to the fitted genotype pair is accessed by the relative error in the fit, as well as by the consistency in the corresponding relative DNA mass ratio. A final composite genotype profile for each of the two contributors is formed after reviewing and processing the LSD results for all loci. Finally, the reference profile, when available, is brought in to check the hypothesized profiles and to pick among alternative choices when multiple possible profile pairs exist at a locus.

From a sensitivity study, it was observed that about 20% imbalance between the relative allele peak areas (heights) can be tolerated before a wrong fit genotype combination pair presents itself as the best-fit one. Further testing of LSD with 60 real mixture samples show that LSD is capable of resolving mixtures correctly provided the following conditions are met by the samples:

- Only true alleles are called in the mixture; that no artifact allele is identified as a true allele.

- No allele drop-out has occurred in the mixture.
- No peak area saturation has occurred due to overloading of samples to distort the inherent proportionality between allele peak areas and DNA mass.
- No severe imbalance exists between the peak areas of heterozygous alleles of a locus and the peak area remains more or less proportional to the corresponding relative DNA mass in the mixture.
- The mass ratio of the DNA mixture is away from 1:1.

Since the LSD algorithm was developed and publicly presented in 2001, more and more forensic scientists realize its strong potential to be used as an analytical tool in resolving DNA mixture samples. Currently, it is being beta-tested by a group of forensic scientists around the country, selected by the FBI. In the near future, it will be released to all the forensic community. It is expected that LSD will greatly strengthen the confidence and enhance the capability of the forensic scientists in the identification of the perpetrator of the crime, when used in conjunction with a DNA analyst's own experience in resolving mixture samples.

To date, LSD has been formulated to analyze two-people DNA mixture samples. Extension to resolving 3-people mixture DNA is underway in our laboratory. There are more possible genotype combinations to consider for a three-person DNA mixture. Therefore, deconvolving and interpreting three-people mixtures is much more complicated. The mathematical formulation for 3-people LSD would have to be heavily

modified from that of 2-people, in order to avoid the high sensitivity of fit to allele peak imbalance, as well as limitations inherent with the underlying linear algebra mathematics.

LIST OF REFERENCES

- [1]: J. Samuels and C. Asplen, *The Future of Forensic DNA Testing: Predictions of the Research and Development Working Group*, The National Institute of Justice, November, 2000.
- [2]: National Research Council, *The Evaluation of Forensic DNA Evidence*, National Academy Press, 1996.
- [3]: T. Clayton, J. Whitaker, R. Sparkes, and P. Gill, "Analysis and Interpretation of Mixed Forensic Stains Using DNA STR Profiling," *Forensic Science International*, 91(1998): 55-70.
- [4]: JM. Curran, CM. Triggs, J. Buckleton, and BS. Weir, "Interpreting DNA Mixtures in Structured Populations," *Journal of Forensic Science*, 44-5(1999): 987-995.
- [5]: P. Gill, R. Sparkes, R. Pinchin, T. Clayton, J. Whitaker, and J. Buckleton, "Interpreting Simple STR Mixtures Using Allele Peak Areas," *Forensic Science International*, 91(1998): 41-53.
- [6]: T. Wang, N. Xue, M. Radar, and J.D. Birdwell, "Least Square Deconvolution (LSD) of Mixture STR/DNA Using Peak Area / Height Data," The 7th CODIS User's Conference, Arlington, VA, October, 2001.
- [7]: M. Farley and J. Harrington, *Forensic DNA Technology*, Lewis Publishers, INC., 1991.
- [8]: K. Inman and N. Rudin, *An Introduction to Forensic DNA Analysis*, CRC Press, 1997.
- [9]: P. Gill, J. Whitaker, C. Flaxman, N. Brown, and J. Buckleton, "An Investigation of the Rigor of Interpretation Rules for STRs Derived from Less Than 100pg of DNA," *Forensic Science International*, 112(2000): 17-40.

- [10]: J. Butler, *Forensic DNA Typing*, Academic Press, 2001.
- [11]: B. Budowle, "Probabilities of Alleles' Distribution at 13 Loci", Personal Communication in October 1997.
- [12]: P. Gill, R. Sparkes, and C. Kimpton, "Development of Guidelines to Designate Alleles Using an STR Multiplex System," *Forensic Science International*, 89(1997): 185-197.
- [13]: B. Brown, Director of FBI National CODIS Project, Personal Communication in January, 2002.
- [14]: P. Billings, *DNA on Trial, Genetic Identification and Criminal Justice*, Cold Spring Harbor Laboratory Press, 1992.
- [15]: I. Evett, and B. Weir, *Interpreting DNA Evidence, Statistical Genetics for Forensic Scientists*, Sinauer Associates Inc., 1998.
- [16]: G. Strang, *Linear Algebra and Its Applications (Third edition)*, Harcourt Brace Jovanovich College Publisher, 1986.
- [17]: J. Birdwell, "Analysis of Forensic Mixtures," *LIT/FBI Weekly Seminar Series*, April 27, 2001.
- [18]: M. Perlin and B. Szabady, "Linear Mixture Analysis: A Mathematical Approach to Resolving Mixed DNA Samples," *Journal of Forensic Science*, 46-6(2001): 1372-1378.
- [19]: P. Graham, "Allele Peak Data of DNA Mixture from Real Cases," Texas Department of Public Safety, Corpus Christi, October, 2001.
- [20]: "Web-LSD Software: [https://www.lit.net/lsd/\(restricted\)](https://www.lit.net/lsd/(restricted))", Developed by Laboratory for Information Technologies, the University of Tennessee, Knoxville, May, 2002.

- [21]: I. Evett, P. Gill, and J. Lambert, “ Taking Account of Peak Areas When Interpreting Mixed DNA Profiles,” *Journal of Forensic Science*, 43-6(1998): 62-69.
- [22]: R. Wickenheiser, “Allele Peak Data of DNA Mixture from Real Cases,” Acadiana Crime Laboratory, Louisiana, September, 2001.
- [23]: P. Newman, “Allele Peak Data of DNA Mixture”, Promaga Corporation, September, 2001.
- [24]: Personal Communication with DNA analysts around the country and through our work in resolving mixed DNA samples.

VITA

Ning Xue was born in Fushun, China, on September 6th, 1977. After graduating from Fushun No. 2 High School in July 1995, she passed the National College Entrance Examination and enrolled in a Bachelor of Science program in the Chemical Engineering Department at Zhejiang University. During her undergraduate program, she spent half a year working as a research assistant in the Polymer Reaction National Key Laboratory at Zhejiang University and obtained her Bachelor of Science degree in Chemical Engineering in July of 1999. From August 1999 to May 2000, she worked as a chemical engineer in Fushun Petrochemical Research Institution, China. In August, 2000, she entered the graduate program of the chemical engineering department at the University of Tennessee, Knoxville, and received her Master of Science Degree in Chemical Engineering in August of 2002.