

To the Graduate Council:

I am submitting herewith a thesis written by Puneet Yadav entitled “Increasing the speed and efficiency of search in FBI/CODIS DNA database through multivariate statistical clustering approach and development of a similarity ranking scheme.” I have examined the final 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:

J.D. Birdwell

J.R. Collier

Mark Rader

Accepted for the Council:

Dr. Anne Mayhew
Interim Vice Provost and
Dean of The Graduate School

(Original signatures are on file in the Graduate Admissions and Records Office).

**INCREASING THE SPEED AND EFFICIENCY OF SEARCH IN
FBI/CODIS DNA DATABASE THROUGH MULTIVARIATE
STATISTICAL CLUSTERING APPROACH AND
DEVELOPMENT OF A SIMILARITY RANKING SCHEME**

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

Puneet Yadav

May 2001

ACKNOWLEDGEMENTS

This work has been supported by the U.S. Department of Justice, Federal Bureau of Investigation (FBI) under contract J-FBI-98-083. 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.

I would like to thank my advisor, Dr. Tsewei Wang for her continued guidance and support during the preparation of this thesis. I also want to thank the other members of my thesis committee, Dr. J.D. Birdwell, Dr. Mark Rader, and Dr. J.R. Collier for their advice and help.

ABSTRACT

A new method has been developed to create and maintain a search tree-structured index to multidimensional data using naturally occurring patterns and clusters within the data, and thereby allows the implementation of efficient search and retrieval strategies in a database. This method was applied to a DNA database, which was developed by the FBI for forensic uses. A set of 10,000 DNA/STR profiles based on the STR allele probability distribution density for the Caucasians has been generated for the sixteen loci. The resulting allele distribution has been analyzed using Multivariate Statistical analysis, in specific; the Principal Component Analysis (PCA) approach was employed to detect clustering patterns among the profiles. The analysis revealed that with the choice of some loci-pairs (such as d13s17 and d16s539) good and distinct clusters were obtainable. Members within each distinct cluster were further studied to determine the attributes that made them distinct from all members of other clusters. The PCA analysis results with a real DNA/STR dataset also showed similar clustering patterns.

In order to rank order the profiles from a search process as to their similarity to that of the target profile, a new Similarity Index (SI) parameter has been developed. The Similarity Index was successfully tested on a small (126) and a large (1026) dataset. Further, a Shuffling Index was developed to study the sensitivity of the Similarity Index to the selection of weights used in the similarity index sub-parameters. Results show that the similarity ranking of profiles remain stable over a wide range of weights.

TABLE OF CONTENTS

CHAPTER	PAGE
1. INTRODUCTION.....	1
1.1. MOTIVATION.....	2
1.2. PROBLEM STATEMENT.....	4
2. BACKGROUND.....	5
2.1. CODIS: THE NATIONAL DNA DATABASE.....	5
2.2. BASIC GENETIC PRINCIPLES.....	7
2.3. DNA TYPING FOR HUMAN IDENTIFICATION PURPOSES.....	11
2.4. APPLICATIONS OF DNA TYPING AND PROFILING.....	13
2.4.1. PCR AMPLIFICATIONS.....	13
2.4.2. SHORT TANDEM REPEATS (STRs).....	15
2.4.3. PROBABILITY DISTRIBUTION OF ALLELES IN THE SIXTEEN CODIS CORE LOCI.....	17
2.5. DNA AND DATABASE.....	18
2.6. DNA IN COURTS: LEGAL ISSUES.....	37
3. REVIEW OF THE LITERATURE.....	39
3.1. MULTIVARIATE STATISTICAL ANALYSIS USED IN CLUSTERING.....	39
3.2. MEASURES OF GENETIC DISTANCE.....	40

CHAPTER	PAGE
4. METHODOLOGY OF APPROACH.....	43
4.1. CLUSTERABILITY STUDY: OVERVIEW OF APPROACH.....	43
4.1.1. SYNTHETIC DNA PROFILES GENERATION AND DNA PROFILES REPRESENTATION.....	45
4.1.2. DNA/STR CLUSTERABILITY STUDY.....	50
4.1.2.1. CLUSTERING BY PRINCIPAL COMPONENT ANALYSIS (PCA).....	50
4.1.2.2. CLUSTER MEMBERSHIP ANALYSIS USING THE K-MEANS ALGORITHM.....	58
4.1.2.3. ALLELE DISTRIBUTION PATTERN OF MEMBERS CONTAINED WITHIN EACH CLUSTER.....	65
4.2. DEVELOPMENT OF A SIMILARITY INDEX (SI) TO RANK ORDER PROFILES.....	65
4.2.1. THE SIMILARITY INDEX.....	66
4.2.2. DNA PROFILE GENERATION FOR TESTING THE SIMILARITY INDEX.....	74
4.3. SENSITIVITY ANALYSIS OF SIMILARITY INDEX TO THE SELECTION OF WEIGHTS FOR THE SIMILARITY INDEX SUB- PARAMETERS.....	77
4.3.1. SELECTION OF WEIGHTS.....	78

CHAPTER	PAGE
4.3.2. RANKING SENSITIVITY STUDY: THE SHUFFLING INDEX AND SHUFFLING INDEX CONTOUR PLOTS.....	82
5. RESULTS AND DISCUSSIONS.....	84
5.1. CLUSTERING BY MULTIVARIATE STATISTICAL ANALYSIS.....	84
5.1.1. IDENTIFICATION OF CLUSTERS FOR DNA/STR PROFILES FOR CERTAIN TWO-LOCI COMBINATIONS.....	86
5.1.2. EXAMPLES OF LOCI PAIRS THAT GIVE GOOD CLUSTERS ...	88
5.1.3. STUDY OF THE PCA OF DNA/STR PROFILES WITH INFORMATION AT THE d13s17-d16s539 LOCI-PAIR THAT GIVE GOOD CLUSTERABILITY.....	95
5.1.4. CLUSTER MEMBERSHIP VALIDATION USING K-MEANS ALGORITHM FOR PROFILES WITH DNA/STR INFORMATION AT THE d13s17-d16s539 LOCI-PAIR.....	113
5.1.4.1. TESTING THE THEORY OF CLUSTER FORMATION BY PREDICTING THE APPROXIMATE CENTERS OF THE CLUSTERS.....	135
5.1.5. CLUSTERING ANALYSIS OF A SMALL REAL DNA/STR DATASET.....	139
5.1.6. EXAMINATION OF LOCI-PAIR THAT GIVE POOR CLUSTER FORMATION.....	147

CHAPTER	PAGE
5.1.7. INCORPORATION OF CLUSTERS INTO THE DECISION TREE FOR SEARCH AND RETRIEVAL OF DNA PROFILES.....	153
5.1.8. SUMMARY FOR THE CLUSTERABILITY OF PROFILES BY THE PRINCIPAL COMPONENT ANALYSIS (PCA) APPROACH.....	158
5.2. DEVELOPMENT AND UTILITY OF A SIMILARITY INDEX (SI) PARAMETER FOR RANKING CANDIDATE PROFILES.....	159
5.2.1. ROBUST RANKING OF A SET OF SYNTHETIC DATA PROFILES.....	159
5.2.2. SENSITIVITY OF RANKING TO THE CHOICE OF WEIGHTS FOR SIMILARITY INDEX SUB-PARAMETERS.....	176
6. SUMMARY, CONCLUSION AND RECOMMENDATIONS FOR FUTURE WORK.....	194
BIBLIOGRAPHY.....	197
VITA.....	205

LIST OF TABLES

TABLE	PAGE
2.1. Relative Frequency of Occurrence of Alleles at the csf1po Locus for Caucasians.....	19
2.2. Relative Frequency of Occurrence of Alleles at the d18s51 Locus for Caucasians.....	20
2.3. Relative Frequency of Occurrence of Alleles at the d13s317 Locus for Caucasians.....	21
2.4. Relative Frequency of Occurrence of Alleles at the d16s539 Locus for Caucasians.....	22
2.5. Relative Frequency of Occurrence of Alleles at the dqalpha Locus for Caucasians.....	23
2.6. Relative Frequency of Occurrence of Alleles at the d3s1358 Locus for Caucasians.....	24
2.7. Relative Frequency of Occurrence of Alleles at the d5s818 Locus for Caucasians.....	25
2.8. Relative Frequency of Occurrence of Alleles at the d7s820 Locus for Caucasians.....	26
2.9. Relative Frequency of Occurrence of Alleles at the d8s1179 Locus for Caucasians.....	27
2.10. Relative Frequency of Occurrence of Alleles at the th01 Locus for Caucasians.....	28
2.11. Relative Frequency of Occurrence of Alleles at the tpox Locus for Caucasians.....	29
2.12. Relative Frequency of Occurrence of Alleles at the vWa Locus for Caucasians.....	30
2.13. Relative Frequency of Occurrence of Alleles at the fga Locus for Caucasians.....	31

TABLE	PAGE
2.14. Relative Frequency of Occurrence of Alleles at the d1s80 Locus for Caucasians.....	33
2.15. Relative Frequency of Occurrence of Alleles at the d21s11 Locus for Caucasians.....	35
4.1. Example of Binary Representation of DNA/STR Profile for Twenty Individuals at the d13s317 Locus. ‘1’ Represents the Presence of and ‘0’ Indicates the Absence of an Allele.....	51
4.2. Scores of Candidate Profiles onto the Two Principal Components (V2 and V3).....	63
4.3. Types of Synthetic Profiles (126) used to Study the Feasibility of Similarity Index.....	75
4.4. Types of Synthetic Profiles (1026) used to Study the Feasibility of Similarity Index.....	76
5.1. Example of Binary Representation of Twenty DNA/STR Profiles with Alleles at the d13s317Locus.....	85
5.2. Combinations of Two-Loci that give Good Clusters after Visual Inspection of the Scores Plot.....	87
5.3. Relative Frequency of Occurrence of Alleles at the d13s317 Locus for Caucasian Population.....	96
5.4. Relative Frequency of Occurrence of Alleles at the d16s539 Locus for Caucasian Population.....	97
5.5. Possible Combinations of Presence/Absence of Allele 11 and 12 that Result in Clustering for Loci-Pair d13s317 and d16s539.....	101
5.6. Cumulative Contribution of all Principal Components in the PCA Analysis of the Data Matrix.....	107
5.7. The 2 nd and 3 rd Principal Components of the PCA Model of 10,000 Synthetic Caucasian DNA Profiles with Alleles in d13s317 and d16s539.....	110

TABLE**PAGE**

5.8.	Number/Percentage of Samples in Each Cluster of the 10,000 Synthetically Generated Profiles for d13s317 and d16s539 Loci-Pair.....	114
5.9.	The Presence (1) or Absence (0) of Alleles 11 and 12 in the d13s317 and d16s539 Loci for each Scores Clusters as shown in Figure 5.1.....	121
5.10.	Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 1 for d13s317-16s539 Loci Pair. The Allele Distribution Pattern at the Four Dominant Sites is (1 0 0 1).....	123
5.11.	Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 2 for d13s317-16s539 Loci Pair. The Allele Distribution Pattern at the Four Dominant Sites is either (1 0 0 1) or (1 0 1 1).....	124
5.12.	Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 3 for d13s317-16s539 Loci Pair. The Allele Distribution Pattern at the Four Dominant Sites is (1 0 1 0).....	125
5.13.	Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 4 for d13s317-16s539 Loci Pair. The Allele Distribution Pattern at the Four Dominant Sites is either (1 1 1 0) or (0 0 0 1).....	126
5.14.	Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 5 for d13s317-16s539 Loci Pair. The Allele Distribution Pattern at the Four Dominant Sites is either (1 1 1 1), (0 0 1 1), (1 1 0 0) or (0 0 0 0).....	127
5.15.	Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 6 for d13s317-16s539 Loci Pair. The Allele Distribution Pattern at the Four Dominant Sites is either (1 1 1 0) or (0 0 1 0).....	128
5.16.	Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 7 for d13s317-16s539 Loci Pair. The Allele Distribution Pattern at the Four Dominant Sites is (0 1 0 1).....	129
5.17.	Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 8 for d13s317-16s539 Loci Pair. The Allele Distribution Pattern at the Four Dominant Sites is either (0 1 1 1) or (0 1 0 0).....	130
5.18.	Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 9 for d13s317-16s539 Loci Pair. The Allele Distribution Pattern at the Four Dominant Sites is (0 1 1 0).....	131

TABLE	PAGE
5.19. The Coordinates of the True Centers of the Nine Clusters shown in Figure 5.1.....	136
5.20. Calculation of the Approximate Cluster Center Coordinates from the Sixteen Possible Allele Distribution Patterns at the Four Dominant Sites.....	137
5.21. Comparison of Number/Fraction of Profiles in Each Cluster of the Synthetic and Real Data Set for d13s317 and d16s539 Loci Pair.....	146
5.22. Types of Synthetic Profiles (126) used to Study the Feasibility of Similarity Index.....	162
5.23. Types of Synthetic Profiles (1026) used to Study the Feasibility of Similarity Index.....	163
5.24. Results Obtained by Applying the Similarity Index on the Synthetic Data Set of 126 Profiles for Six Loci and Weights ($\alpha_1=0.70$, $\alpha_2=0.20$, $\alpha_3=0.10$), Showing Ranks from 1-20.....	164
5.25. Results Obtained by Applying the Similarity Index on the Synthetic Data Set of 126 Profiles for Six Loci and Weights ($\alpha_1=0.70$, $\alpha_2=0.20$, $\alpha_3=0.10$), Showing Ranks from 21-40.....	165
5.26. Results Obtained by Applying the Similarity Index on the Synthetic Data Set of 126 Profiles for Six Loci and Weights ($\alpha_1=0.70$, $\alpha_2=0.20$, $\alpha_3=0.10$), Showing Ranks from 41-60.....	166
5.27. Results Obtained by Applying the Similarity Index on the Synthetic Data Set of 126 Profiles for Six Loci and Weights ($\alpha_1=0.70$, $\alpha_2=0.20$, $\alpha_3=0.10$), Showing Ranks from 61-80.....	167
5.28. Results Obtained by Applying the Similarity Index on the Synthetic Data Set of 126 Profiles for Six Loci and Weights ($\alpha_1=0.70$, $\alpha_2=0.20$, $\alpha_3=0.10$), Showing Ranks from 81-100.....	168
5.29. Top Twenty Five Ranked Profiles obtained from Similarity Index for Different Set of Weights (Table 5.30).....	172
5.30. Different Set of Weights for PML, PMA, and PBPD ^c used to rank 126 Candidate Profiles using Similarity Index Parameter.....	173

TABLE	PAGE
5.31. Comparison of Allele Distribution of Candidate 4 and 21 for the Set of Weights of ($\alpha_1=0.90$, $\alpha_2=0.00$, $\alpha_3=0.10$) (i.e. Case 1, Table 5.29).....	175

LIST OF FIGURES

FIGURE	PAGE
2.1. CODIS Architecture.....	6
2.2. Location of the Two Types of DNA [42].....	8
2.3. DNA HELIX [16].....	10
2.4. DNA bases line up in Pairs; An Example of a Short Length of Double Stranded DNA.....	10
2.5. Flowchart for Forensic DNA typing [16].....	12
2.6. Overview of the Steps of Polymerase Chain Reaction (PCR) [42].....	14
2.7. Allele Frequency Distribution of csf1po Locus of Caucasian DNA.....	19
2.8. Allele Frequency Distribution of d18s51 Locus of Caucasian DNA.....	20
2.9. Allele Frequency Distribution of d13s317 Locus of Caucasian DNA.....	21
2.10. Allele Frequency Distribution of d16s539 Locus of Caucasian DNA.....	22
2.11. Allele Frequency Distribution of dqalpha Locus of Caucasian DNA.....	23
2.12. Allele Frequency Distribution of d3s1358 Locus of Caucasian DNA.....	24
2.13. Allele Frequency Distribution of d5s818 Locus of Caucasian DNA.....	25
2.14. Allele Frequency Distribution of d7s820 Locus of Caucasian DNA.....	26
2.15. Allele Frequency Distribution of d8s1179 Locus of Caucasian DNA.....	27
2.16. Allele Frequency Distribution of th01 Locus of Caucasian DNA.....	28
2.17. Allele Frequency Distribution of tpox Locus of Caucasian DNA.....	29
2.18. Allele Frequency Distribution of vWa Locus of Caucasian DNA.....	30

FIGURE	PAGE
2.19. Allele Frequency Distribution of fga Locus of Caucasian DNA.....	32
2.20. Allele Frequency Distribution of d1s80 Locus of Caucasian DNA.....	34
2.21. Allele Frequency Distribution of d21s11 Locus of Caucasian DNA.....	36
4.1 Flowchart for Clusterability Study Using Multivariate Data Analysis for DNA/STR Locus.....	46
4.2 Flowchart for the Development of Similarity Index (SI) for Ranking Candidate Profiles.....	48
4.3 Illustration of Principal Component Analysis Concept, in which PC1 and PC2 Represent the New Principal Component Basis Vectors.....	52
4.4 Normalized Scores of 30 Caucasian-DNA/STR Profiles onto the 2 nd and 3 rd Principal Components in d13s317 and d16s539 Loci.....	64
4.5 Utilization of Similarity Index (SI) in the Search Decision Tree.....	67
4.6 Examples of Ambiguous Profiles.....	69
4.7 Weights of the Similarity Index Sub-Parameter Expressed on a Ternary Equilateral Triangle; PML=Percentage of Matching Loci; PMA=Percentage of Matching Alleles; PBPD ^c =1-PBPD (Percentage of Base Pair Difference), the Sides pq, qr, and rp Represent the Weight (α_1 , α_2 , and α_3) for PML, PMA, and PBPD ^c Respectively.....	79
5.1. Scores of 10,000 Caucasian-DNA/STR Profiles onto the 2 nd and 3 rd Principal Components in d13s317 and d16s539 Loci.....	89
5.2. Scores of 10,000 Caucasian-DNA/STR Profiles onto the 2 nd and 3 rd Principal Components in d13s317 and d1s80 Loci.....	90
5.3. Scores of 10,000 Caucasian-DNA/STR Profiles onto the 2 nd and 3 rd Principal Components in d16s539 and d1s80 Loci.....	91
5.4. Scores of 10,000 Caucasian-DNA/STR Profiles onto the 2 nd and 3 rd Principal Components in d16s539 and d5s818 Loci.....	92
5.5. Scores of 10,000 Caucasian-DNA/STR Profiles onto the 2 nd and 3 rd Principal Components in d13s317 and d5s818 Loci.....	93

FIGURE**PAGE**

5.6.	Scores of 10,000 Caucasian-DNA/STR Profiles onto the 2 nd and 3 rd Principal Components in csf1po and fga Loci.....	94
5.7.	Allele Frequency Distribution of d13s317 Locus of the Caucasian DN.....	96
5.8.	Allele Frequency Distribution of d16s539 Locus of the Caucasian DNA.....	97
5.9.	Joint Probability Density of Two-Allele Distribution of d13s317 Locus.....	99
5.10.	Joint Probability Density of Two-Allele Distribution of d16s539 Locus.....	99
5.11.	Joint Probability Distribution of Allele Pairs for d13s317 and d16s539.....	100
5.12.	Allele Frequency Distribution of the d18s51 Locus of Caucasian DNA Exhibit a More Diffused Probability Density Distribution Pattern and Yield Poor Clusterability in Pair with Another Locus.....	102
5.13.	Scree Plot of the Cumulative Contribution of the Principal Components for 10,000 synthetic profiles for d13s317 and d16s539 Loci-Pair.....	106
5.14.	Make up of the 2 nd Principal Component.....	109
5.15.	Make up of the 3 rd Principal Component.....	109
5.16.	Using K-Means Algorithm to Identify Clusters of the Scores Projected onto the Second and Third Principal Components of d13s317 and d16s539 Loci Pair for 10,000 Synthetically Generated Sample Profiles.....	112
5.17.	Percent of Profile Samples in Cluster 1 that has Each of the Alleles in d13s317 and d16s539 Loci.....	116
5.18.	Percent of Profile Samples in Cluster 2 that has Each of the Alleles in d13s317 and d16s539 Loci.....	116
5.19.	Percent of Profile Samples in Cluster 3 that has Each of the Alleles in d13s317 and d16s539 Loci.....	117
5.20.	Percent of Profile Samples in Cluster 4 that has Each of the Alleles in d13s317 and d16s539 Loci.....	117

FIGURE**PAGE**

5.21.	Percent of Profile Samples in Cluster 5 that has Each of the Alleles in d13s317 and d16s539 Loci.....	118
5.22.	Percent of Profile Samples in Cluster 6 that has Each of the Alleles in d13s317 and d16s539 Loci.....	118
5.23.	Percent of Profile Samples in Cluster 7 that has Each of the Alleles in d13s317 and d16s539 Loci.....	119
5.24.	Percent of Profile Samples in Cluster 8 that has Each of the Alleles in d13s317 and d16s539 Loci.....	119
5.25.	Percent of Profile Samples in Cluster 9 that has Each of the Alleles in d13s317 and d16s539 Loci.....	120
5.26.	Comparison of True and Predicted Cluster Centers for Scores of 10,000 Synthetic Profiles for d13s317 and d16s539 Loci-Pair.....	138
5.27.	Scores Clusters of 10,000 Synthetic Caucasian DNA Data Plotted over Scores of 176 Real Data Samples in d13s317 and d16s539.....	142
5.28.	Scores Clusters of 176 Real Caucasian DNA Data Plotted over Scores of 10,000 Synthetic Data Samples in d13s317 and d16s539.....	143
5.29.	Comparison of Percentages of Population within Each Cluster in d13s317 and d16s539 of Real (176 Samples) and Synthetic (10,000 Samples) Data Set...145	
5.30.	Scores of 10,000 Synthetic Profiles onto the 2 nd and 3 rd Principal Components of DNA Samples in Loci d18s51 and d7s820.....	148
5.31.	Scores of 10,000 Synthetic Profiles onto the 2 nd and 3 rd Principal Components of DNA Samples in Loci d7s820 and fga.....	149
5.32.	Scores of 10,000 Synthetic Profiles onto the 2 nd and 3 rd Principal Components of DNA Samples in Loci fga and tho1.....	150
5.33.	Allele Frequency Distribution of d18s51.....	151
5.34.	Allele Frequency Distribution of d7s820.....	151

FIGURE	PAGE
5.35. Joint Probability Density of 2-Allele Pairs for d18s51 and d7s820 Locus....	152
5.36. Allele Frequency Distribution of fga.....	154
5.37. Joint Probability Density of 2-Allele Pairs for d7s820 and fga Locus.....	155
5.38. Decision Tree Structure Based on Clusters Obtained Using Multivariate Statistical Analysis.....	156
5.39. Utilization of Similarity Index (SI) in the Search Decision Tree.....	160
5.40. Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Change in the Shuffling Index Value for the Top 10 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=1/3$, $\alpha_2=1/3$, $\alpha_1=1/3$], Sample Size =126.....	179
5.41. Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 10 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=1/3$, $\alpha_2=1/3$, $\alpha_1=1/3$], Sample Size =126.....	180
5.42. Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 20 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=1/3$, $\alpha_2=1/3$, $\alpha_1=1/3$], Sample Size =126.....	181
5.43. Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 100 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=1/3$, $\alpha_2=1/3$, $\alpha_1=1/3$], Sample Size =126.....	182
5.44. Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 10 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=0.8$, $\alpha_2=0.1$, $\alpha_1=0.1$], Sample Size =126.....	183
5.45. Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 10 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=0.1$, $\alpha_2=0.8$, $\alpha_1=0.1$], Sample Size =126.....	184

5.46.	Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 10 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=0.1$, $\alpha_2=0.1$, $\alpha_1=0.8$], Sample Size =126.....	185
5.47.	Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 20 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=0.8$, $\alpha_2=0.1$, $\alpha_1=0.1$], Sample Size =126.....	186
5.48.	Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 20 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=0.1$, $\alpha_2=0.8$, $\alpha_1=0.1$], Sample Size =126.....	187
5.49.	Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 20 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=0.1$, $\alpha_2=0.1$, $\alpha_1=0.8$], Sample Size =126.....	188
5.50.	Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 100 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=0.8$, $\alpha_2=0.1$, $\alpha_1=0.1$], Sample Size =126.....	189
5.51.	Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 100 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=0.1$, $\alpha_2=0.8$, $\alpha_1=0.1$], Sample Size =126.....	190
5.52.	Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 100 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=0.1$, $\alpha_2=0.1$, $\alpha_1=0.8$], Sample Size =126.....	191
5.53.	Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 10 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=1/3$, $\alpha_2=1/3$, $\alpha_1=1/3$], Sample Size =1026.....	192

5.54.	Contour plot of Shuffling Index Representing Effect of Changing Weights on the Similarity Index Sub Parameter by Studying the Percentage Change in the Shuffling Index Value for the Top 20 Ranked Candidate Profiles. Reference Weight: [$\alpha_1=1/3$, $\alpha_2=1/3$, $\alpha_3=1/3$], Sample Size =102.....	193
-------	--	-----

CHAPTER 1

INTRODUCTION

DNA analysis is the newest tool for unique personal identification, and has recently found its application in a court of law. Nearly a decade has passed since the potential power of DNA typing was first used in solving crimes, identify paternity, and towards other forensic applications. This new technique has come under severe public scrutiny and tremendous amount of work has to be achieved to show competency for its success. The legal caseworkers and other law enforcement agencies are now taking stern steps in understanding the hidden ideas of molecular biology, genetics and statistics to provide justice to the trials. DNA typing methods are extremely powerful in differentiating one person from another within population and across populations. Thus, it acts as a strong tool in forensic applications. The developed technique of population genetic statistics can be employed to interpret the results of forensic DNA typing. The statistical interpretation of DNA typing lacks 100% clarity, and is always put under rigorous scrutiny. The future of DNA typing depends upon the achievements and success in the field of automation. In forensics, we know that as the size of DNA database grows, improvements and development of new computer technology (software/hardware) will play an important role in the processing, management, and searching of large number of DNA profiles in a database in solving crimes.

1.1. MOTIVATION

The key to harnessing the crime fighting power of DNA profiles is in building up and development of the ability to search through increasingly larger DNA databases. The FBI database of DNA of up to a million (10^6)-convicted criminals from all fifty states has been recently activated to help solve past and future crimes. “In an experiment in 17 states over three years, 193 convicted criminals were matched to DNA profiles taken from crime scenes” [39]. Those successes are likely to occur more and more often as increasingly more DNA samples are entered and stored in the database. The next big step is linking the states DNA database. The FBI has started providing states with the CODIS (Combined DNA Index System) software, which searches for DNA profiles from the stored profiles in the database matching that of a target profile.

The size of the national DNA database is expected to grow eventually to the order of 10^8 profiles in the coming years, and therefore a new search method is necessary to manage the storage and utilization of DNA profile information for the national CODIS database. The present search scheme cannot accommodate efficient search to much beyond a million. The objective of the new search algorithm is also to incorporate the complexities involved in the match specifications. Present search software (CODIS) is based on sequential search and is a time-consuming operation and would break down with increase in the database size. A new search engine for the FBI CODIS DNA database is under development at the University of Tennessee, Knoxville, under the direction of Dr. J.D. Birdwell (Department of Electrical Engineering) and Dr. T.W.

Wang (Department of Chemical Engineering) [38]. The main features of the new search method are that it is based on a decision tree structure, and has a logarithmic scale up in speed, and supports parallelization.

The search algorithm will result in a list of profiles matching exactly or close to the target profile depending upon the match stringency criteria specified. The final list of probable candidates resulting after the search is arranged in an order according to their place of entry in the database and thus is difficult for caseworkers to identify the top closest candidates to follow up, as time plays an important role in solving crimes. The match returned by a search maybe on the order of 100's/1000's and it becomes difficult for case workers/analysts to look into all the returned searches. Therefore, developing the capability of rank ordering the final list of candidate profile becomes a priority. Such a similarity measure will rank and identify the top few candidate profiles that match or are closest to the target profile, and lessen the work of caseworkers by short listing the number of probable suspects to the most probable suspects. The similarity index should be designed to reflect the rationale used by the caseworkers if they were to rank order the profiles manually. In reviewing relevant literature reference, such a similarity parameter does not appear to exist.

A potential application of such a similarity measure down the road is in using the search engines where the match stringency requirements can be removed and the entire search based on the similarity measure of all profiles to that of target. The

objective is to organize the storage of DNA profile information to minimize the time required to locate all DNA profiles within the database that satisfy a set of user selected criteria when compared against a target profile.

1.2. PROBLEM STATEMENT

The purpose of this research work is to develop a clustering method to partition the DNA profiles stored in the national CODIS database. A second objective is to develop a Similarity Index (SI) parameter to rank order candidate profiles, resulting from a search of a large DNA database, and to identify the sensitivity of the subsequent rankings to changes in the relative weights for the subparameters used in the development of the SI.

CHAPTER 2

BACKGROUND

2.1. CODIS: THE NATIONAL DNA DATABASE

Two separate DNA databases have been developed and activated online by the FBI; one containing DNA profiles of convicted criminals (The Convicted Offender Index) and the other comprising of profiles obtained from crime scene evidences (The Forensic Index). These two indices can be accessed and profiles compared and searched by local, state and federal law enforcement entities using search engines. This highly sophisticated index system is known as the COmbined DNA Index System (CODIS) [41, 43]. This powerful system can be highly effective in providing rapid solution possibilities to criminal cases. The larger and the more complete the database, the easier it is to search for matching profiles. Of course, it may also take a longer time.

The CODIS database contains relevant DNA information and the actual STR/DNA loci characteristics. The profiles from different crime scenes can be shared easily. In addition, CODIS also contains a database of anonymous profiles (Population Index) to test the statistical significance of the match.

CODIS is divided into three levels; local, state and the national level, as shown in Figure 2.1[43]. Each level may contain the Convicted Offender Index, the

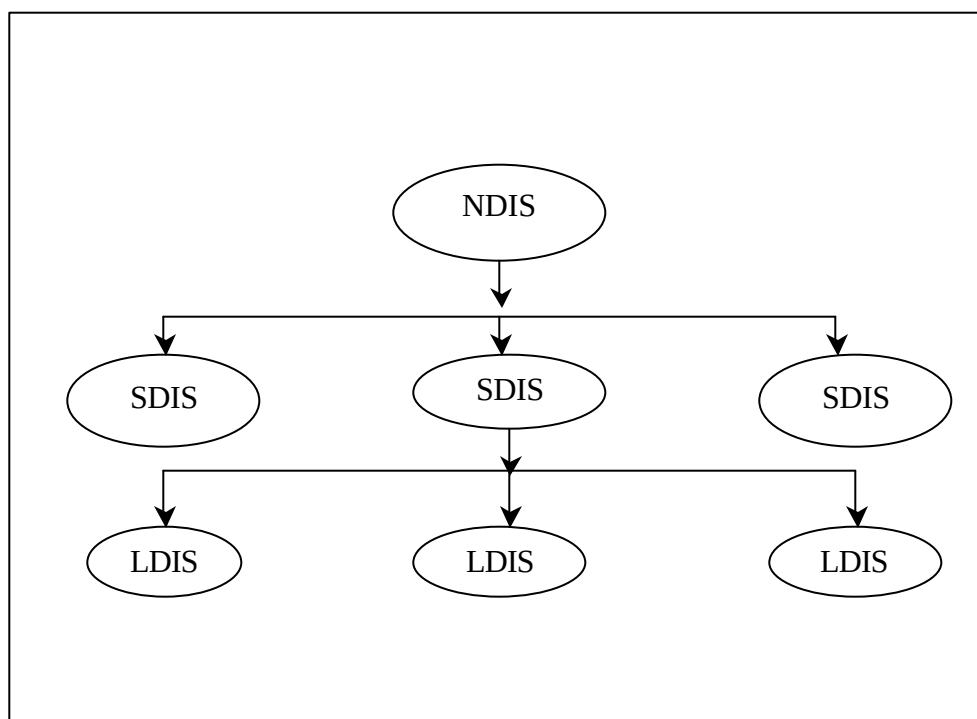


FIGURE 2.1: CODIS Architecture. NDIS: National DNA Index System, SDIS: State DNA Index System, LDIS: Local DNA Index System.

Forensic Index, as well as several Population indices. The last is to provide population statistical information to assess the rarity of a DNA profile. Each level has control over its own data. The police departments, the sheriff's offices and other state police agencies, use the Local DNA Index System (LDIS). This database is shared with other LDIS by forwarding it to the state level. At the state level, a single State DNA Index System (SDIS) exists where the forwarded local Forensic Indices obtained from all local areas are compiled and then compared against each other. These local forensic profiles are also compared against the Convicted Offender Index. This state level data is then forwarded to the national level.

The National DNA Index System (NDIS) is operated and maintained by the FBI. The DNA Identification Act of 1994 formalized the FBI's authority to establish a national DNA index system for law enforcement purposes.

2.2. BASIC GENETIC PRINCIPLES

Each human body consists of cells that originate from a successive division from a single fertilized egg. Deoxyribonucleic acid (DNA) is present in the nucleus of each cell [42]. There are two types of DNA, nuclear and mitochondrial as shown in Figure 2.2. The DNA is organized into chromosomes. The process of cell division and chromosome replication is such that each cell in the human body has the same chromosomal makeup as any other cell. Blood and semen are the usual body tissues from which the nuclear DNA can be easily isolated and typed to establish lineage or

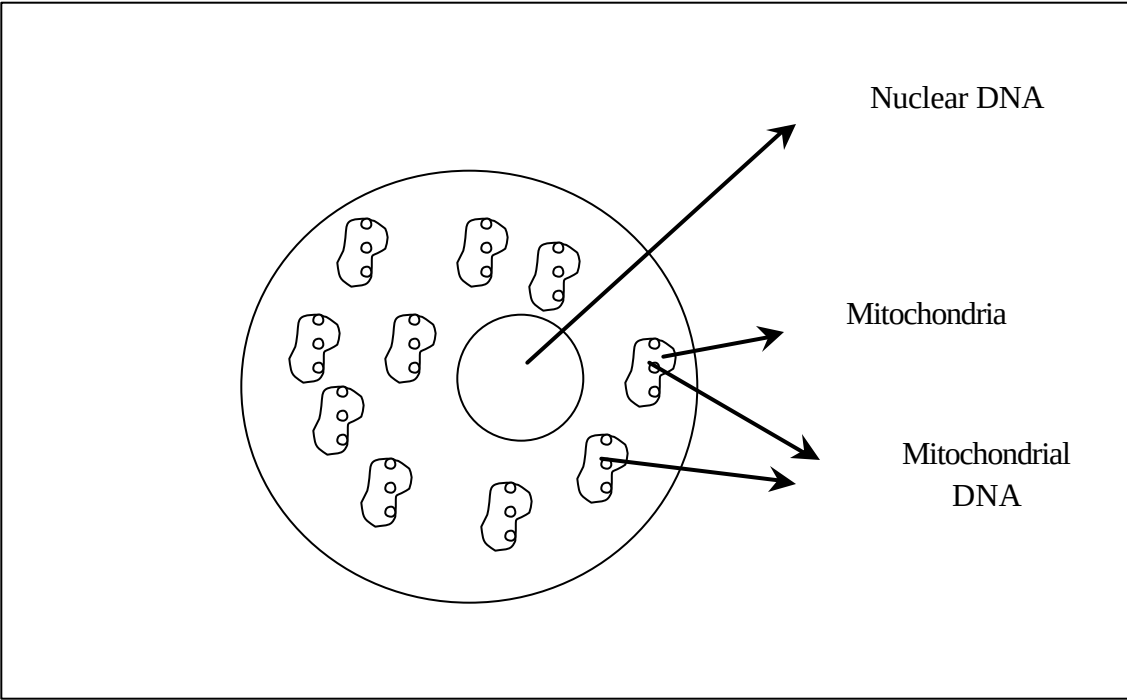


FIGURE 2.2: Location of the Two Types of DNA [42].

for identification purposes.

A chromosome is a very long and thin thread of DNA, surrounded mainly by protein. The DNA in the chromosome has double helical structure as shown in Figure 2.3. Each strand consists of a string of nucleotide bases strung together by sugar-phosphate bonds and forms the backbone of the chain. The phosphate bond is between the 3' carbon position of one ribose base and the 5' carbon of the adjacent ribose base. In one direction of the DNA strand, the direction is referred to as 3' to 5', whereas the other direction is referred to as 5' to 3'. The bases are one of four kinds: A (adenosine), T (thymine), G (guanine), and C (cytosine). In double stranded DNA, the bases line up in pairs, an 'A' from one strand is bonded to a 'T' in the other strand and a 'G' is bonded to a 'C', as seen in Figure 2.4. Thus, if a sequence in one strand is known, the other is determined. A gene is a sequence of DNA consisting of base pairs, and the position that a gene occupies along the DNA is defined as its *locus*. Each gene is unique from the other gene because it consists of a specific sequence of bases. Alternative forms of gene sequence at a site are called alleles for that locus. There are two copies of alleles present in a locus: one inherited from the father, and the other from the mother. If the same kind of allele is present at a specific locus, then the DNA profile is referred to as homozygous for that locus, and the DNA profile is heterozygous if it has two different alleles present at that locus.

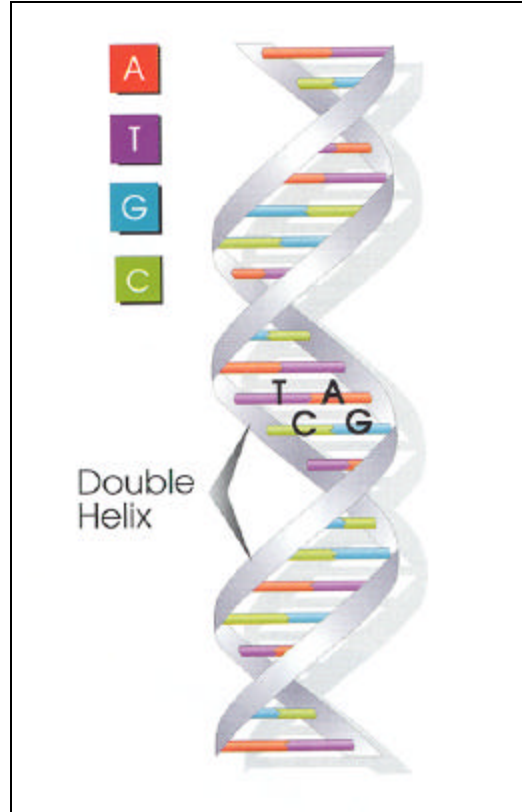


FIGURE 2.3: DNA HELIX [16].

CATTAGACTGAT
GTAATCTGACTA

FIGURE 2.4: DNA Bases Line Up in Pairs; An Example of a Short Length of Double Stranded DNA.

2.3. DNA TYPING FOR HUMAN IDENTIFICATION PURPOSES

DNA typing for forensic purposes is based on the fundamental principles of diagnosis and gene mapping with the application of efficient statistical tools. These methods analyze a portion of the DNA profile in a tissue sample and extract the genetic makeup of that region of a person to render identity. An important property of DNA is that it has strong resistance to the surrounding conditions, and so is, more stable than other biological compounds. For DNA typing, only a very small amount of DNA is required if the PCR (Polymerase Chain Reaction) amplification method is employed. Thus, an amount enough for subsequent DNA analysis can be obtained from just a tiny amount collected from crime scenes. It has been shown [1, 16] that saliva from cigarette butt would yield enough DNA samples that can be amplified by PCR to yield enough DNA for subsequent identification. Direct DNA typing/analysis gives more direct and unique results, in comparison to identification by other methods that use blood groups and enzymes. With the advent of new DNA typing methods many pending and upcoming forensic trial cases can be unambiguously resolved. DNA technology is very useful in resolving cases involving falsely accused innocent suspects. According to the FBI, about one-third of individuals named as primary suspect in rape cases were later on excluded by DNA evidence, thus signifying its importance in forensics. Figure 2.5 shows the flowchart of steps involved in DNA typing. The following section provides some necessary genetic background and a minimum vocabulary to assist in further understanding of the basic fundamentals.

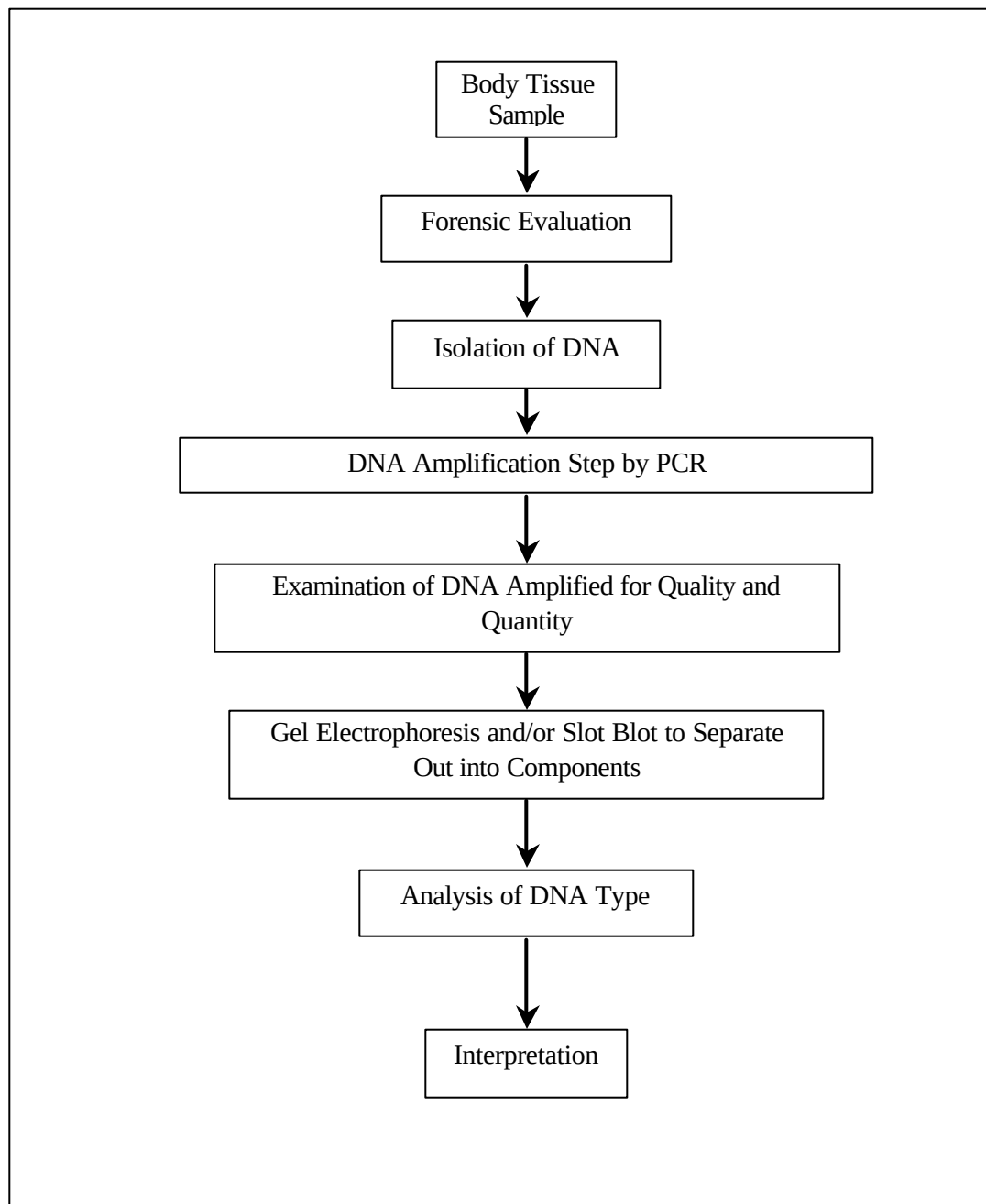


FIGURE 2.5: Flowchart for Forensic DNA Typing [16].

2.4. APPLICATIONS OF DNA TYPING AND PROFILING

2.4.1 PCR AMPLIFICATIONS

PCR (Polymerase Chain Reaction) is a general technique for DNA amplification that increases and duplicates the amount of DNA in a sample. Usually the DNA samples that are available are either limited or degraded. By this technique, at each round of amplification, multiple copies of each existing copy are made. The copies have all the properties of the original sample, and so these copies can be used for subsequent analysis. PCR is done in three stages – sample treatment, PCR process, and product analysis. Figure 2.6 shows the geometric increase in DNA copies during the PCR step.

The PCR technique is also time saving and economical when compared to other techniques commonly used in forensic DNA applications. Because of its advantages, almost all DNA labs are converting to PCR as the preferred method of doing DNA typing. The DNA product obtained using the PCR method can be analyzed in a variety of different ways. Commercial kits are available that analyze for several DNA loci information using only one sample preparation [1, 16, 57].

In cases of sexual assault the evidence sample often contains a mixture of DNA from the victim and the suspect(s). A differential extraction procedure can be applied to separate the DNA female cells from that of the sperm cells. Then only the sperm cells are analyzed for its DNA typing by the PCR procedure.

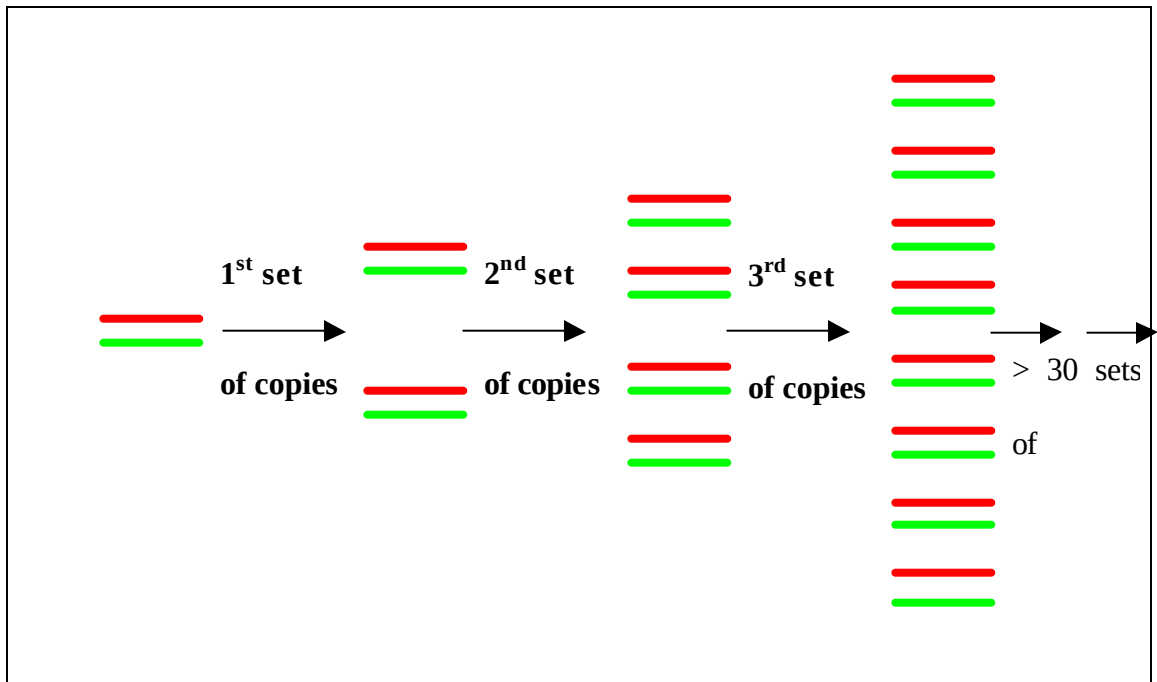


FIGURE 2.6: Overview of the Steps of Polymerase Chain Reaction (PCR) [42].

DNA extracted from blood from various substrates consisting of clothing, wood, paper, human skin, rubber, leather, and steel has been shown to amplify reliably using different techniques to detect the amplified DNA sample [15,16]. Further, DNA from cigarette butts can be a potential source of evidence [15,16].

PCR can also be used to identify unknown human remains [15,16]. Due to the effects of temperature, moisture, and autolytic process within the body, the identification of human remains by conventional forensic means is not always possible. A potential source of DNA from human remains is compact bones, but soft tissue samples can also be used if available. In the identification of human remains, reference blood samples from relatives are essential for interpretation of the results.

2.4.2 SHORT TANDEM REPEATS (STRs)

Short Tandem Repeat (STR) is a recent DNA fragmentation system that has strengthened the use of DNA in forensic applications, because it can be amplified into a great amount of identical DNA from just a tiny amount of forensic sample DNA. It is short length in nature (usually just around 100 base pairs long). However, because of its relatively small size, these fragments have to be first amplified into millions of identical copies in order to yield enough material that can be analyzed for its makeup. The STR sequence is consisted of multiple repeat of some particular 4base sequence. The repeat number can be from 2 units to tens of units. STR loci are detected as discrete alleles. Each allele has a certain probability of occurrence among a

population, and is known for several world populations. The allele distribution pattern of occurrence can therefore be compared directly to a standard allele ladder ruler made up of all possible alleles (allelic ladder) after separation by gel electrophoresis.

DNA typing refers to the identification of the type of alleles present at several STR loci from a DNA sample preparation. At present, CODIS requires a sample to have allele information at the thirteen loci before permitting a search. Because of the varying probability of allele occurrences in the population, differentiation of individuals can be achieved by examining the joint allele distribution pattern across all thirteen loci. This is the basis, which makes it possible to uniquely identify an individual. The larger the population, the more loci are required to allow for differentiation of individuals. In England a total of six STR loci are required for DNA typing. This low number of loci requirement for identification has the potential of having more than one individual to exhibit identical DNA pattern at these STR loci. Indeed, in year 2000, a case was reported in England, where a search produced a DNA profile that matches that of a forensic sample, when the owner of the profile has been shown unambiguously that he was not where the scene of the crime at the time of the crime was committed.

The loci chosen for forensic use to identify individuals generally have a tandem repeat unit of 4 base pair (bp) and may be repeated from a few to dozens of times among individuals. But the number of repeats at a locus is fixed in any particular individual,

but maybe different between individuals. The number of repeats is termed as the 'allele'. The number of different alleles at each locus present in a population varies from about 5 to 20, depending on the type of the locus. Each allele name corresponds to the number of the tandem repeat units present at the locus. For instance, allele 11 designates the presence of eleven 4base pair repeat units, whereas an allele name with a decimal number, such as 10.3 represents the presence of ten 4base repeat units plus a 3-base partial unit. The STR fragments comprise of just a few hundred base pairs.

A STR locus from several of the 23 chromosomes in humans has been identified to be suitable for use in human identification because of the different number of variant alleles exhibited by each of those loci. The frequency distribution of the alleles at these loci for several populations in the world have also been derived by scientists and are well known and accepted by the scientific community. These populations include the four major ethnic groups of Caucasian, African American, Hispanics and Asians. Every year, STR information on a variety of ethnic groups are obtained and reported in the literature.

2.4.3. PROBABILITY DISTRIBUTION OF ALLELES IN THE SIXTEEN CODIS CORE LOCI

This section gives the allele probability distribution of the sixteen STR loci provided by the CODIS [41] for the Caucasian population, thirteen of which are presently used in conducting searches on CODIS. The thirteen loci used are csf1po, fga, th01, tpox,

vWa, d3s1358, d5s818, d7s820, d8s1179, d13s317, d16s539, d18s51, and d21s11. The number after the letter 'd' in the locus name refers to the chromosome number the locus resides in. The probability densities for alleles at these loci were provided for the Caucasians, African Americans, Hispanics and some other populations. The probability distribution at the sixteen loci is shown in Tables 2.1 - 2.15 and Figures 2.7 - 2.21.

2.5. DNA AND DATABASE

Most of us are aware of the usefulness and effectiveness of the fingerprints database. The Automated Fingerprint Identification System (AFIS) contains millions of people's fingerprints in computer files and is useful in solving thousands of criminal cases every year, worldwide. The DNA identifying technique is an even more powerful tool in solving multiple crimes and countries around the world have been busy in establishing 'DNA fingerprint' databases for forensic applications.

All 50 states have already passed legislation making the collection and analysis of samples for DNA databanks compulsory [39, 40]. DNA samples are generally collected from convicted criminals and then forwarded to the state DNA laboratory for typing and storage. The entire DNA/STR information can be represented by a set of numbers, for example, in binary format. The numerical representation of DNA/STR profiles facilitates the storage and search of DNA profiles by computer. The effectiveness of such databanks can be further increased by proper standardization of the systems used to create these profiles.

TABLE 2.1: Relative Frequency of Occurrence of Alleles at the csf1po Locus for Caucasians.

Bin	Allele	Fraction
1	7	0.000
2	8	0.003
3	9	0.028
4	10	0.268
5	10.3	0.003
6	11	0.258
7	12	0.358
8	13	0.070
9	14	0.015
10	15	0.000
Total Probability		1.000

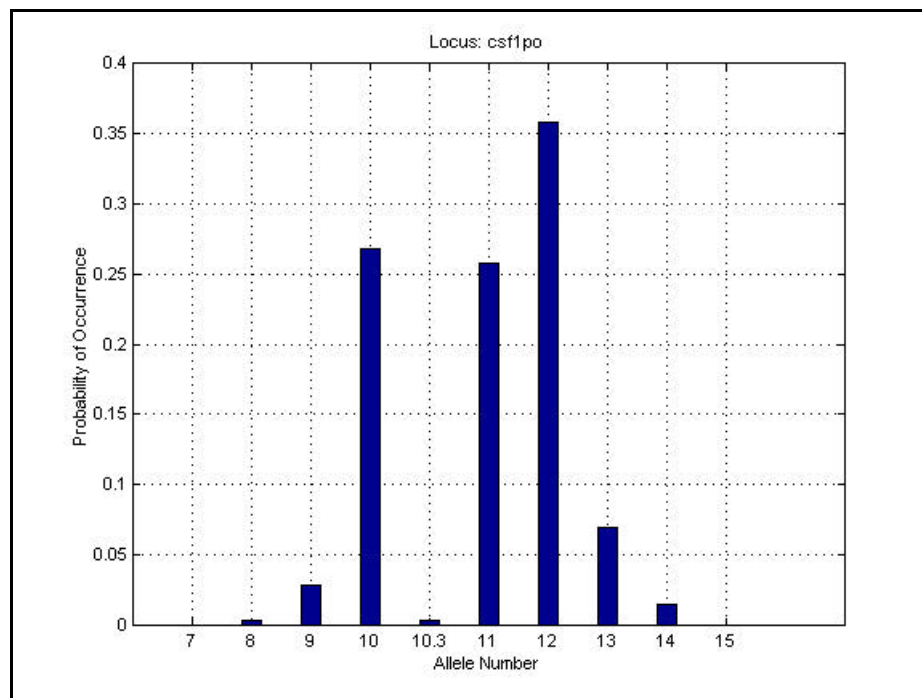


FIGURE 2.7: Allele Frequency Distribution of the csf1po Locus of Caucasian DNA.

TABLE 2.2: Relative Frequency of Occurrence of Alleles at the d18s51 Locus for Caucasians.

Bin	Allele	Fraction
1	8	0.000
2	9.2	0.001
3	10	0.008
4	11	0.012
5	12	0.139
6	13	0.125
7	14	0.164
8	14.2	0.000
9	15	0.145
10	16	0.137
11	17	0.115
12	18	0.080
13	19	0.041
14	19.2	0.000
15	20	0.017
16	21	0.010
17	22	0.005
18	23	0.001
19	24	0.002
Total Probability		1.000

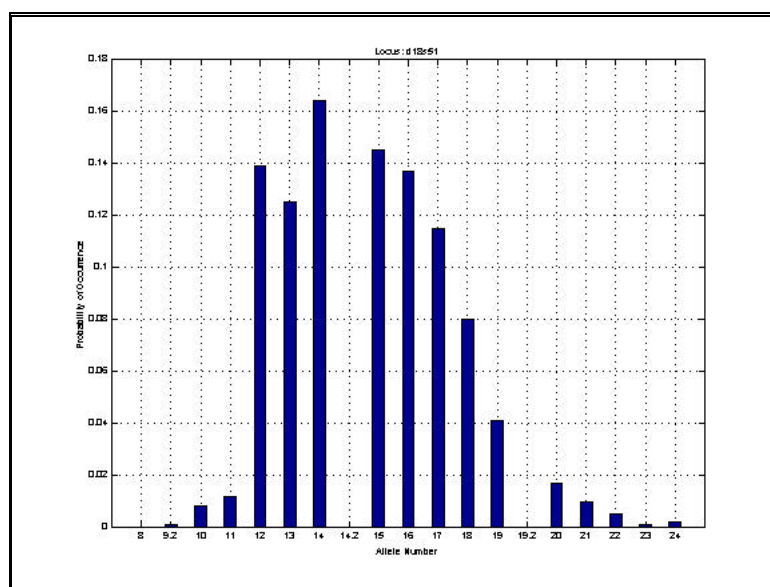


FIGURE 2.8: Allele Frequency Distribution of the d18s51 Locus of Caucasian DNA.

TABLE 2.3: Relative Frequency of Occurrence of Alleles at the d13s317 Locus for Caucasians.

Bin	Allele	Fraction
1	7	0.000
2	8	0.115
3	9	0.078
4	10	0.070
5	11	0.313
6	12	0.283
7	13	0.098
8	14	0.043
9	15	0.003
Total Probability		1.000

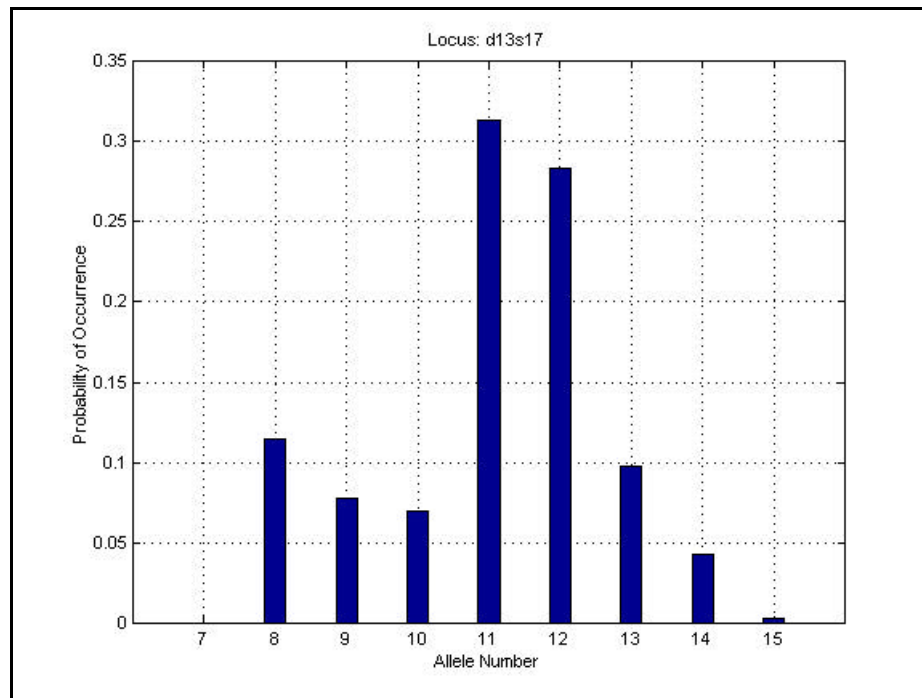


FIGURE 2.9: Allele Frequency Distribution of the d13s317 Locus of Caucasian DNA.

TABLE 2.4: Relative Frequency of Occurrence of Alleles at the d16s539 Locus for Caucasians.

Bin	Allele	Fraction
1	5	0.000
2	6	0.000
3	7	0.000
4	8	0.026
5	9	0.107
6	10	0.079
7	11	0.319
8	12	0.269
9	13	0.167
10	14	0.031
11	15	0.002
Total Probability		1.000

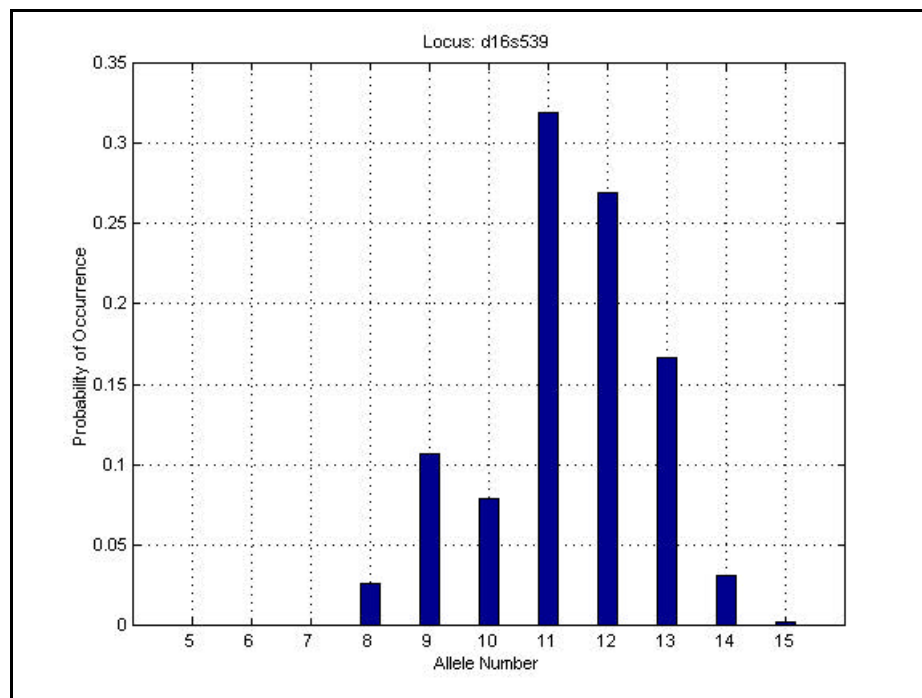


FIGURE 2.10: Allele Frequency Distribution of the d16s539 Locus of Caucasian DNA.

TABLE 2.5: Relative Frequency of Occurrence of Alleles at the dqalpha Locus for Caucasians.

Bin	Allele	Fraction
1	1.1	0.122
2	1.2	0.178
3	1.3	0.064
4	2	0.133
5	3	0.183
6	4	0.320
Total Probability		1.000

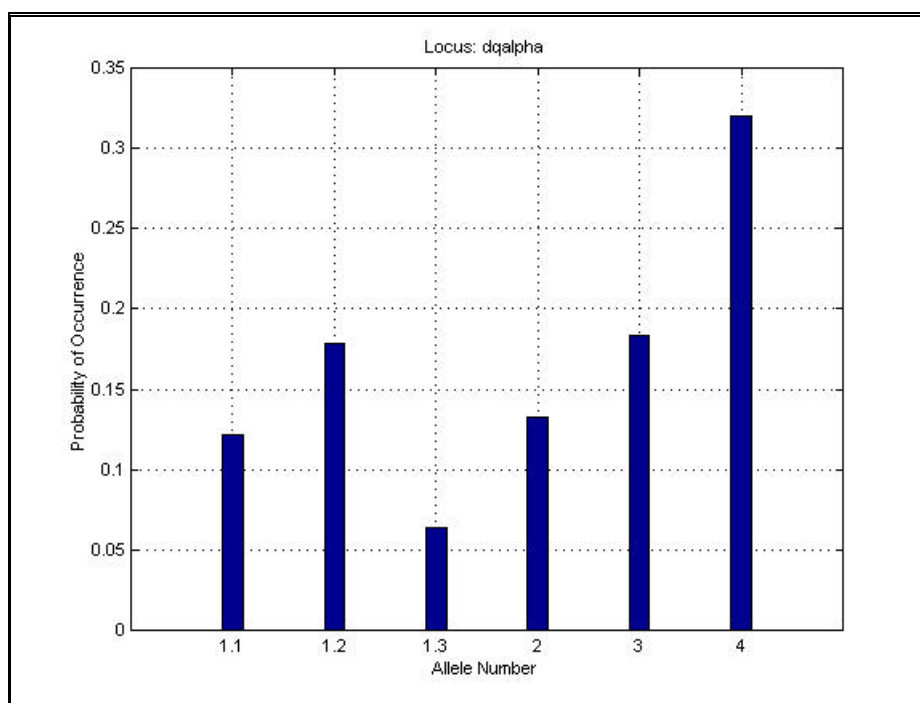


FIGURE 2.11: Allele Frequency Distribution of the dqalpha Locus of Caucasian DNA.

TABLE 2.6: Relative Frequency of Occurrence of Alleles at the d3s1358 Locus for Caucasians.

Bin	Allele	Fraction
1	9	0.000
2	11	0.003
3	12	0.003
4	13	0.005
5	14	0.113
6	15	0.283
7	15.2	0.000
8	16	0.223
9	17	0.223
10	18	0.145
11	19	0.005
Total Probability		1.000

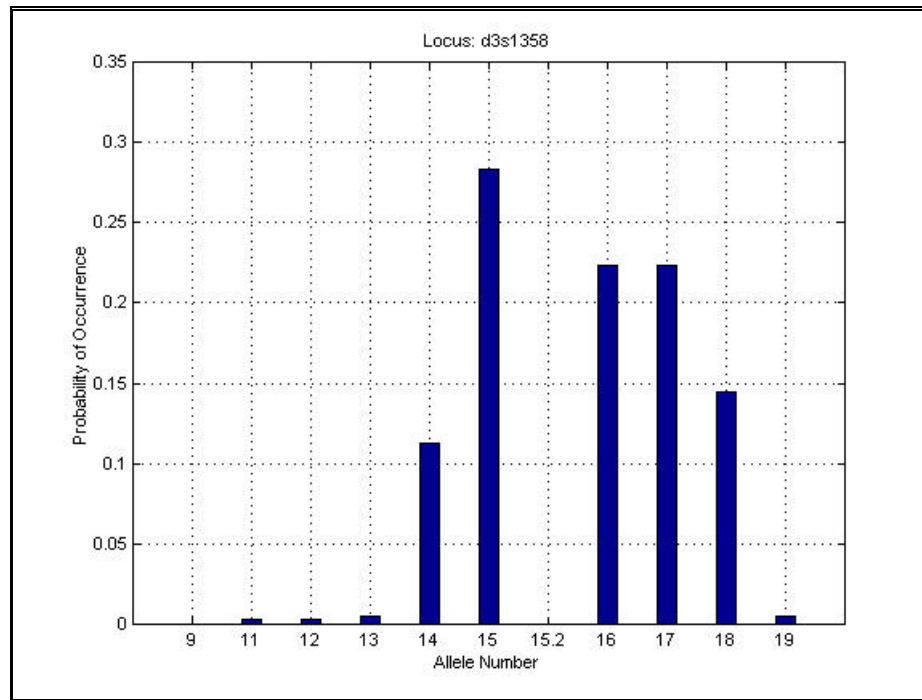


FIGURE 2.12: Allele Frequency Distribution of the d3s1358 Locus of Caucasian DNA.

TABLE 2.7: Relative Frequency of Occurrence of Alleles at the d5s818 Locus for Caucasians.

Bin	Allele	Fraction
1	7	0.003
2	8	0.005
3	9	0.023
4	10	0.068
5	11	0.393
6	12	0.333
7	13	0.165
8	14	0.010
9	15	0.000
10	16	0.003
Total Probability		1.000

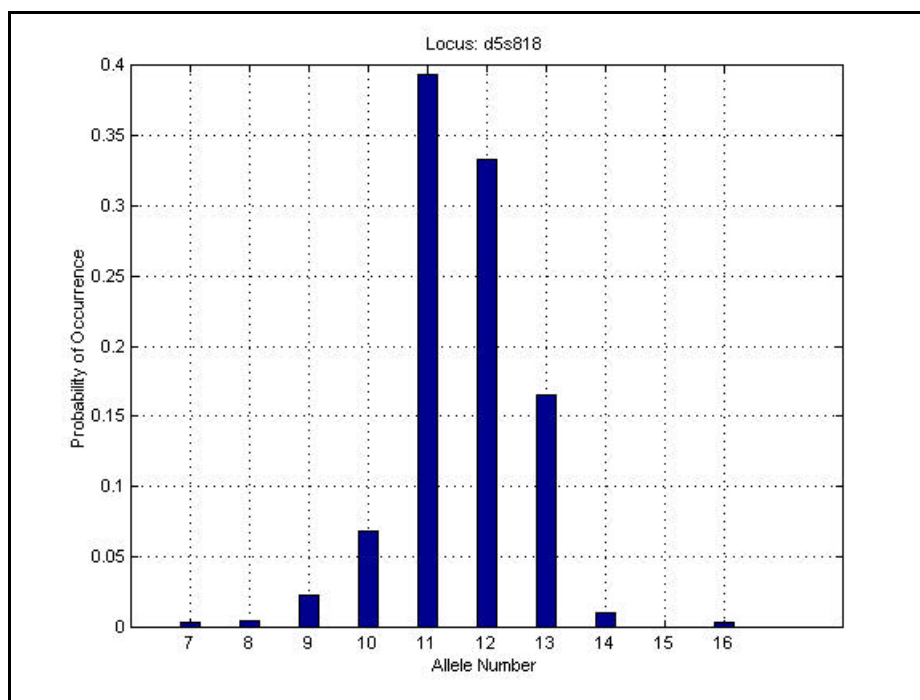


FIGURE 2.13: Allele Frequency Distribution of the d5s818 Locus of Caucasian DNA.

TABLE 2.8: Relative Frequency of Occurrence of Alleles at the d7s820 Locus for Caucasians.

Bin	Allele	Fraction
1	6	0.000
2	6.3	0.003
3	7	0.025
4	8	0.175
5	9	0.130
6	10	0.240
7	11	0.230
8	12	0.160
9	13	0.028
10	14	0.008
11	15	0.003
Total Probability		1.000

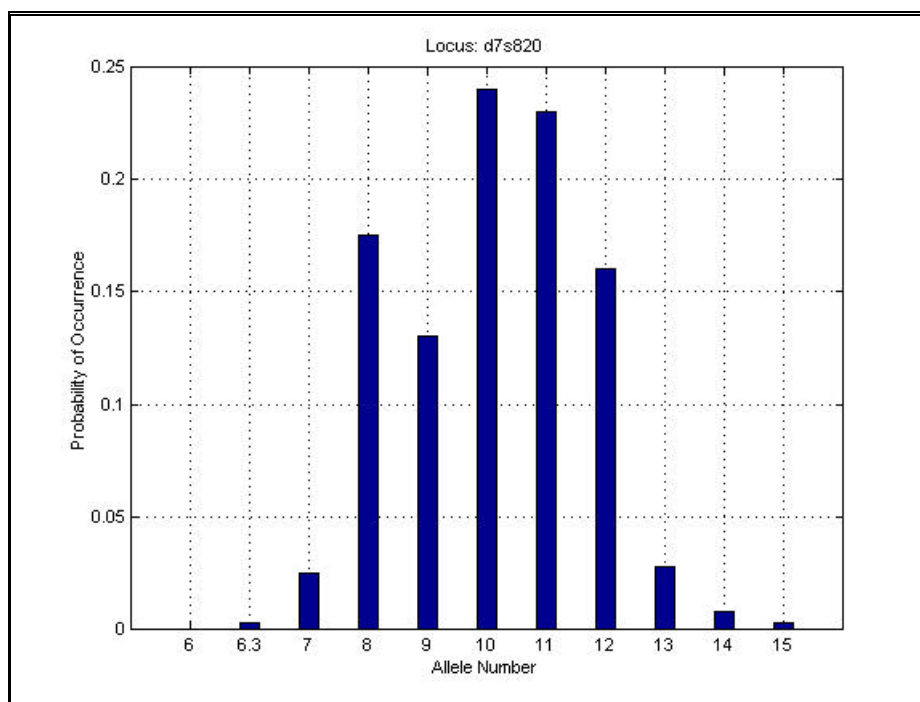


FIGURE 2.14: Allele Frequency Distribution of the d7s820 Locus of Caucasian DNA.

TABLE 2.9: Relative Frequency of Occurrence of Alleles at the d8s1179 Locus for Caucasians.

Bin	Allele	Fraction
1	8	0.018
2	9	0.013
3	10	0.094
4	11	0.066
5	12	0.143
6	13	0.333
7	14	0.209
8	15	0.088
9	16	0.031
10	17	0.004
11	18	0.000
Total Probability		1.000

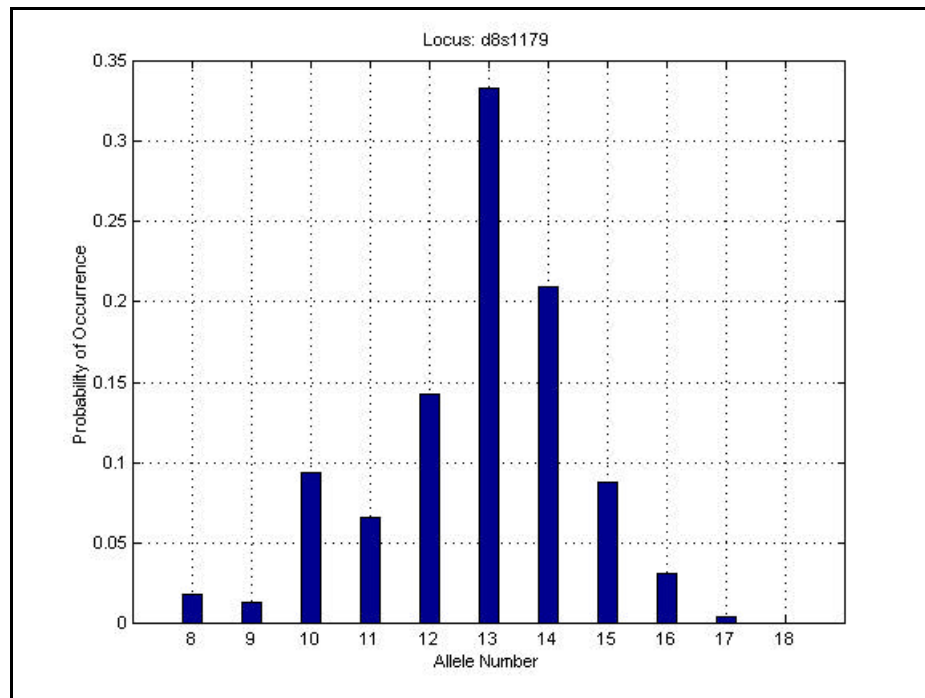


FIGURE 2.15: Allele Frequency Distribution of the d8s1179 Locus of Caucasian DNA.

TABLE 2.10: Relative Frequency of Occurrence of Alleles at the th01 Locus for Caucasians.

Bin	Allele	Fraction
1	5	0.000
2	6	0.253
3	7	0.160
4	8	0.093
5	9	0.138
6	9.3	0.350
7	10	0.008
8	11	0.000
Total Probability		1.000

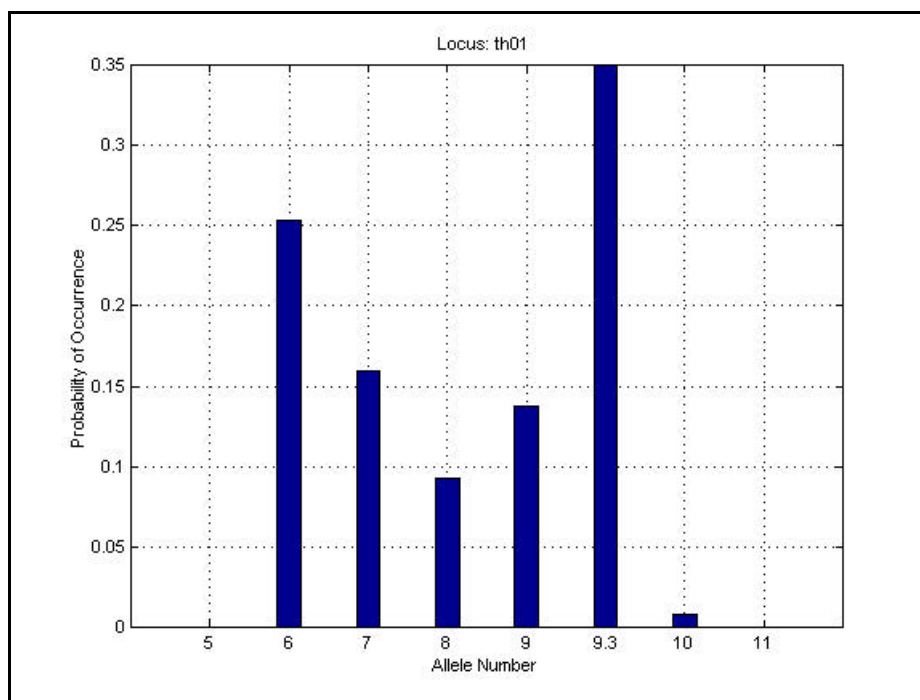


FIGURE 2.16: Allele Frequency Distribution of the th01 Locus of Caucasian DNA.

TABLE 2.11: Relative Frequency of Occurrence of Alleles at the tpox Locus for Caucasians.

Bin	Allele	Fraction
1	6	0.000
2	7	0.000
3	8	0.583
4	9	0.113
5	10	0.048
6	11	0.225
7	12	0.030
8	13	0.003
Total Probability		1.000

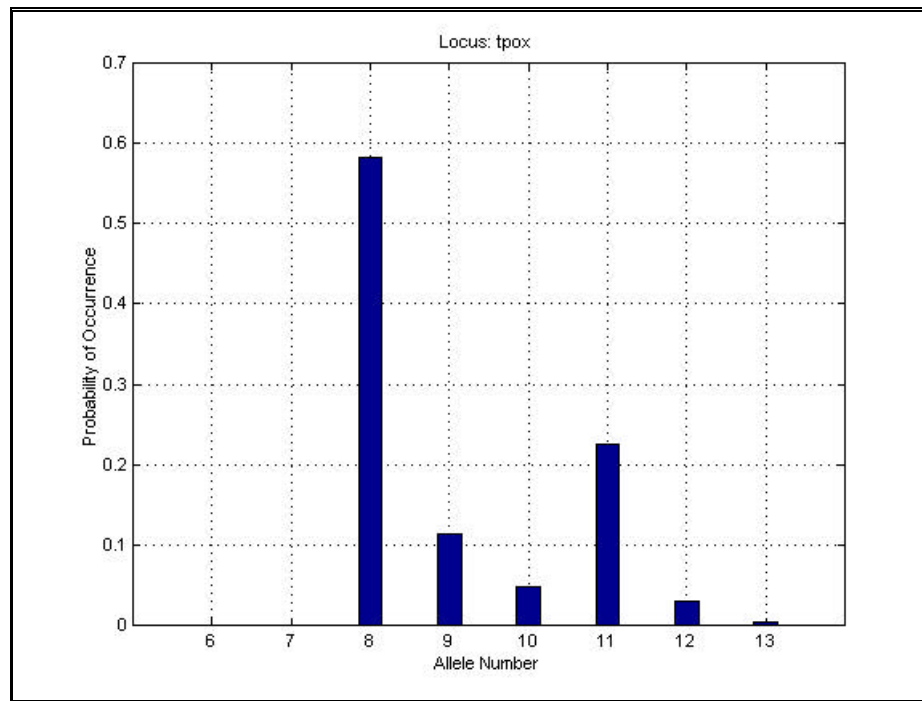


FIGURE 2.17: Allele Frequency Distribution of the tpox Locus of Caucasian DNA.

TABLE 2.12: Relative Frequency of Occurrence of Alleles at the vWa Locus for Caucasians.

Bin	Allele	Fraction
1	11	0.000
2	12	0.000
3	13	0.000
4	14	0.085
5	15	0.083
6	16	0.198
7	17	0.250
8	18	0.258
9	19	0.110
10	20	0.015
11	21	0.003
Total Probability		1.000

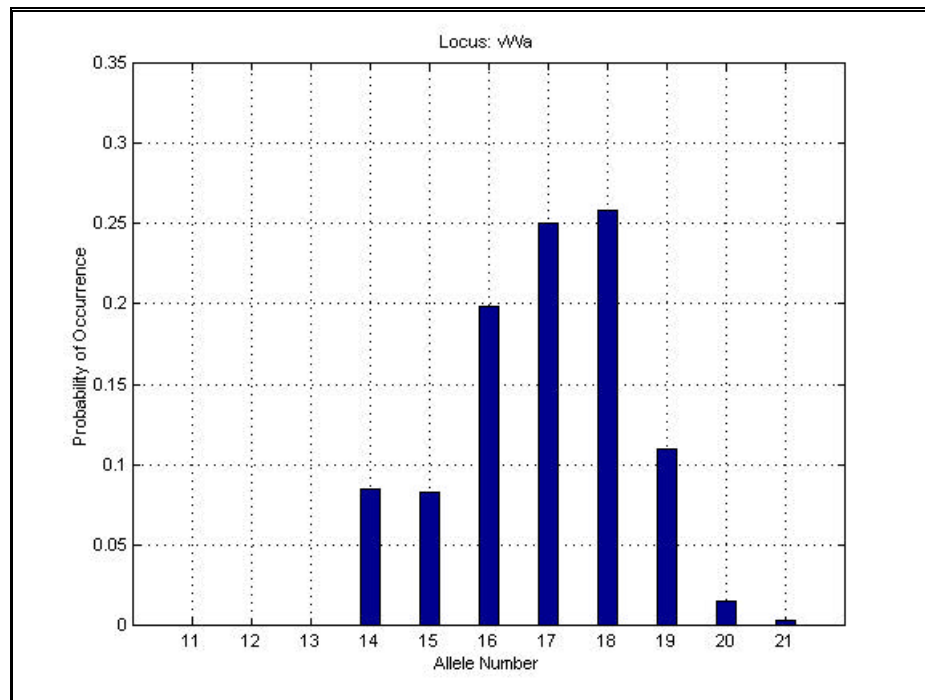


FIGURE 2.18: Allele Frequency Distribution of the vWa Locus of Caucasian DNA.

TABLE 2.13: Relative Frequency of Occurrence of Alleles at the fga Locus for Caucasians.

Bin	Allele	Fraction
1	16	0.000
2	17	0.000
3	17.2	0.000
4	18	0.015
5	18.2	0.000
6	19	0.063
7	20	0.170
8	20.2	0.000
9	21	0.178
10	22	0.170
11	22.2	0.000
12	23	0.140
13	24	0.133
14	25	0.113
15	26	0.015
16	27	0.005
17	28	0.000
18	29	0.000
19	30	0.000
20	31	0.000
21	>31	0.000
Total Probability		1.000

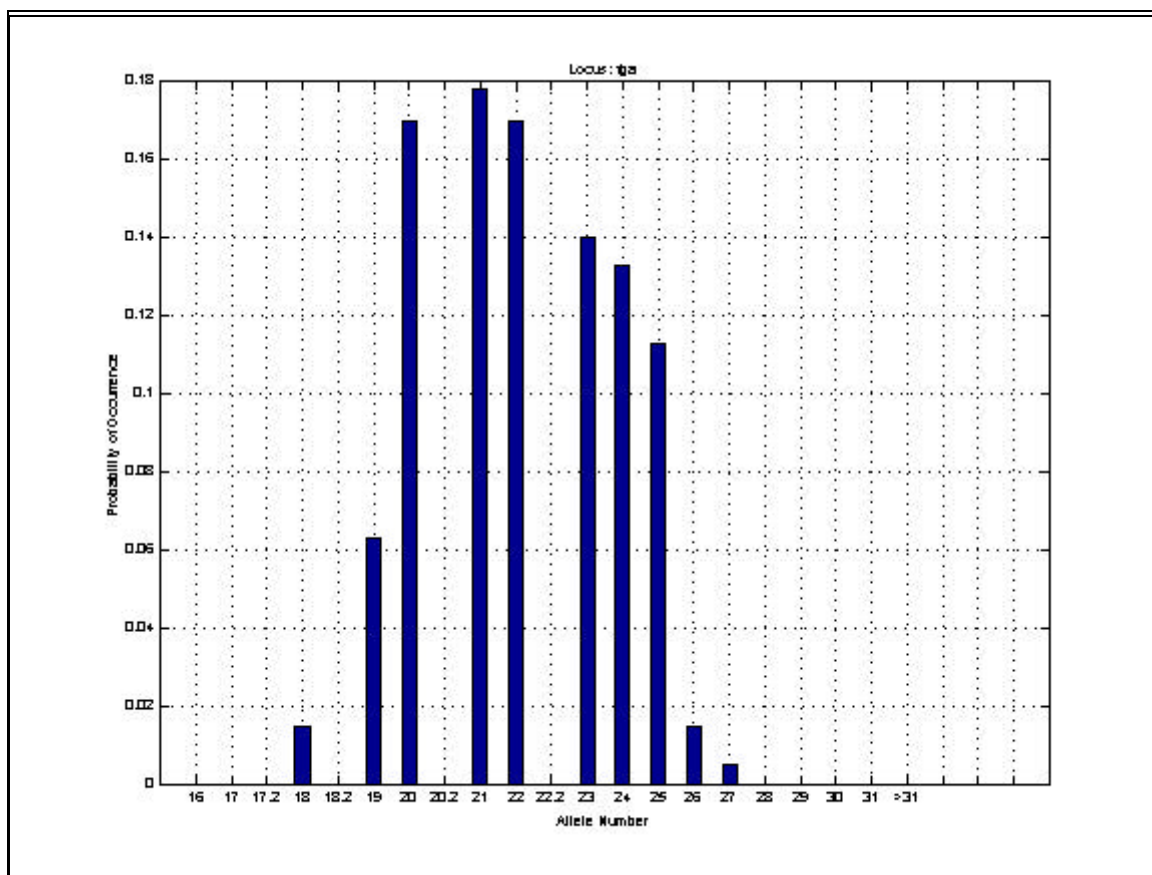


FIGURE 2.19: Allele Frequency Distribution of fga Locus of Caucasian DNA.

TABLE 2.14: Relative Frequency of Occurrence of Alleles at the d1s80 Locus for Caucasians.

Bin	Allele	Fraction
1	15	0.000
2	16	0.001
3	17	0.002
4	18	0.237
5	19	0.003
6	20	0.018
7	21	0.021
8	22	0.038
9	23	0.012
10	24	0.378
11	25	0.046
12	26	0.020
13	27	0.007
14	28	0.068
15	29	0.052
16	30	0.008
17	31	0.072
18	32	0.006
19	33	0.003
20	34	0.001
21	35	0.003
22	36	0.004
23	37	0.001
24	38	0.000
25	39	0.003
26	40	0.000
27	41	0.000
28	>41	0.001
Total Probability		1.000

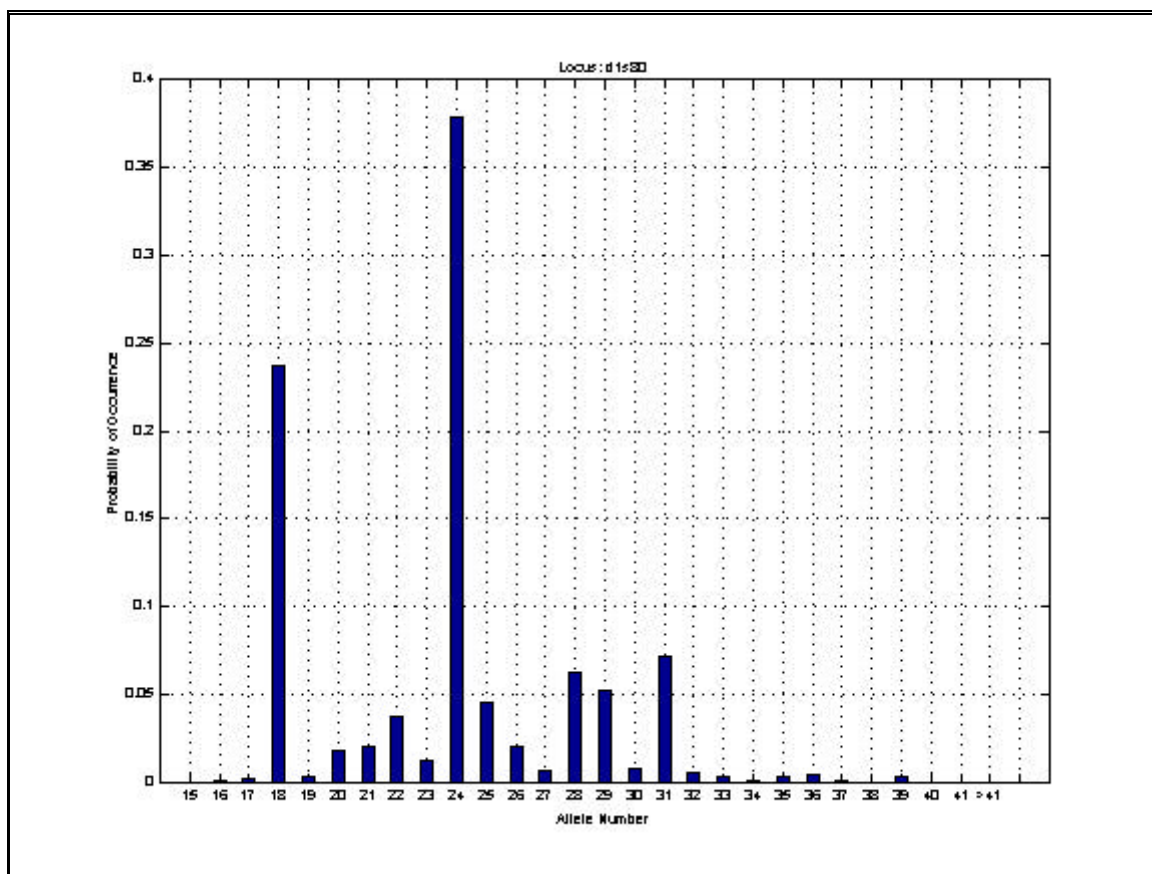


FIGURE 2.20: Allele Frequency Distribution of the d1s80 Locus of Caucasian DNA.

TABLE 2.15: Relative Frequency of Occurrence of Alleles at the d21s11 Locus for Caucasians.

Bin	Allele	Fraction
1	53	0.000
2	54	0.001
3	57	0.001
4	59	0.031
5	61	0.160
6	63	0.226
7	64.1	0.000
8	64	0.000
9	65	0.258
10	66	0.027
11	67	0.069
12	68	0.093
13	69	0.018
14	70	0.090
15	71	0.001
16	72	0.022
17	73	0.000
18	74	0.022
19	75	0.000
20	77	0.000
Total Probability		1.000

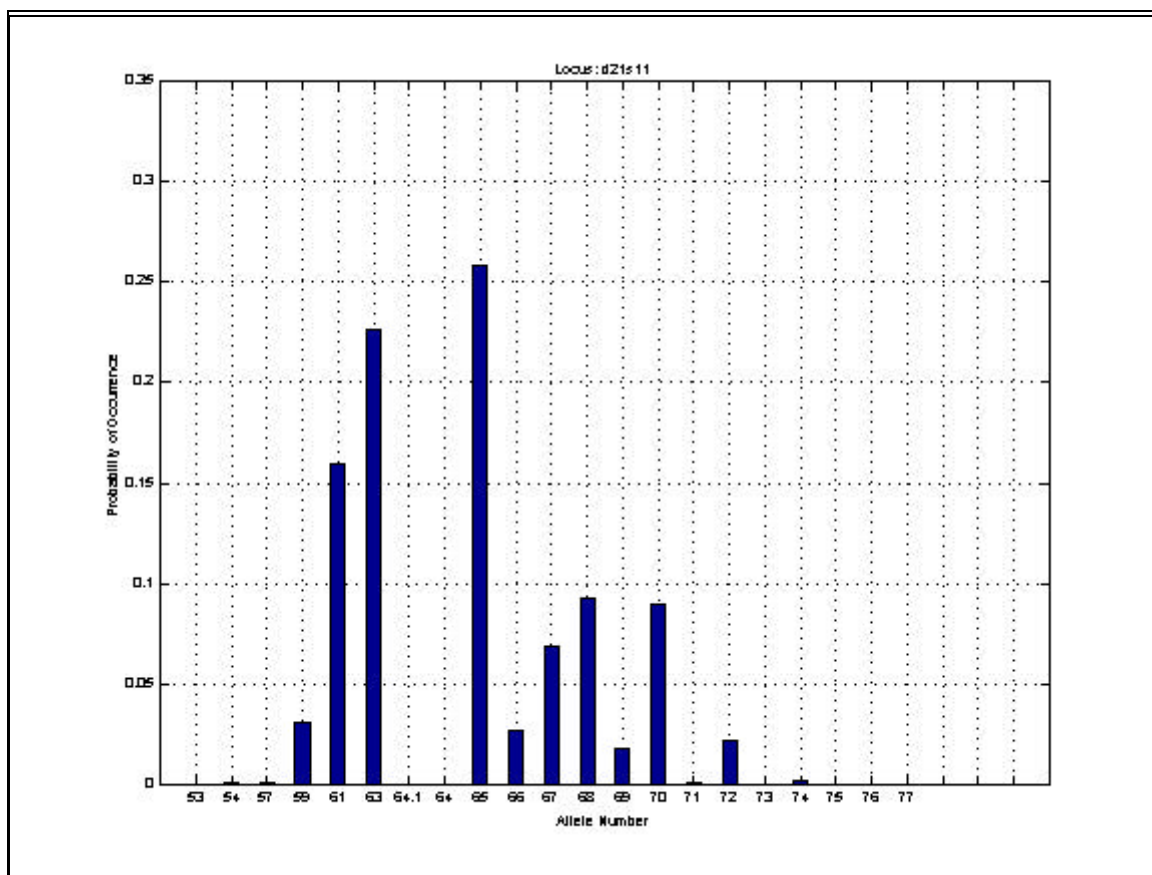


FIGURE 2.21: Allele Frequency Distribution of the d21s11 Locus of Caucasian DNA.

Currently, the FBI is leading in the effort to create a national DNA database. The national database will be connected to all the states that would contribute, as well as use, the information across databases. At present, mostly all states are linked in a network called NDIS (National DNA Index System) to share the DNA information (Figure 2.1).

Databanks may also be created from non-suspect individuals that would be useful in identifying unknown bodies, for instance, in a mass disaster situations, such as airplane crash, earthquake and bombings. The Armed Forces Identification Laboratory has already setup a program to collect DNA samples from military personal to help in the identification of victims of war.

2.6. DNA IN COURTS: LEGAL ISSUES

DNA identification is the most reliable technique to compare and check if the samples collected from different sites have the same possible source. If the recurrence of the DNA pattern is unique enough to be linked to one individual, then the validity in using DNA for identification purposes is boosted. Currently, the criminal justice system usually accepts this DNA profiling technique to improve its criminal justice efficiency by virtue of the tremendous benefits offered by this technique. In some situations, the DNA evidence serves as the only evidence based on which a verdict is rendered [60, 61, 62].

In order to be effective and reliable, any forensic method of identification needs to be highly accurate in its ability to differentiate between various sources. Sound scientific basis and procedures must exist in order to store, process, compare and link the physical evidence to their sources.

In this regard, most of the courts demand a more convincing proof of individualization by DNA tests. Some courts go even further and demand a proof that the DNA profiles do indeed differ from person to person. In addition, they also ask experts in population statistics to provide quantitative estimates about the rarity of the identifying patterns in various population groups and subgroups.

DNA profiling is most useful in civil disputes involving paternity and immigration applications. The main advantage of DNA profiling is such that it is hoped that such legal matters could be resolved without invoking the need for formal hearings. Since the emergence of using DNA for personal identification, wrongly convicted suspects and some death-row inmates have been exonerated and released after testing of their DNA data and comparing them to those found at the crime scene. The Federal government has a project called 'The Innocence Project', precisely dedicated to the exoneration of convicted criminals.

CHAPTER 3

REVIEW OF THE LITERATURE

3.1. MULTIVARIATE STATISTICAL ANALYSIS USED IN CLUSTERING

In recent years a number of important developments have occurred in the statistical analysis of data. These developments have been in the areas of applied research in multivariate statistical analysis and computer technology. Multivariate statistical analysis is being widely used in diverse fields of applications such as education, psychology, public health, business, science, and lately in engineering.

Multivariate statistical analysis is concerned with analyzing data that consists of sets of related measurements on several non-independent process variables or characteristics for a given process or population. Common multivariate approaches are cluster analysis, factor analysis, and principal component analysis. The goal is to reduce the dimensionality of the data and through a change of basis, represent the reduced data in a low dimensional and mutually independent framework. Multivariate analysis is now being used for process monitoring and fault diagnosis [54, 55, 56] in the field of engineering. These approaches often depend upon calculating some characteristic parameters that are then used to group data into clusters.

The technique of Principal Component Analysis (PCA) will be used to analyze the allele/loci distribution information contained in the DNA database. Clusters will be

established that partition the entire database. Then, during search one of the clusters will be searched for possible matching profiles to the profile of the target of interest, thus reduces the time for search from that required if the entire database is searched through.

3.2. MEASURES OF GENETIC DISTANCE

In recent years, considerable attention has been devoted to measuring the genetic distance between populations as a whole, and several methods of distance measures have been proposed and used. The goal is to develop genetic linkage between population groups so as to trace their evolutionary pathways and to establish the estimated time of branching in the evolutionary process.

In 1968, Wright [1, 44] developed a distance measure to compute Euclidean or angular distance between two populations. Between 1966-1968, Sanghvi and Steinberg [1, 8, 44] developed various methods for applying the Euclidean distance to represent genetic differences between populations. Around the same time, Cavalli-Sforza [1, 9, 44] and his co-workers developed a genetic distance using angular transformations. In mid 1960's, Malecot [44] obtained a relationship connecting genetic similarity to geographic distance. Morton [44] tested the genetic index developed by Malecot, on various datasets. Edwards proposed another important genetic distance in early 1970's [1, 7, 9, 44].

A large number of measures have been proposed for studying the genetic similarity between populations. In 1972, Masatoshi Nei developed the most important and most widely used distance measure [1, 12, 28, 44]. This was the period when many researchers were working on the development of either Euclidean and/or angular genetic distance measures. Nei's similarity measure has been the most frequently used genetic distance because it identifies the biological unit to be measured.

The distance measures proposed by other researchers are modifications of Nei's distance measure, which is based on the frequency of the occurrence of alleles at each locus of the DNA profile. Other big names in the above field of genetic distance are Hillis, J.S. and Rogers [1].

A large number of studies in the late 1980's and 90's have been done to improve upon the earlier genetic distance measures developed in the 1960's and 70's. The new/modified methods developed by Lynch [4, 13, 22, 23, 24, 25] and Chakraborty [2, 10, 21, 35] separately are increasingly used to make inferences about the genetic variations within and between populations.

The purpose of most of the genetic distances developed to date is to measure the global evolutionary difference between populations and does not measure the distance between two individual DNA profiles. Thus they are not applicable in providing a similarity measure to rank order individual DNA profiles with respect to their

similarity to that of a target profile. As a result, during the course of our research, a *new* similarity index measure is developed that is to be used in ranking candidate profiles produced after a search is carried out.

CHAPTER 4

METHODOLOGY OF APPROACH

4.1. CLUSTERABILITY STUDY: OVERVIEW OF APPROACH

This section provides a description of an approach that uses multivariate statistical methods to study clusterability of DNA profiles. The application of this method to cluster profiles of Short Tandem Repeat (STR) DNA locus data is presented. The developed method is applied to DNA profiles containing sixteen CODIS loci with allele distribution for that of Caucasians [Section 2.4.3].

The DNA/STR profiles in the synthetic database are to be represented in binary form first, using a ‘1’ to denote the presence of, and a ‘0’, the absence of an allele at a locus. The binary data matrix is generated using the allele probability distribution specifications for each of the sixteen loci for that of the Caucasian. We then determine the combinations of various two-loci pairs of the sixteen CODIS loci that would give good cluster formation. Clusterability is studied by performing principal component analysis on the synthetic data profiles of 10,000 samples, followed by examination of the associated scores plots and other relevant information to determine factors that contribute to the good clusterability.

The clusters are validated by using the k-means clustering algorithm [40, 48]. The membership of each cluster is then identified and recorded followed by a study of the

allele distribution patterns that characterize each cluster. Each DNA/STR profile belongs to one and only one of these clusters. Therefore, based on the allele distribution at the two chosen loci, the entire DNA database can be partitioned into these distinct clusters. Further validation of the developed principal component analysis model was carried out with a real DNA/STR dataset. When performing a search for candidate profiles to match that of target profile, the target's DNA profile can be classified into one of these clusters by knowing its allele distribution pattern at the two loci under consideration. Thus, search can be restricted to members within this cluster only, thereby enhancing the search speed and thereby reducing the search time.

Once the search has reached a leaf node of the database tree, a list of candidate profiles is returned satisfying the match specifications. The matches returned by a search may be on the order of tens to hundreds, depending on the match requirement specification and it becomes difficult for caseworkers and analysts to examine all the returned profiles in order to identify the top suspects. Therefore, there comes the need for the development of a similarity measure to rank order the candidate profiles.

A Similarity Index (SI) parameter would reflect the degree of similarity of a candidate's DNA profile to that of the target. To test the feasibility of using the similarity index, a small as well as a large synthetic data set of candidate profiles of known rankings were created. Sensitivity analysis of the developed similarity index was also carried out to study the region of robustness of the similarity index in

response to changes in the weight parameter used in the similarity index. A flowchart showing the various steps used in the overall approach is shown in Figures 4.1 and 4.2.

4.1.1. SYNTHETIC DNA PROFILES GENERATION AND DNA PROFILES REPRESENTATION

The allele probability data were provided by the CODIS software for each of the sixteen STR loci for Caucasians, African Americans and Hispanic populations. The DNA/STR profiles are represented by a binary coding with a '1' designating the presence, and '0' the absence of an allele at a locus. The binary representation results in a sparse matrix of mostly zeros and some ones. For a profile consisted of all heterozygous loci (presence of two different alleles at a locus) we can have the presence of '1' a maximum of twice the number of loci present in that profile, and the minimum number of 1's present in a DNA/STR profile is equal to the number of loci present in that profile; that is if all loci exhibit homozygosity (presence of the same allele twice at a locus).

The synthetic binary data set of 10,000 profiles used in this research was generated from the known allele frequency distribution at the sixteen STR loci for the Caucasian population. Note that all loci have different allele lengths because the possible number of alleles available for each locus is different. The data matrix has a dimension of

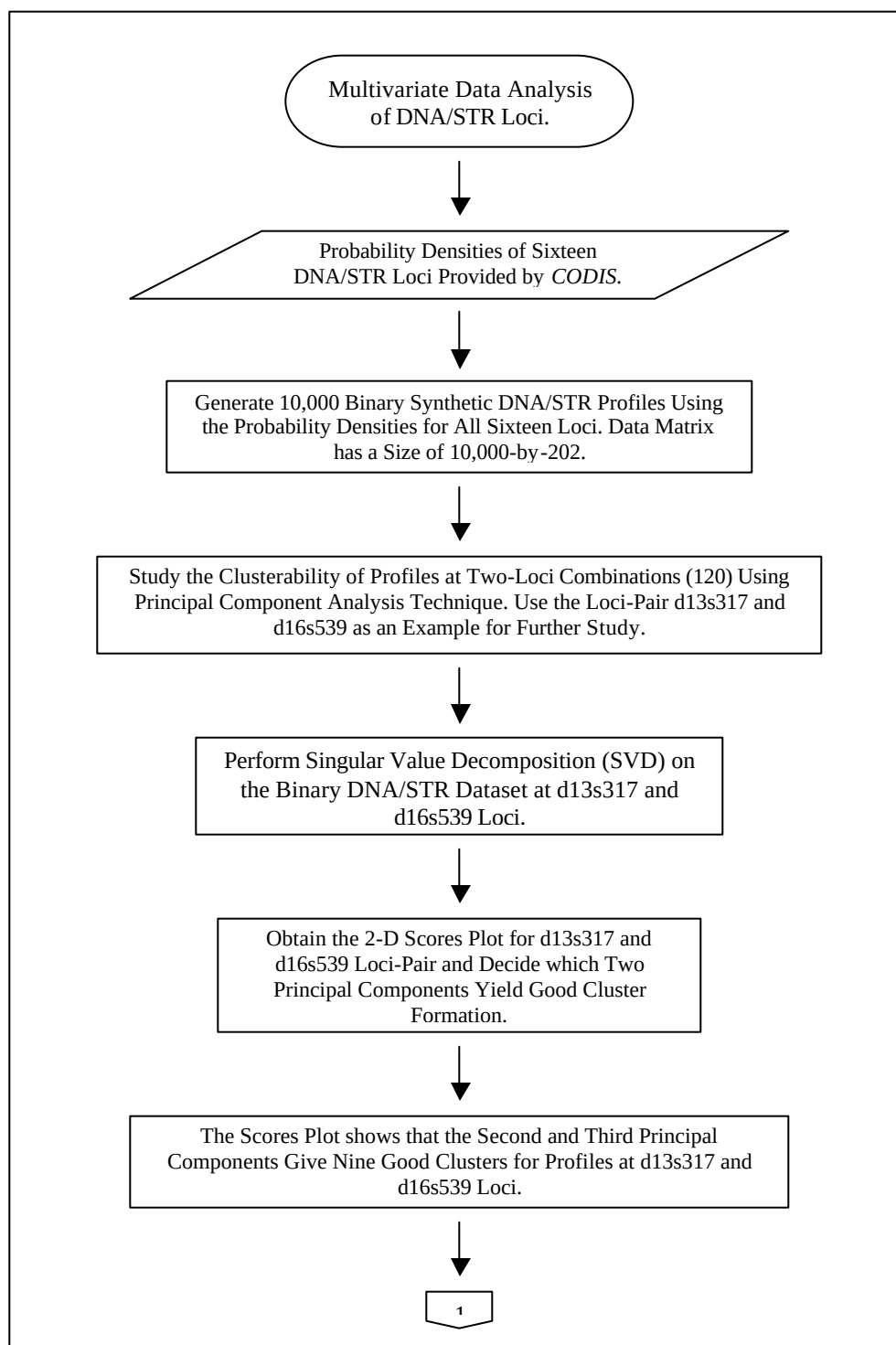


FIGURE 4.1: Flowchart for Clusterability Study Using Multivariate Data Analysis for DNA/STR Loci.

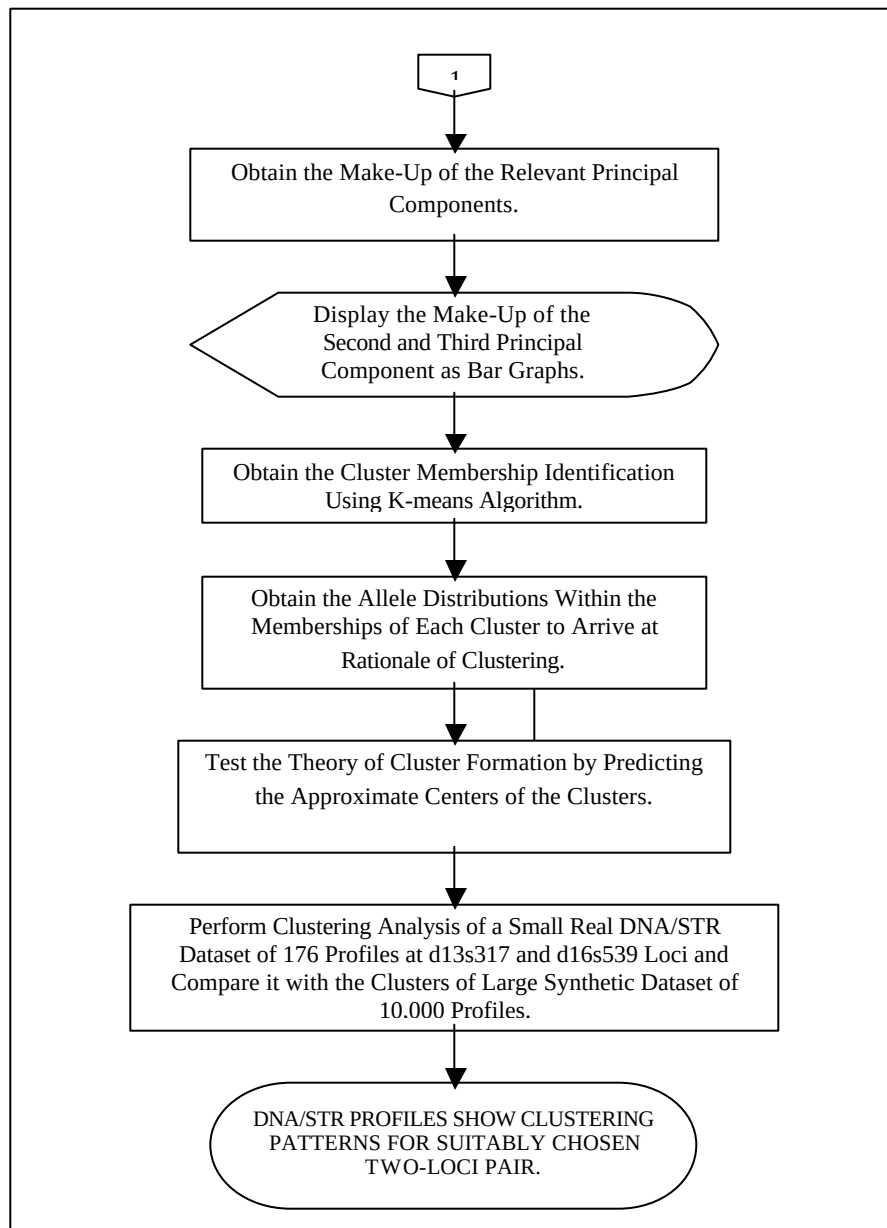


FIGURE 4.1: (Continued)

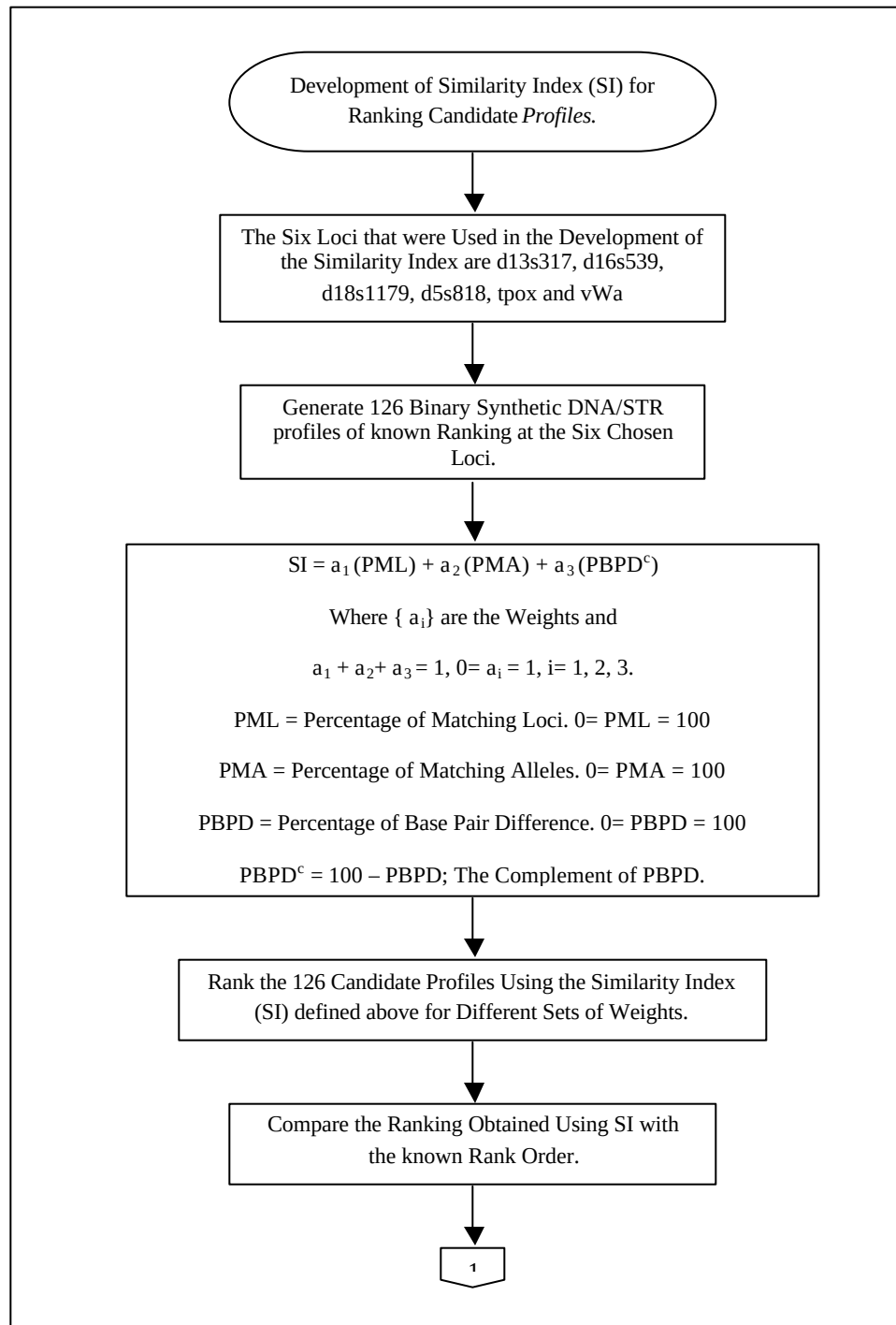


FIGURE 4.2: Flowchart for the Development of Similarity Index (SI) for Ranking Candidate Profiles.

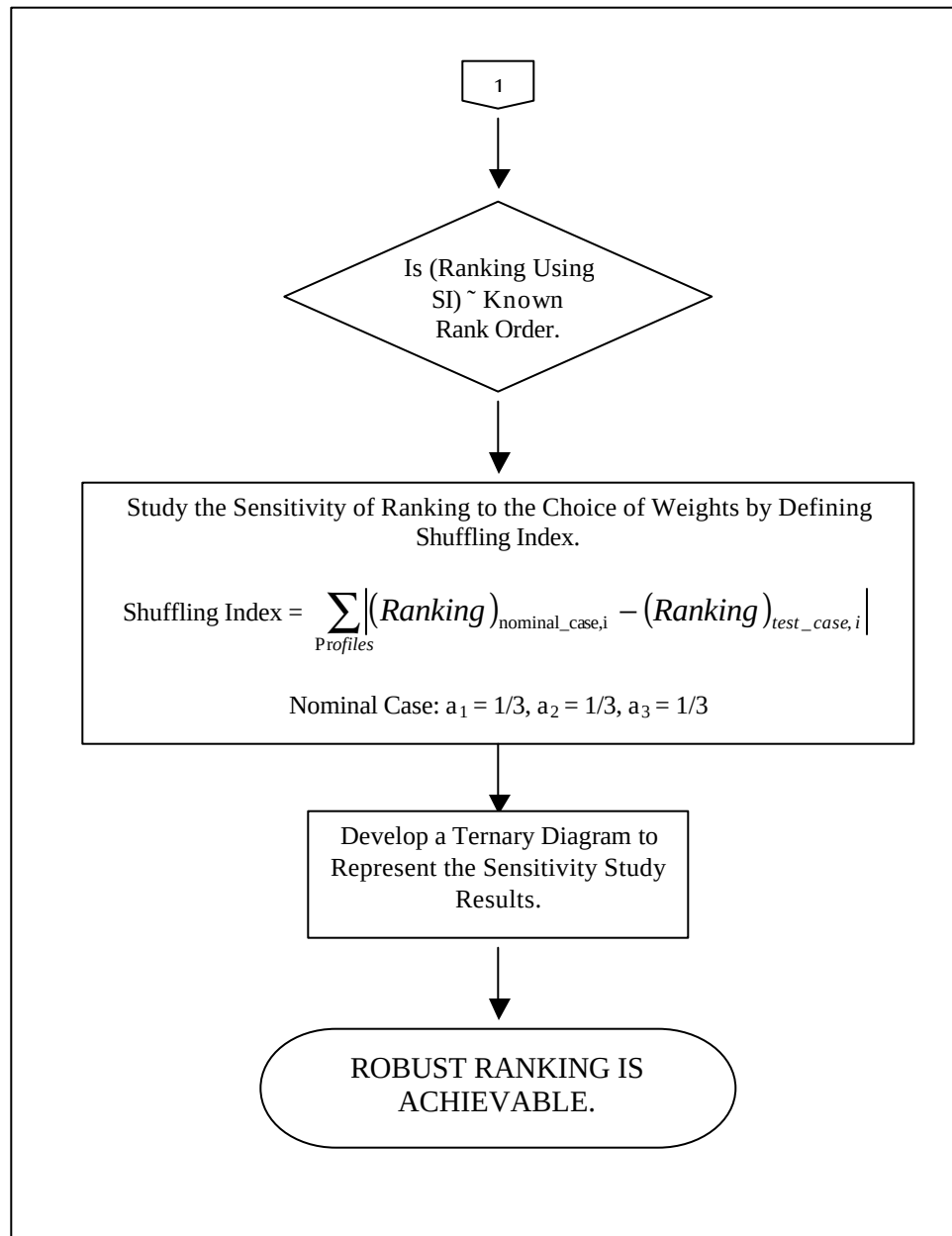


FIGURE 4.2: (Continued)

10000-by-202. (There are a total of 202 possible alleles from the sixteen loci.) Each row in the data matrix denotes a candidate DNA profile containing allele information at the 202-allele sites. The i^{th} column represents the presence '1', or the absence '0' of the corresponding i^{th} allele for all 10,000 profiles. When clusterability is studied using the loci pair, say of d13s317 (9 alleles)-d16s539 (11 alleles), the dimension of the sub-data matrix will then be reduced to 10000-by-20. Singular Value Decomposition (SVD) and principal component analysis are applied to the submatrix having only information on the loci pair under study. Table 4.1 gives an example of the binary matrix of twenty profiles for the d13s317 locus. Note that profiles 16 and 17 have only one allele present, thus representing the homozygous case.

4.1.2. DNA/STR CLUSTERABILITY STUDY

4.1.2.1. CLUSTERING BY PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal Component Analysis is an effective tool for data compression, information extraction, and cluster analysis. PCA finds combinations of the original system variables that describe trends in the data. PCA involves a mathematical procedure that transforms a number of correlated variables into a reduced number of uncorrelated variables called *principal components*. Figure 4.3 shows a two-dimensional principal component concept. The two principal components represent a change of coordinate basis vectors from those of the original. The first principal component is chosen such that the sum of the squares of the projections of the data points onto the chosen axis is

TABLE 4.1: Example of Binary Representation of DNA/STR Profile for Twenty Individuals at the d13s317 Locus. ‘1’ Represents the Presence of and ‘0’ Indicates the Absence of an Allele.

Profile ID	Alleles for Locus d13s317									# of Alleles
	7	8	9	10	11	12	13	14	15	
1	0	1	0	0	1	0	0	0	0	2
2	0	0	0	0	1	1	0	0	0	2
3	0	0	1	0	1	0	0	0	0	2
4	0	0	0	1	0	1	0	0	0	2
5	0	0	0	0	0	0	1	1	0	2
6	0	1	0	0	1	0	0	0	0	2
7	0	0	0	1	0	0	1	0	0	2
8	0	0	0	0	1	1	0	0	0	2
9	0	0	0	1	0	1	0	0	0	2
10	0	0	0	1	1	0	0	0	0	2
11	0	0	0	0	1	1	0	0	0	2
12	0	0	1	0	0	1	0	0	0	2
13	0	0	0	0	1	0	0	1	0	2
14	0	0	0	0	0	0	1	1	0	2
15	0	0	1	0	0	1	0	0	0	2
16	0	1	0	0	0	0	0	0	0	1
17	0	0	0	0	0	1	0	0	0	1
18	0	0	0	0	1	1	0	0	0	2
19	0	1	0	0	1	0	0	0	0	2
20	0	1	0	0	0	0	1	0	0	2

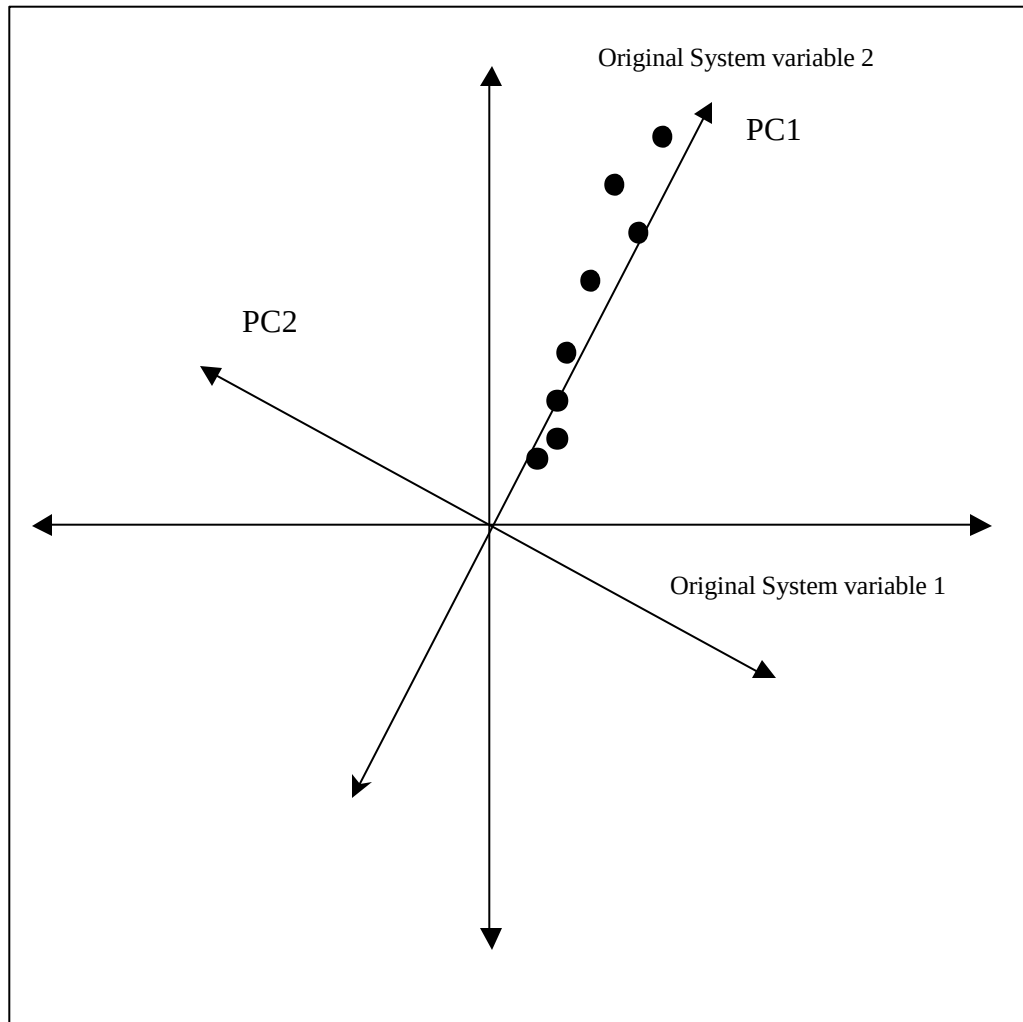


FIGURE 4.3: Illustration of Principal Component Analysis Concept, in which PC1 and PC2 Represent the New Principal Component Basis Vectors.

maximum of all possible new co-ordinate axis; The second principal component is chosen to be orthogonal to the first, and the sum of the squares of the projections of the data points onto this second axis is the next maximum, and so on for each of the subsequent principal components in a multidimensional principal component framework. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. PCA reduces the dimensionality of the data matrix and amounts to a rotation of coordinate axes to form a new but smaller coordinate system that has the described properties. There are two main objectives of doing the principal component analysis on a data matrix:

- To find the new co-ordinate axis to represent a reduced dimensional data space.
- To look for cluster formation when original data points are projected onto the leading principal component axes.

Implementing PCA involves several steps. First, the data may be column mean-centered, and often normalized by the standard deviation of each column. Mean centering implies that the mean of the centered value of each column is zero. Normalization is usually necessary to avoid problems associated with having some measurements with large values and others with small ones, perhaps because of the different physical units involved, that is, to prevent the problem of comparing ‘oranges to apples’. In this research, neither mean centering, nor normalization is performed because all measurements, that is, the 1’s and 0’s are all on the same footing, thereby no scaling is necessary. Without mean centering, the first principal component would

be interpreted to represent the ‘average’ size of the matrix; therefore not useful in cluster differentiation.

The preferred method to implement PCA is to use the Singular Value Decomposition (SVD) [45] of the data matrix. SVD decomposes the data matrix, X , into the product of three matrices, U , Σ , and V^T as seen in Equation 4.1, in which ^{T} denotes transpose. In Equation 4.1, the orthogonal vectors of the columns of the matrix, V , are referred to as the principal components of the data matrix, X , and they are all of unit length.

$$X = U \Sigma V^T \quad (4.1)$$

X = Original data matrix, size m -by- n , $n \ll m$ in this study.

U = Left singular matrix, size m -by- n , in which the columns are mutually orthogonal and of unit length.

V = Right singular matrix, size n -by- n , in which the columns are mutually orthogonal and of unit length.

Σ = Diagonal matrix, size n -by- n , with positive diagonal elements arranged in descending order of magnitude.

The columns of V , being the principal components, represent the new coordinate axes or basis vectors of the axes of the new frame of reference with respect to the old coordinate axes. The diagonal elements of the Σ matrix are referred to as the singular values of X . The percentage of total variation of the information contained in X , contributed by each principal component is given by the ratio of the square of each of the corresponding singular value, to the sum of squares of all the singular values. A

scree plot can be developed to show the cumulative ratio of this measure, plotted against the index number. The singular values are used in making a decision as to how many principal components to keep in building the PCA model to represent the original data. The data represented by the remaining principal components may be discarded for data analysis, as their contribution towards capturing the total variance in the original data is negligible. Due to correlation among the original variables, not all the principal components need to be used to fully represent the original data. This accounts for the dimension reducing effect of using the principal component analysis approach.

After the principal components are identified, each data profile (a row of the X matrix) is to be projected onto each of the chosen principal components. This step is carried out by multiplying the data matrix, X , by V_i , where V_i represents the i^{th} column of the V matrix, which is the i^{th} principal component vector (i.e. $\text{scores} = X \cdot V$, Equation (4.3)). The vector formed by the projections of profiles onto a principal component vector is referred to as the *scores* vector onto that principal component. The bigger the length of the projection of a profile onto the scores vector, the better this principal component represents that particular data profile. Thus, data profiles with comparable projections onto this principal component can be regarded as *similar* to each other, with respect to this principal component. Therefore, profiles which have high projections onto this principal component indicate that this principal component aligns more with these profiles, thus exhibiting the pattern of contribution by the original

variables to this principal component. Similarly, projections of data profiles onto each of the succeeding principal components can be carried out to obtain the respective scores.

As the degree of variation exhibited by the data profiles along the different principal components is not the same, normalization of the scores vector is important in order to make judicious comparison and analysis of the profiles as to their similarity. In order to obtain meaningful normalization the Mahalanobis distance measure is employed [48]. The normalization is carried out by multiplying each projection by the inverse of the corresponding singular value - the projection onto the i^{th} principal component is divided by the i^{th} singular value. Equation 4.2 shows the procedures.

$$\begin{aligned}\text{Mahalanobis Scores} &= \mathbf{XVS}^{-1} = (\mathbf{USV}^T)\mathbf{VS}^{-1} = (\mathbf{US})(\mathbf{V}^T\mathbf{V})\mathbf{S}^{-1} \\ &= \mathbf{U}(\mathbf{SS}^{-1}) = \mathbf{U}\end{aligned}\tag{4.2}$$

Scores or Projection of Profiles onto the Principal Component Vectors:

$$\text{Scores} = \mathbf{X*V}\tag{4.3}$$

In the above equation, $\mathbf{X}$ represents the original data matrix and $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$ are obtained from the SVD of $\mathbf{X}$, as shown in Equation 4.1. Since the $\mathbf{V}$ matrix is orthogonal, $\mathbf{V}^T$ is also equal to $\mathbf{V}^{-1}$.

All data profiles can be projected onto a two-dimensional subspace spanned by two of the leading principal components (i.e. PC3 & PC4, or PC2 & PC 3, or PC2 & PC4) to see if any distinct clustering pattern is detectable. Data profiles that are projected into the same cluster would imply that they are very similar in its variations in the original variables, which are the allele distribution patterns at the two loci. In the following sections it will be observed that projections onto the second and third principal components for DNA/STR profiles at the loci pair of d13s317 and d16s539 yield 9 distinct clusters. Our task is now to study the allele distribution pattern that is associated with the candidate profiles contained in each cluster to see what makes them cluster in the manner observed.

A numerical example consisting of thirty DNA/STR samples containing alleles for the loci-pair d13s317 and d16s539 is used to illustrate the principal component analysis procedures.

The binary data matrix, X (30-by-20), is shown in Equation 4.4, where 30 denotes the number of candidate profiles and 20 denotes the total number of alleles present in the loci pair d13s317 and d16s539. The three matrices U , Σ , and V obtained after SVD (Equation 4.1) are shown in Equations 4.5, 4.6, and 4.7 respectively. Recall that the columns of V are the principal components. Now the scores (Equation 4.3) or the projections of the profiles onto the second and the third principal component is determined as shown in Table 4.2. The obtained 2-D scores plot is shown in Figure

4.4. Distinct clusters are not seen because our example sample size is very small to capture all possible arrangements of alleles at these two loci. The purpose of this example is to illustrate the overall steps involved in clusterability study.

4.1.2.2.CLUSTER MEMBERSHIP ANALYSIS USING THE K-MEANS

ALGORITHM

When a distinct clustering pattern is discerned, the well-known k-means clustering algorithm [40, 48] was applied to validate the clustering of the normalized Mahalanobis scores of a large dataset (Equation 4.2), projected onto the 2-D principal component subspace. K-means is an iterative procedure, and in order to use the k-means algorithm *a-priori* information on the number of clusters has to be supplied. This is not a problem, as the number of clusters obtained can be obtained by visual inspection of the scores plot. The K-means algorithm consists of the following steps.

- An arbitrary set of m initial cluster centers is first chosen (in our case $m=9$). These are generally selected as the first m samples of the given sample set.
- Each data point is assigned to that cluster to which it is closest using some distance/similarity measure, depending upon the requirement of the problem. Usually the Euclidian distance measure is employed.
- New cluster centers are determined from the updated clusters.
- In the next round of iteration, clusters are re-formed by assigning the data to the new cluster center to which it is closest.
- Iteration continues until some convergence criteria are met.

[illegible]

$$U = \begin{pmatrix} 0.21 & -0.21 & 0.19 & -0.11 & 0.09 & -0.13 & 0.18 & 0.02 & 0.14 & -0.40 & -0.04 & 0.11 & -0.07 & -0.01 & 0.02 & -0.01 & 0.00 & 0.78 & 0.00 & 0.00 \\ 0.17 & -0.15 & -0.36 & 0.05 & 0.03 & 0.21 & -0.25 & 0.23 & 0.01 & -0.03 & 0.09 & 0.07 & 0.03 & 0.14 & 0.10 & -0.13 & -0.15 & 0.06 & -0.23 & -0.02 \\ 0.18 & 0.06 & 0.31 & -0.31 & -0.11 & 0.03 & -0.14 & -0.04 & -0.12 & 0.24 & 0.06 & 0.03 & 0.07 & -0.22 & -0.02 & -0.21 & 0.27 & 0.04 & -0.30 & -0.20 \\ 0.23 & -0.08 & -0.29 & -0.11 & -0.03 & 0.00 & 0.24 & 0.14 & -0.04 & 0.04 & 0.29 & 0.22 & 0.00 & -0.18 & 0.08 & -0.46 & -0.11 & -0.08 & 0.12 & -0.04 \\ 0.16 & -0.12 & -0.21 & 0.23 & -0.26 & -0.14 & -0.06 & -0.34 & 0.32 & 0.11 & -0.02 & 0.05 & 0.27 & 0.07 & 0.24 & 0.20 & -0.17 & 0.05 & -0.08 & -0.17 \\ 0.12 & -0.03 & 0.31 & 0.24 & 0.17 & 0.03 & 0.04 & 0.14 & -0.08 & -0.22 & -0.14 & 0.04 & 0.51 & 0.02 & -0.03 & -0.10 & -0.30 & -0.17 & -0.09 & -0.28 \\ 0.17 & -0.22 & -0.17 & -0.09 & -0.17 & -0.32 & 0.20 & 0.26 & -0.08 & 0.25 & -0.30 & -0.06 & -0.11 & -0.28 & 0.02 & 0.16 & 0.03 & -0.04 & 0.01 & -0.22 \\ 0.20 & 0.11 & 0.04 & -0.12 & 0.16 & 0.02 & 0.12 & -0.04 & 0.35 & 0.14 & 0.10 & 0.06 & 0.39 & -0.41 & 0.03 & -0.07 & 0.00 & -0.06 & -0.04 & 0.27 \\ 0.20 & -0.06 & -0.25 & 0.25 & 0.03 & 0.16 & 0.13 & 0.09 & -0.37 & 0.06 & 0.13 & -0.05 & 0.08 & 0.26 & 0.16 & -0.06 & 0.18 & 0.14 & 0.08 & 0.03 \\ 0.22 & -0.21 & -0.03 & -0.19 & 0.18 & -0.20 & 0.10 & 0.10 & -0.19 & 0.14 & 0.07 & 0.12 & 0.09 & 0.17 & -0.36 & 0.22 & -0.27 & -0.10 & 0.04 & 0.17 \\ 0.25 & 0.20 & 0.09 & 0.08 & 0.19 & -0.01 & 0.16 & 0.03 & 0.10 & 0.12 & 0.02 & -0.22 & 0.09 & 0.15 & 0.10 & -0.01 & -0.04 & 0.00 & 0.09 & 0.03 \\ 0.23 & 0.06 & -0.15 & -0.20 & -0.20 & 0.10 & 0.10 & -0.20 & -0.04 & -0.25 & -0.20 & -0.02 & -0.06 & 0.18 & 0.17 & -0.20 & 0.02 & -0.15 & -0.03 & 0.13 \\ 0.19 & 0.14 & 0.20 & 0.11 & 0.14 & -0.44 & -0.08 & -0.07 & -0.07 & 0.05 & 0.28 & 0.06 & -0.04 & 0.24 & 0.30 & -0.07 & 0.32 & -0.09 & -0.14 & 0.03 \\ 0.09 & -0.20 & -0.09 & 0.17 & -0.24 & -0.34 & -0.12 & -0.28 & 0.11 & 0.07 & -0.14 & -0.03 & -0.07 & -0.03 & -0.26 & -0.24 & -0.03 & -0.02 & 0.10 & 0.05 \\ 0.19 & 0.19 & 0.05 & -0.22 & -0.32 & 0.24 & -0.01 & 0.00 & 0.03 & 0.14 & 0.28 & 0.13 & -0.02 & 0.23 & -0.31 & 0.23 & -0.11 & 0.11 & -0.15 & -0.36 \\ 0.08 & -0.24 & 0.03 & 0.07 & -0.08 & 0.06 & 0.11 & 0.00 & 0.00 & -0.50 & -0.02 & 0.14 & -0.12 & -0.11 & -0.07 & 0.11 & 0.13 & -0.38 & -0.48 & 0.04 \\ 0.19 & 0.25 & -0.05 & 0.11 & -0.01 & 0.00 & 0.23 & 0.14 & 0.04 & -0.02 & 0.12 & -0.20 & -0.34 & -0.23 & -0.07 & 0.08 & -0.12 & -0.03 & -0.18 & 0.25 \\ 0.26 & -0.05 & -0.05 & -0.12 & 0.13 & 0.03 & 0.15 & -0.21 & 0.25 & -0.01 & 0.05 & -0.14 & -0.12 & 0.17 & 0.06 & 0.29 & 0.22 & -0.17 & 0.00 & -0.14 \\ 0.16 & -0.33 & 0.29 & 0.03 & -0.02 & 0.20 & -0.04 & -0.07 & 0.04 & -0.12 & 0.32 & 0.07 & -0.16 & -0.14 & 0.08 & 0.12 & -0.01 & -0.23 & 0.59 & -0.08 \\ 0.05 & -0.17 & 0.24 & 0.13 & -0.35 & 0.08 & 0.03 & 0.26 & 0.25 & 0.15 & 0.00 & 0.09 & 0.14 & 0.32 & -0.22 & -0.07 & 0.20 & -0.01 & -0.01 & 0.51 \\ 0.19 & 0.23 & -0.12 & -0.21 & -0.14 & 0.12 & -0.28 & 0.18 & 0.18 & -0.22 & -0.27 & 0.02 & 0.10 & -0.02 & -0.02 & -0.11 & 0.26 & -0.05 & 0.30 & -0.08 \\ 0.17 & -0.19 & 0.20 & -0.15 & -0.11 & 0.22 & -0.25 & -0.33 & -0.29 & 0.13 & -0.06 & -0.16 & -0.06 & -0.11 & 0.25 & -0.02 & -0.32 & 0.07 & -0.11 & 0.31 \\ 0.15 & 0.03 & 0.11 & -0.19 & 0.21 & -0.16 & -0.28 & 0.24 & 0.27 & -0.02 & -0.14 & -0.19 & -0.30 & 0.27 & 0.06 & -0.16 & -0.35 & -0.14 & -0.02 & -0.04 \\ 0.23 & -0.04 & -0.01 & 0.24 & 0.19 & 0.18 & 0.05 & -0.26 & -0.07 & 0.02 & -0.10 & -0.41 & -0.03 & -0.02 & -0.49 & -0.25 & 0.16 & 0.05 & 0.00 & -0.14 \\ 0.15 & 0.20 & -0.13 & 0.35 & -0.01 & -0.03 & -0.43 & 0.06 & 0.12 & -0.12 & 0.30 & 0.01 & -0.10 & -0.27 & -0.14 & 0.14 & -0.07 & 0.10 & -0.03 & 0.06 \\ 0.18 & 0.14 & 0.08 & 0.26 & 0.24 & 0.22 & 0.02 & -0.12 & 0.03 & 0.26 & -0.40 & 0.66 & -0.31 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.22 & 0.32 & -0.01 & 0.00 & -0.15 & 0.07 & 0.10 & 0.04 & -0.19 & -0.12 & -0.23 & -0.10 & 0.15 & -0.06 & 0.08 & 0.34 & -0.06 & 0.03 & 0.13 & 0.18 \\ 0.09 & -0.08 & 0.29 & 0.33 & -0.32 & 0.06 & 0.06 & 0.34 & 0.00 & 0.13 & -0.08 & -0.19 & -0.15 & -0.08 & 0.22 & -0.05 & -0.02 & 0.02 & -0.02 & -0.18 \\ 0.16 & 0.25 & 0.09 & 0.03 & -0.19 & -0.37 & -0.13 & -0.06 & -0.36 & -0.20 & 0.02 & 0.19 & 0.02 & 0.00 & -0.15 & -0.03 & -0.05 & -0.05 & 0.15 & 0.03 \\ 0.17 & -0.29 & -0.10 & -0.04 & 0.25 & 0.00 & -0.39 & 0.19 & -0.14 & 0.07 & -0.12 & -0.03 & 0.12 & -0.11 & 0.00 & 0.26 & 0.32 & 0.02 & -0.04 & 0.08 \end{pmatrix} \quad (4.5)$$

30 × 20

(4.6)

$$V = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0.11 & -0.28 & 0.39 & 0.22 & -0.17 & 0.13 & 0.18 & 0.34 & 0.19 & -0.58 & 0.03 & 0.29 & 0.24 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.20 & 0.24 & 0.11 & -0.33 & -0.41 & 0.16 & -0.29 & -0.20 & -0.43 & -0.17 & -0.38 & 0.09 & 0.31 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.17 & -0.27 & -0.35 & -0.04 & 0.10 & -0.06 & 0.02 & 0.49 & -0.43 & 0.32 & 0.15 & 0.30 & 0.34 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.40 & -0.18 & 0.48 & -0.09 & 0.58 & -0.01 & -0.16 & -0.22 & 0.12 & 0.24 & -0.11 & -0.02 & 0.27 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.46 & 0.32 & -0.40 & 0.19 & -0.05 & 0.49 & 0.13 & -0.13 & 0.39 & 0.07 & 0.13 & 0.07 & 0.21 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.08 & -0.21 & 0.02 & 0.25 & -0.45 & -0.28 & 0.05 & 0.11 & 0.33 & 0.42 & -0.51 & -0.18 & 0.15 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.11 & 0.07 & -0.04 & 0.28 & -0.19 & -0.56 & -0.38 & -0.34 & 0.06 & -0.06 & 0.41 & 0.32 & 0.13 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.39 & 0.92 & -0.04 & -0.01 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.66 & -0.26 & 0.56 & -0.43 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.61 & 0.29 & 0.72 & -0.15 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.22 & -0.07 & 0.40 & 0.89 & 0 & 0 & 0 \\ 0.03 & 0.04 & 0.02 & 0.08 & 0.08 & 0.09 & 0.01 & -0.06 & 0.02 & 0.16 & -0.37 & 0.75 & -0.49 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.30 & 0.35 & 0.18 & 0.64 & 0.10 & -0.06 & 0.08 & 0.16 & -0.46 & -0.03 & -0.08 & -0.24 & -0.19 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.12 & 0.00 & -0.18 & -0.01 & 0.12 & 0.06 & -0.75 & 0.45 & 0.23 & -0.19 & -0.12 & -0.14 & -0.22 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.49 & 0.28 & 0.06 & -0.49 & -0.07 & -0.44 & 0.30 & 0.26 & 0.14 & -0.08 & 0.08 & -0.01 & -0.24 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.41 & -0.62 & -0.30 & 0.01 & -0.07 & 0.02 & 0.07 & -0.34 & -0.18 & -0.25 & -0.05 & -0.17 & -0.32 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.13 & -0.14 & 0.41 & -0.06 & -0.41 & 0.35 & -0.16 & 0.08 & -0.05 & 0.40 & 0.47 & -0.04 & -0.30 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}_{20 \times 20} \quad (4.7)$$

TABLE 4.2: Scores of Candidate Profiles onto the Two Principal Components (V2 and V3)

CANDIDATE PROFILE ID	PRINCIPAL COMPONENT 1 (V2)	PRINCIPAL COMPONENT 2 (V3)
1	-0.21	0.19
2	-0.15	-0.36
3	0.06	0.31
4	-0.08	-0.29
5	-0.12	-0.21
6	-0.03	0.31
7	-0.22	-0.17
8	0.11	0.04
9	-0.06	-0.25
10	-0.21	-0.03
11	0.20	0.09
12	0.06	-0.15
13	0.14	0.20
14	-0.20	-0.09
15	0.19	0.05
16	-0.24	0.03
17	0.25	-0.05
18	-0.05	-0.05
19	-0.33	0.29
20	-0.17	0.24
21	0.23	-0.12
22	-0.19	0.20
23	0.03	0.11
24	-0.04	-0.01
25	0.20	-0.13
26	0.14	0.08
27	0.32	-0.01
28	-0.08	0.29
29	0.25	0.09
30	-0.29	-0.10

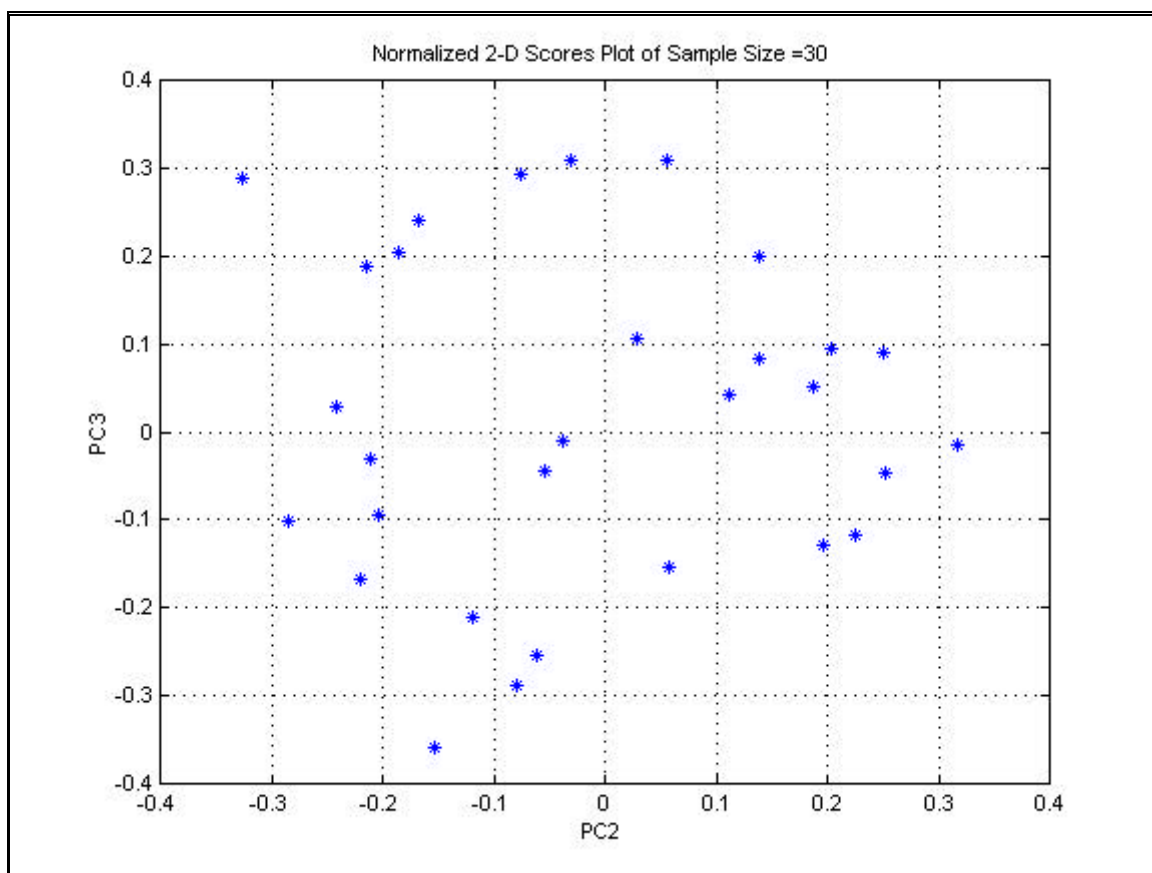


FIGURE 4.4: Normalized Scores of 30 Caucasian-DNA/STR Profiles onto the 2nd and 3rd Principal Components in d13s317 and d16s539 Loci.

Although no general proof of convergence exists for this algorithm, it can be expected to yield acceptable results when the data exhibit characteristic clustering, which are relatively distinct from each other. In most practical cases the application of this algorithm will require experimenting with various values of m but in this case the information on the number of clusters to start with is already known.

4.1.2.3. ALLELE DISTRIBUTION PATTERN OF MEMBERS CONTAINED WITHIN EACH CLUSTER

Using the k-means algorithm the memberships of each cluster is determined and how the allele distribution patterns of the member profiles resemble each other is then studied. The size of each cluster is also determined. Later it turns out that the size of each cluster is comparable to each other, an important finding that is beneficial in generating a more balanced decision tree for the search engine.

The description of how the development of the similarity index for ranking profiles after a search is described next.

4.2. DEVELOPMENT OF A SIMILARITY INDEX (SI) TO RANK ORDER PROFILES

The search algorithm under development will result at the lowest level of the decision tree, in a list of profiles, identified to be **similar** to that of the target profile, matching exactly or close to the target profile depending upon the match stringency criteria

specified. Figure 4.5 shows where the SI is to be employed. The list of probable candidates resulting after a search is otherwise arranged in an order according to their place of entry in the database, and it is difficult for caseworkers to identify the top closest candidates to follow up, as time plays an important role in solving crimes. The matches returned by a search maybe in the order of tens to hundreds, and it becomes difficult for caseworkers/analysts to manually look into all the returned profiles to identify the top mostly likely candidates. Developing the capability of rank ordering the candidate profiles with respect to their similarity to that of the target profile thus becomes important. Such a similarity measure will rank order the top few candidate profiles that match or are closest to the target profile and thus lessen the workload of caseworkers by short listing the number of probable suspects to the most probable suspects. The following section will talk about the factors relevant in the development of a similarity index.

4.2.1. THE SIMILARITY INDEX

The four factors relevant to the formulation of a Similarity Index (SI) are:

- The total number of matched loci between target and candidate profiles: The more loci that match the more similar the two profiles are.

The total number of matched alleles between target and candidate profiles: In general, the more alleles that match between the two profile the more they maybe similar. However it is not always true. There might be some inconsistencies. It would depend

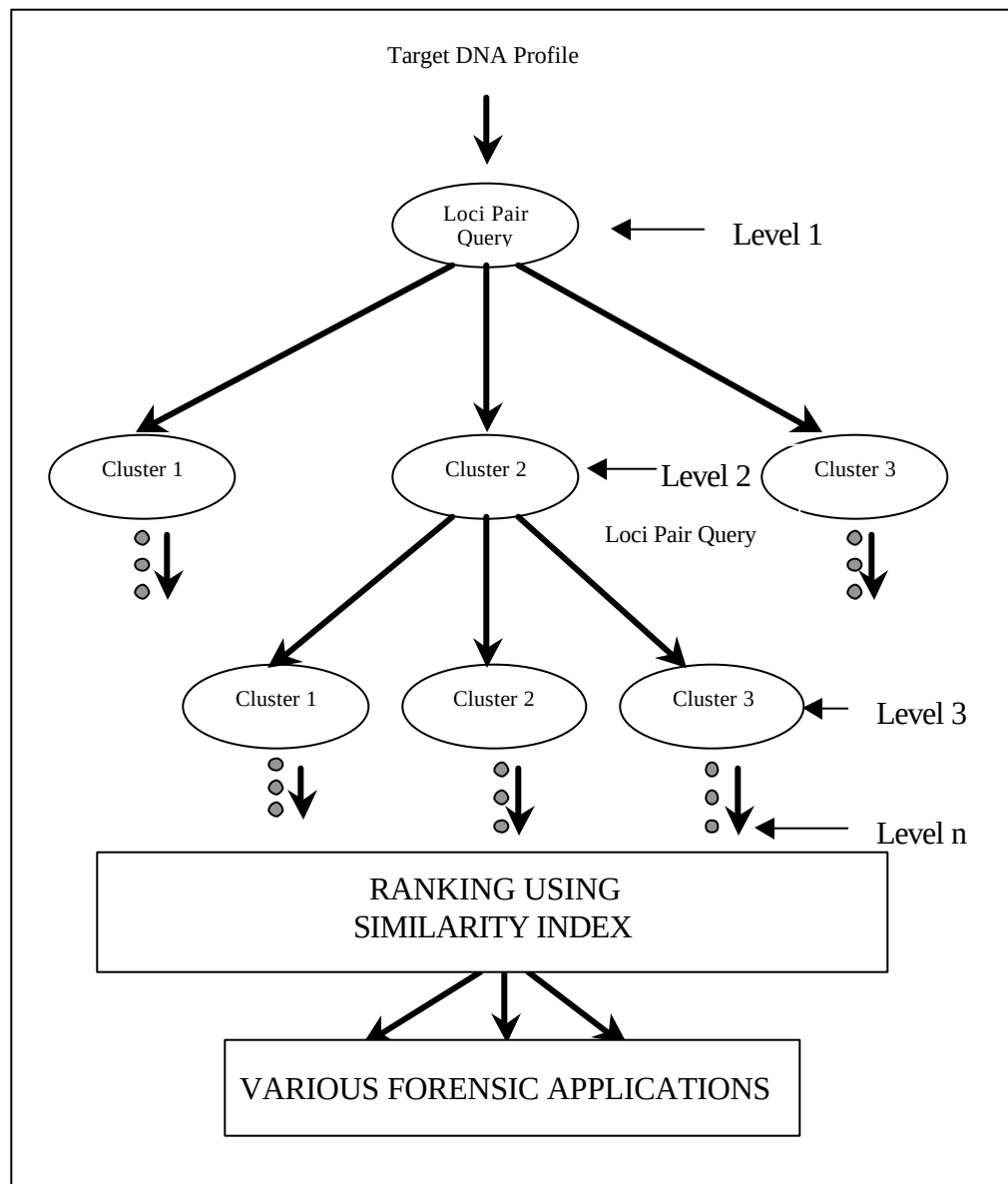


FIGURE 4.5: Utilization of Similarity Index (SI) in the Search Decision Tree.

on the distribution of the unmatched alleles. If all the unmatched alleles were dispersed across all different loci (i.e. 1 mismatched allele/locus) then the same number of loci as the number of unmatched alleles would be unmatched.

- This case is to be contrasted to that in which all the unmatched alleles disperse into only about half as many loci (i.e. 2 unmatched alleles/locus). In the latter case, a similar number of unmatched alleles would result in a different similarity measure. Therefore, the weight assigned to the percent of matching alleles should be much smaller than that assigned to the percent of matching loci in contributing to the overall similarity index (SI).
- The sum of differences in base pairs between all the unmatched alleles of target and candidate profiles. The further apart they are, the less similar they are.
- Special Consideration of mixture profiles in either the target or candidate profile: The profiles with mixtures are not ranked using Similarity Index parameter but are flagged for further analysis by the caseworkers, because a numerical SI value would not have the meaning it is intended when mixture profiles are considered.

In the application of the Similarity Index parameter, there might be situations when the SI may not be able to rank the candidate profiles efficiently because either candidate/target is a mixture, or the profiles are of unequal lengths. In such cases the candidate profiles that fall under this category would be flagged (Figure 4.6) to alert the caseworkers that they might want to examine these profiles manually. The blue bands in the figure correspond to the alleles that match the alleles present in the target profile (Column 2) and the red bands correspond to the mismatched alleles. This

Ambiguous Profiles

Allele	Target	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
7																	
8																	
9																	
10																	
11																	
12																	
13																	
14																	
15																	
16																	
17																	
18																	
19																	
20																	
21																	
22																	
23																	
24																	
25																	
26																	
27																	
28																	
29																	
30																	
31																	
32																	
33																	
34																	
35																	
36																	
37																	
38																	
39																	
40																	
41																	
42																	
43																	
44																	
45																	
46																	
47																	
48																	
49																	
50																	
51																	
52																	
53																	
54																	
55																	
56																	
57																	
58																	
59																	
60																	
61																	
62																	
63																	
64																	
65																	
66																	
67																	
68																	
69																	
70																	
71																	
72																	
73																	
74																	
75																	
76																	
77																	
78																	
79																	
80																	
81																	
82																	
83																	
84																	
85																	
86																	
87																	
88																	
89																	
90																	
91																	
92																	
93																	
94																	
95																	
96																	
97																	
98																	
99																	
100																	
Total # of Loci		4	4	4	4	4	4	4	4	4	4	2	3	4	4	4	4
Total # of mut. Loci		2	2	2	2	2	2	3	3	0	1	2	1	3	2	3	3
Total # of alleles		7	6	7	6	8	8	7	7	8	7	4	5	8	8	8	8
Total # of mut. Alleles		6	6	6	5	6	6	6	6	3	5	4	4	7	5	6	7

Assumption: The profiles are similar at other loci locations.

FIGURE 4.6: Examples of Ambiguous Profiles in which the Target and Candidate Profiles have Unequal Number of Loci or Alleles.

example shows a target profile with four loci for simplicity, assuming that profiles are similar at other loci locations. Candidates 5 and 9 may represent a mixture profile, as one of the loci has more than two alleles present. Incomplete/missing profile examples are shown in candidates 11, 12 and 15. There are many different arrangements that lead towards ambiguity that one must consider in the development of the similarity index.

Therefore, the Similarity Index is to have contributions from the degree of matching loci, matching alleles, and base pair difference of the unmatched alleles between that of the target and the candidate, weighted differently to reflect their degree of impact on the overall measure of similarity. The overall Similarity Index is to be calculated as in Equation 4.8. It is a weighted combination of the three relevant factors (expressed in percentage) discussed earlier.

In the equations 4.8 and 4.9, the ' α_i ' denotes the weight for each of the subparameters of the Similarity Index. The maximum base pair difference in Equation 4.12 is determined by calculating the maximum base pair difference possible between the two pairs of alleles in each locus. For example in the locus d13s317 (Table 2.4), it is seen that the maximum base pair spread occurs when the target DNA/STR profile has allele present at allele 7 (Bin 1) and allele 15 (Bin 9) respectively and the candidate's profile has allele present at allele 11 (Bin 5), i.e. it is homozygous.

$$SI = a_1(PML) + a_2(PMA) + a_3(PBPD^C) \quad (4.8)$$

Where

$$a_1 + a_2 + a_3 = 1; \quad 0 \leq a_i \leq 1, i = 1, 2, 3 \quad (4.9)$$

PML = Percentage of Matching Loci

$$= \frac{\text{Total Number of Matching Loci Across Candidate and Target}}{\text{Total Number of Loci Present Across Candidate and Target}} \times 100 \quad (4.10)$$

$$0 \leq PML \leq 100$$

PMA = Percentage of Matching Alleles

$$= \frac{\text{Total Number of Matching Alleles Across Candidate and Target}}{\text{Total Number of Alleles Present Across Candidate and Target}} \times 100 \quad (4.11)$$

$$0 \leq PMA \leq 100$$

PBPD = Percentage of Base Pair Difference

$$= \frac{\sum_{\text{unmatched alleles}} \text{Base Pair Difference}}{\sum_{\text{unmatched loci}} \text{Maximal Base Pair Difference}} \times 100 \quad (4.12)$$

$$0 \leq PBPD \leq 100$$

$$PBPD^C = 100 - PBPD; \text{The complement of } PBPD. \quad (4.13)$$

Therefore, in obtaining the maximum base pair deviation between the mismatched alleles, the difference between the mismatched alleles of the candidate with that of the target profile is calculated. For the homozygous case there are two alleles at the same allele location and in order to determine the difference the same alleles are considered twice. In our example, the maximum difference is computed as $[4 \times (|7-11| + |15-11|)] =$

32 bp (base pairs)]. A multiplying factor of 4 is used because each allele difference represents a difference of 4 base pairs. Note that all other possible distribution of alleles in the target/candidate profile will exhibit a base pair difference less than or equal to the calculated maximum base pair difference (32 bp). In the above example, there are an odd number of alleles present in the locus d13s317. One is to be added to the allele number difference obtained in case the number of alleles in the locus is even (for example locus d5s818). A simpler way to compute the maximum base pair difference for each locus is to take the difference between the last and the first allele index numbers. For example, consider the same locus d13s317, the difference in the last and the first allele index numbers is equal to 8 (the ninth and the first). Thus 32 bp (4×8) is the maximum base pair difference. This method gives the same number as before but the calculation is independent of the odd/evenness of alleles at each locus.

In interpreting the SI value, the higher the SI value the more similar to the target the profile is regarded. In practice, α_1 would be chosen to be quite a bit larger than α_2 , which would be chosen to be larger than α_3 to reflect the dominance of PML in determining the overall similarity.

As both PML and PMA are percentage based, if two profiles possess unequal number of loci, but with the same number of loci or alleles matching to that of the target profile, the one with fewer total loci will end up with a higher number for PML and PMA (because it would have a smaller number for the denominator in the calculation of PML and PMA), a misleading situation that would lead to an erroneous

interpretation of similarity. Similarity Index and ranking among profiles are meaningful only when the denominator in either the PML and PMA calculation for all the profiles whose similarity is to be compared are the same; meaning the same number of loci and alleles between the DNA of the target and that of all candidate profiles. Otherwise, one is comparing ‘oranges to apples’, asking the question, such as ‘Is 5 out of 5 loci matching (PML=100%) better than 10 out of 13 (PML= 77%) matching’. To overcome such contradictory arguments the development of some sort of a scaling scheme is required to apply in the future to the similarity index ranking parameter.

The third related parameter PBPD captures the percentage base pair difference in the mismatched alleles of a candidate profile against that of the target. Proper care should be taken in integrating PBPD into the overall SI. From Equation 4.12, it is seen that the related parameter PBPD has a positive (+) contribution in the calculation of SI, which implies that the PBPD value cannot be used as it is; because the greater the PBPD value the farther apart the candidate profile is from the target in the unmatched alleles, and vice-versa. Therefore, a higher PBPD value should translate to a smaller contribution to the similarity index. To account for the contrasting nature of PBPD, one has to take the complement of the true PBPD value such that a higher PBPD value would lead to a smaller addition to SI. The practice adopted here is to use the **complement** of PBPD from 100 and then to add it to the overall SI (as shown in Equation 4.8). The complement is denoted by $PBPD^c$, as in Equation 4.13.

Note that in general the higher the SI value, the more the profile is regarded to be similar to that of the target. In the application of the Similarity Index parameter, there might be situations when the SI may not be able to rank the candidate profiles meaningfully, such as in the ambiguous cases. In such cases the candidate profiles would be flagged, e.g. Figure 4.6, to alert the caseworkers for manual scrutiny of the profiles during follow-up.

4.2.2. DNA PROFILE GENERATION FOR TESTING THE SIMILARITY INDEX

A set of one hundred and twenty six synthetic profiles, as well as a bigger set of one thousand and twenty six profiles in six loci was generated, to be used in studying the feasibility of using the developed similarity index to rank order the profiles with respect to a given target profile. Table 4.3 and 4.4 show the characteristics of profiles generated for Similarity Index study for the small (126) and large (1026) datasets respectively. The profiles contain loci with either heterozygous or homozygous alleles. Six loci were chosen because it is simpler than handling the entire set of thirteen loci, and we believe that successful ranking of six-loci profiles would imply successful ranking of thirteen-loci profiles. The set also contains mixtures and profiles with unequal number of loci to that of target. In the original DNA data matrix, ‘1’ is used to indicate the presence of an allele and ‘0’ to indicate the absence of an allele at a particular locus. The six loci that were used to create synthetic profiles are as follows: d13s317, d16s539, d5s818, d18s1179, tpox and vWa. These loci were chosen because

Table 4.3: Types of Synthetic Profiles (126) used to Study the Feasibility of Similarity Index.

Total Number of Profiles	126
Target Profile to be Compared With	Heterozygous across all six loci
Total Number of Loci Used	6 (d13s317, d16s539, d18s1179, d5s818, tpox, vWa)
Number of Profiles with Mixtures	9 (ID#: 118-126)
Number of Ambiguous Profiles	12 (ID#: 106-117)
Number of Profiles with 1 mismatched allele to that of the target	1 (ID#: 1)
Number of Profiles with 2 mismatched allele to that of the target	9 (ID#: 2-10)
Number of Profiles with 4 mismatched allele to that of the target	30 (ID#: 11-40)
Number of Profiles with 5 mismatched allele to that of the target	20 (ID#: 41-60)
Number of Profiles with 8 mismatched allele to that of the target	10 (ID#: 61-70)
Number of Profiles with 9 mismatched allele to that of the target	10 (ID#: 71-80)
Number of Profiles with 10 mismatched allele to that of the target	5 (ID#: 81-85)
Number of Profiles with 11 mismatched allele to that of the target	5 (ID#: 86-90)
Number of Profiles with 12 mismatched allele to that of the target	9 (ID#: 91-99)
Number of Profiles that are exact replica of the target	1 (ID#: 100)
Number of Profiles with an extra 7th Locus	5 (ID#: 101-105)

Table 4.4: Types of Synthetic Profiles (1026) used to Study the Feasibility of Similarity Index.

Total Number of Profiles	1026
Target Profile to be Compared With	Heterozygous across all six loci
Total Number of Loci Used	6 (d13s317, d16s539, d18s1179, d5s818, tpox, vWa)
Number of Profiles with Mixtures	9 (ID#: 1018-1026)
Number of Ambiguous Profiles	12 (ID#: 1006-1017)
Number of Profiles with 1 mismatched allele to that of the target	1 (ID#: 1)
Number of Profiles with 2 mismatched allele to that of the target	99 (ID#: 2-100)
Number of Profiles with 4 mismatched allele to that of the target	300 (ID#: 101-400)
Number of Profiles with 5 mismatched allele to that of the target	200 (ID#: 401-600)
Number of Profiles with 8 mismatched allele to that of the target	100 (ID#: 601-700)
Number of Profiles with 9 mismatched allele to that of the target	100 (ID#: 701-800)
Number of Profiles with 10 mismatched allele to that of the target	50 (ID#: 801-850)
Number of Profiles with 11 mismatched allele to that of the target	50 (ID#: 851-900)
Number of Profiles with 12 mismatched allele to that of the target	99 (ID#: 901-999)
Number of Profiles that are exact replica of the target	1 (ID#: 1000)
Number of Profiles with an extra 7th Locus	5 (ID#: 1001-1005)

they have been shown to contribute to good clusterability by the principal component analysis approach.

4.3. SENSITIVITY ANALYSIS OF SIMILARITY INDEX TO THE SELECTION OF WEIGHTS FOR THE SIMILARITY INDEX SUBPARAMETERS

The main objective of performing the sensitivity analysis was to determine the range of weights for the three Similarity Index (SI) subparameters, in which the similarity rankings of the top ranked candidates remain relatively robust. Robustness means that the rank of the top few candidates would remain relatively insensitive to the choice of weights used in equation 4.8, in calculating the Similarity Index. High sensitivity to the choice of weights would render the use of the SI less meaningful and lead to a subjective interpretation of the ranking results.

First, a Shuffling Index needs to be developed to reflect the overall change in the rankings of the profiles. The Shuffling Index is developed as shown in Equation 4.14, which is just the sum of the absolute values of the change of rankings of the corresponding profiles when different weighting is used.

Note that the Shuffling Index is always a positive number and that a larger index implies the more shuffling associated with the profiles using that test_case weights.

Shuffling Index = Overall Sum of the Absolute Value of the Difference between the Respective Rankings of the Profiles in the Nominal Case and that of the Corresponding Profiles in the Test case.

$$= \sum_{\text{Profiles}} |(Ranking)_{\text{nominal_case},i} - (Ranking)_{\text{test_case},i}| \quad (4.14)$$

Where $(Ranking)_{\text{test_case},i}$ denotes the ranking of the i^{th} profile using the test_case weights.

4.3.1. SELECTION OF WEIGHTS

Revisiting the three parameters that make up the overall similarity index.

- PML: Percentage of Matching Loci (Equation 4.10)
- PMA: Percentage of Matching Alleles (Equation 4.11)
- PBPD^c: Complement of the Percentage of Base Pair Difference (Equation 4.12), i.e. (1-PBPD)

A graph would be nice to represent the specification of the three weights ($\alpha_1, \alpha_2, \alpha_3$), to aid in the visual perception of the range of weights over which the shuffling is calculated. Thus, a ternary diagram is chosen to fulfill this role. Ternary diagrams are used in the field of physical chemistry [46, 47] to represent thermodynamic and concentration information associated with a mixture. Figure 4.7 represents such a diagram. Starting from the vertex point; 'p', sides $\overline{pq}$, $\overline{qr}$, and $\overline{rp}$ represent the weight for PML (α_1), PMA (α_2), and PBPD^c (α_3) respectively, ranging from 0 to 100% for all

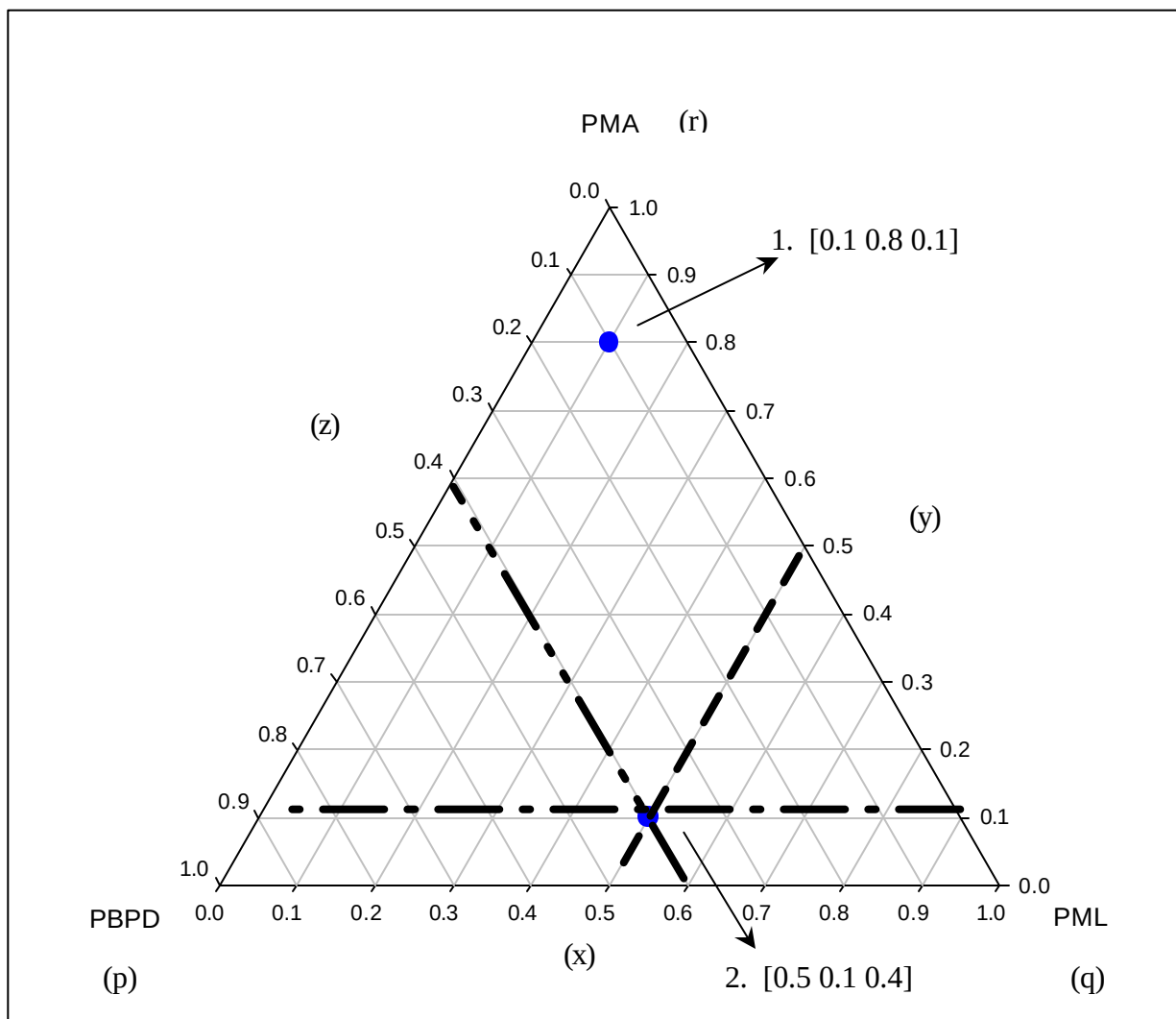


FIGURE 4.7: Weights for the Similarity Index Subparameter Expressed on a Ternary Equilateral Triangle; PML= Percentage of Matching Loci; PMA= Percentage of Matching Alleles; $PBPD^c = 1 - PBPD$ (Percentage of Base Pair Difference), the Sides pq, qr, and rp Represent the Weight (a_1 , a_2 , and a_3) for PML, PMA, and $PBPD^c$ Respectively.

three parameters. Each point in the interior of the triangle of the ternary diagram represents a specific combination of the three-parameter weights.

Here are the rules for determining the three weights represented by a point inside the triangle. At each point of interest, draw three lines, each passing through this point and with a line parallel to each of the three sides of the triangle.

1. To identify the weight for PML, first, locate the vertex that represents 100% PML. This point is labeled as 'q' in the diagram. Then examine the line segments representing the two sides of the triangle that emanate from this vertex. In this case, it is the line segment $\overline{qr}$ and $\overline{qp}$ (They form a 60° angle). From the three lines drawn at the point of interest, pick out the line that, together with the two lines emanating from vertex 'q', forms a closed equilateral triangle. Locate the point of intersection of this line with the 'PML' line ($\overline{pq}$) and read off the corresponding PML value on $\overline{pq}$.
2. Similarly proceed to read off the PMA and PBPD^c values.

For example, consider point '2' shown in the figure near the bottom of the triangle with co-ordinates of ($\alpha_1=0.5$, $\alpha_2=0.1$, $\alpha_3=0.4$). The values 0.5, 0.1, and 0.4 are the weights for PML, PMA, and PBPD^c respectively. Following the steps mentioned above, draw three lines passing through point 2 and parallel to the three sides. To identify the weight for PML, first locate vertex 'q'. Then draw lines $\overline{qx}$ and $\overline{qy}$

emanating from vertex 'q'. Now line $\overline{yx}$ forms a closed equilateral triangle with line segments $\overline{qx}$ and $\overline{qy}$ and intersects $\overline{pq}$ at $0.5(=\alpha_1)$. This gives us the corresponding PML value of 0.5. Similarly locate the weights for PMA ($y=0.1=\alpha_2$) and PBPD^c ($z=0.4=\alpha_3$).

Having constructed the ternary diagram, a large set of discrete weights was generated consisting of 4182 sets of weights in the range shown below.

- $0 \leq \alpha_1 \leq 1.0$; weight for PML.
- $0 \leq \alpha_2 \leq 1.0$; weight for PMA.
- $0 \leq \alpha_3 \leq 1.0$; weight for PBPD^c and
- Subject to $\alpha_1 + \alpha_2 + \alpha_3 = 1.0$ (Equation 4.9).

Having created the matrix (4182-by-3) of weights, the effect of SI subparameter weights on the shuffling of rankings obtained on each of the two sets of synthetic profiles (The small and big sets) was studied. The shuffling of subsequent rankings when compared to that of a reference set was investigated. The new parameter called the *Shuffling Index*, as described in Equation 4.14, was used to calculate the changes in rankings of the top ranked profiles. The reference set of weight chosen was the midpoint of the ternary triangle, (1/3,1/3,1/3). The shuffling index determines the net sum of the shift in the rankings of candidate profiles. Note that, in calculating the shuffling index, it is the absolute difference in the change of rankings of the corresponding profiles that are added together. Therefore, there will be no cancellation between profile rankings being shifted up and others being shifted down.

4.3.2. RANKING SENSITIVITY STUDY: THE SHUFFLING INDEX AND SHUFFLING INDEX CONTOUR PLOTS

Having created the matrix of weights, the task was reduced to study the effect of weight selection on the shuffling of ranks obtained by our similarity measure. The shuffling of ranks was determined, by defining the new parameter called the Shuffling Index to reflect the relative changes in the rankings of the corresponding profiles when different weights are used in forming the Similarity Index (SI). Equation 4.14 shows how the shuffling index is to be calculated. The shuffling is studied using the reference weight ($\alpha_1=1/3$, $\alpha_2=1/3$, $\alpha_3=1/3$) that assigns equal importance to all the three parameters in forming the SI. It is the median point of the equilateral triangle of the ternary diagram. The sensitivity of ranking of the top ten, twenty and hundred candidates is also performed for different set of reference weight sets, as will be seen later. The Shuffling Index determines the juggling or difference in the ranks of the candidate profiles from that obtained using the reference weights, when different set of weights is used in doing the ranking.

The sensitivity study based on percentage of shuffling is performed to make appropriate normalized comparison for different set of weights and different number of top ranked profiles based on which the shuffling is calculated. Percentage shuffling is defined in Equation 4.15.

$$\text{Percentage Shuffling} = \frac{\text{Shuffling Index (Eq.4.14)}}{\text{Maximum Possible Shuffling}} \cdot 100 \quad (4.15)$$

The maximum possible shuffling varies depending upon the sample size, and the number of top ranks one is considering for studying the sensitivity of the ranked profiles to different set of weights. The Shuffling Index or percentages obtained from using the data set of possible weight values can be plotted as contour curves, superimposed on top of the parameter weight ternary diagram. At a glance, the stability of profile rankings to changes in the SI subparameter weights can be discerned by visually examining the contour plot patterns.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1. CLUSTERING BY MULTIVARIATE STATISTICAL ANALYSIS

The synthetic binary data set of 10,000 profiles was generated from the known allele frequency distribution for each of the sixteen STR loci for Caucasian population. Note that all loci have different number of alleles. The overall data matrix has a dimension of 10000-by-202, 10000 profiles and 202 alleles. Each row in the data matrix denotes one candidate DNA profile containing alleles at all sixteen loci. Each column represents the presence '1', or the absence '0' of the corresponding allele for all 10,000 individuals/candidates. If studying the clusterability using the d13s317 (9 alleles)-d16s539 (11 alleles) loci-pair, the dimension of the sub-data matrix is reduced to 10000-by-20: only the twenty columns corresponding to the twenty alleles of the two loci of interest are lifted out of the original overall matrix, to form the submatrix which is then subjected to principal component analysis. The profiles were first represented in a binary form such as that shown in Table 5.1 Based on the allele distribution pattern of two suitably chosen STR loci, clusterable patterns are discernable after principal component analysis (PCA) of the data matrix.

TABLE 5.1: Example of Binary Representation of Twenty DNA/STR Profiles with Alleles at the d13s317 Locus.

Profile ID	Alleles for Locus d13s317									# of Alleles
	7	8	9	10	11	12	13	14	15	
1	0	1	0	0	1	0	0	0	0	2 (Heterozygosity)
2	0	0	0	0	1	1	0	0	0	2
3	0	0	1	0	1	0	0	0	0	2
4	0	0	0	1	0	1	0	0	0	2
5	0	0	0	0	0	0	1	1	0	2
6	0	1	0	0	1	0	0	0	0	2
7	0	0	0	1	0	0	1	0	0	2
8	0	0	0	0	1	1	0	0	0	2
9	0	0	0	1	0	1	0	0	0	2
10	0	0	0	1	1	0	0	0	0	2
11	0	0	0	0	1	1	0	0	0	2
12	0	0	1	0	0	1	0	0	0	2
13	0	0	0	0	1	0	0	1	0	2
14	0	0	0	0	0	0	1	1	0	2
15	0	0	1	0	0	1	0	0	0	2
16	0	1	0	0	0	0	0	0	0	1 (Homozygosity)
17	0	0	0	0	0	1	0	0	0	1
18	0	0	0	0	1	1	0	0	0	2
19	0	1	0	0	1	0	0	0	0	2
20	0	1	0	0	0	0	1	0	0	2

5.1.1. IDENTIFICATION OF CLUSTERS FOR DNA/STR PROFILES FOR CERTAIN TWO-LOCI COMBINATIONS

A rigorous study of the clusterability of two-loci combinations for DNA of 10,000 synthetically generated Caucasians was carried out. A total of sixteen DNA loci were available for profile identification purposes (Currently, a subset of thirteen loci is required by FBI for state agencies to request a search in the national CODIS system). Therefore, a total of 120 ($C_2^{16} = 16*15/2 = 120$) two-loci combinations were analyzed using PCA to decipher the clusterability of the profiles under consideration. The two-loci combinations, the identity of the corresponding principal components and the number of clusters that potentially resulted in good and distinct clusters, were determined by visual inspection of the relevant scores plots and are shown in Table 5.2. K-means clustering algorithm will be used later to identify and validate the memberships of each cluster formed by the DNA information contained in a loci pair. The reason that only certain two loci combinations yield good clusters will be discussed in subsequent sections to show how the distribution of alleles in each locus will or will not lead to cluster formation.

At this point, let us look into some examples of good cluster formation. Table 5.2 shows that in most cases, the second and the third principal components are the ones to differentiate between clusters. In Table 5.2, it is noted that the number of clusters is either seven or nine in number. Let us also keep in mind that *a-priori* information of the number of distinct clusters from principal component analysis is available, and

TABLE 5.2: Combinations of Two-Loci that Potentially give Good Clusters after Visual Inspection of the Scores Plot.

(This is not the final list of loci pair. In some cases one does not need loci pair for good clusters.)

Locus 1	Locus 2	Principal Component	Principal Component	# of Clusters	Locus 1	Locus 2	Principal Component	Principal Component	# of Clusters
csf1po	fga	2	3	7	d18s51	d8s1179	2	4	7
csf1po	tpox	2	3	7	d18s51	fga	2	3	7
d13s317	d16s539	2	3	9	d18s51	tpox	4	5	7
d13s317	d1s80	2	3	9	d18s51	vWa	2	3	7
d13s317	d21s11	2	3	9	d1s80	fga	3	4	7
d13s317	d5s818	2	3	9	d21s11	d5s818	2	3	9
d13s317	d7s820	2	3	9	d21s11	d7s820	2	3	9
d13s317	fga	3	4	7	d21s11	fga	2	5	9
d13s317	tho1	2	3	9	d21s11	tho1	2	3	9
d13s317	vWa	2	3	9	d21s11	vWa	2	3	9
d16s539	d1s80	2	3	9	d3s1358	fga	2	3	7
d16s539	d5s818	2	3	9	d3s1358	tpox	2	3	7
d16s539	d7s820	2	3	9	d5s818	d7s820	2	3	9
d16s539	fga	4	5	7	d5s818	tho1	2	3	9
d16s539	tho1	2	3	9	d5s818	vWa	3	4	7
d16s539	vWa	2	3	9	d7s820	d1s80	2	3	9
d18s51	d1s80	4	5	7	d7s820	d8s1179	2	3	9
d18s51	d21s11	2	3	7	d7s820	vWa	2	4	7
d18s51	d3s1358	2	3	7	d8s1179	fga	3	4	7
d18s51	d5s818	2	3	7	vWa	fga	2	3	7

this information is used later on as an input to the k-means clustering algorithm to start the iteration process.

5.1.2. EXAMPLES OF LOCI PAIRS THAT GIVE GOOD CLUSTERS

From Table 5.1 it is seen that potentially forty possible two-loci pairs of DNA/STR combinations resulted in good and well-separated cluster formation. The allele frequency distribution for all possible two-loci combinations was studied and analyzed to determine the reason behind the formation of distinct clusters. It will be presented in greater details in the following sections.

Now let's look at some examples of good clusters for Caucasians DNA/STR profiles. The scores plot showing clusters for different two-loci cases are nearly of the same structure, resembling to animals 'PAW' prints, though they might differ in the exact appearance. Figures 5.1 – 5.6 show some examples of the paw prints of clusterable two-loci combinations. Figures 5.1-5.5 show nine clusters each, and Figure 5.6 is an example representing seven clusters. Having determined the number of clusters, seven or nine, the memberships of each cluster can then be identified with the analytical K-means clustering algorithm.

Having seen the formation of clusters, the reason behind the clusterability of DNA/STR containing different loci pairs is explored. In the following section, the example of d13s317 and d16s539 loci pair will be used for illustration.

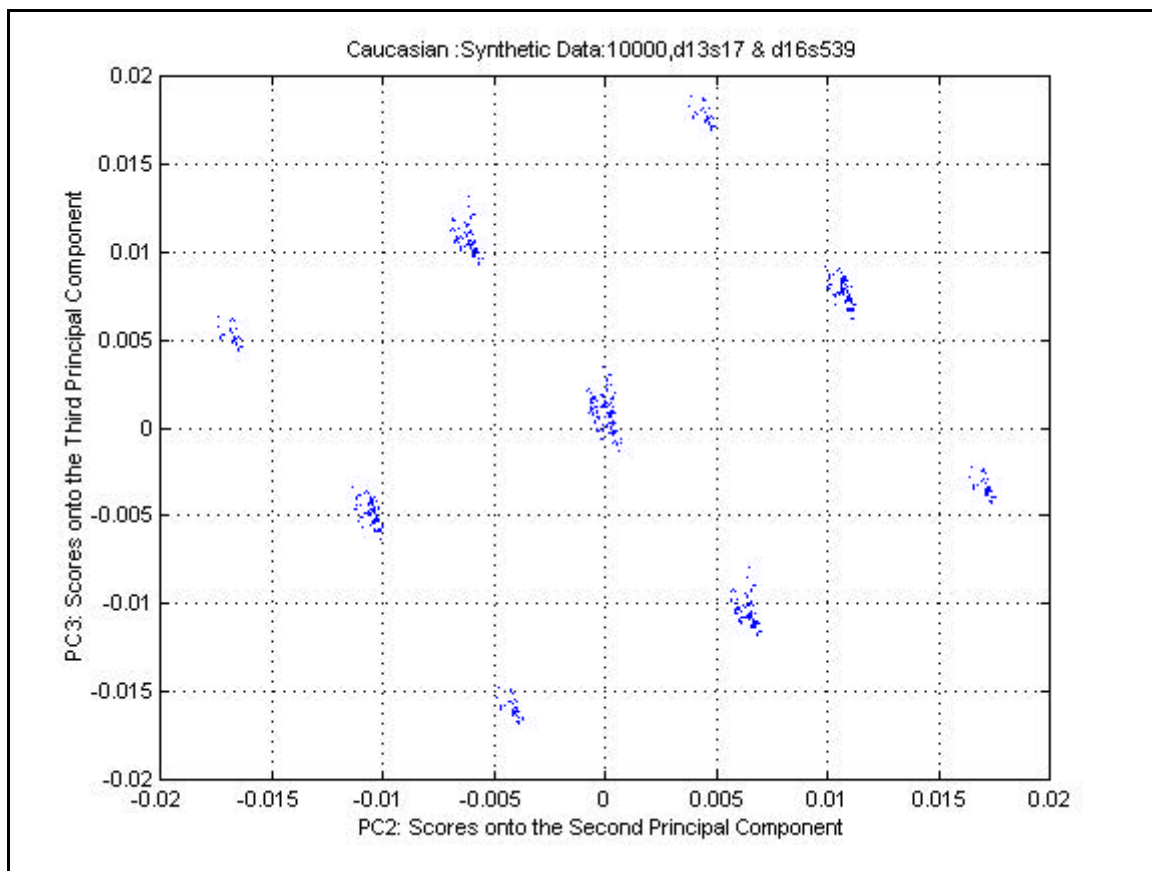


FIGURE 5.1: Scores of 10,000 Caucasian-DNA/STR Profiles onto the 2nd and 3rd Principal Components in d13s317 and d16s539 Loci.

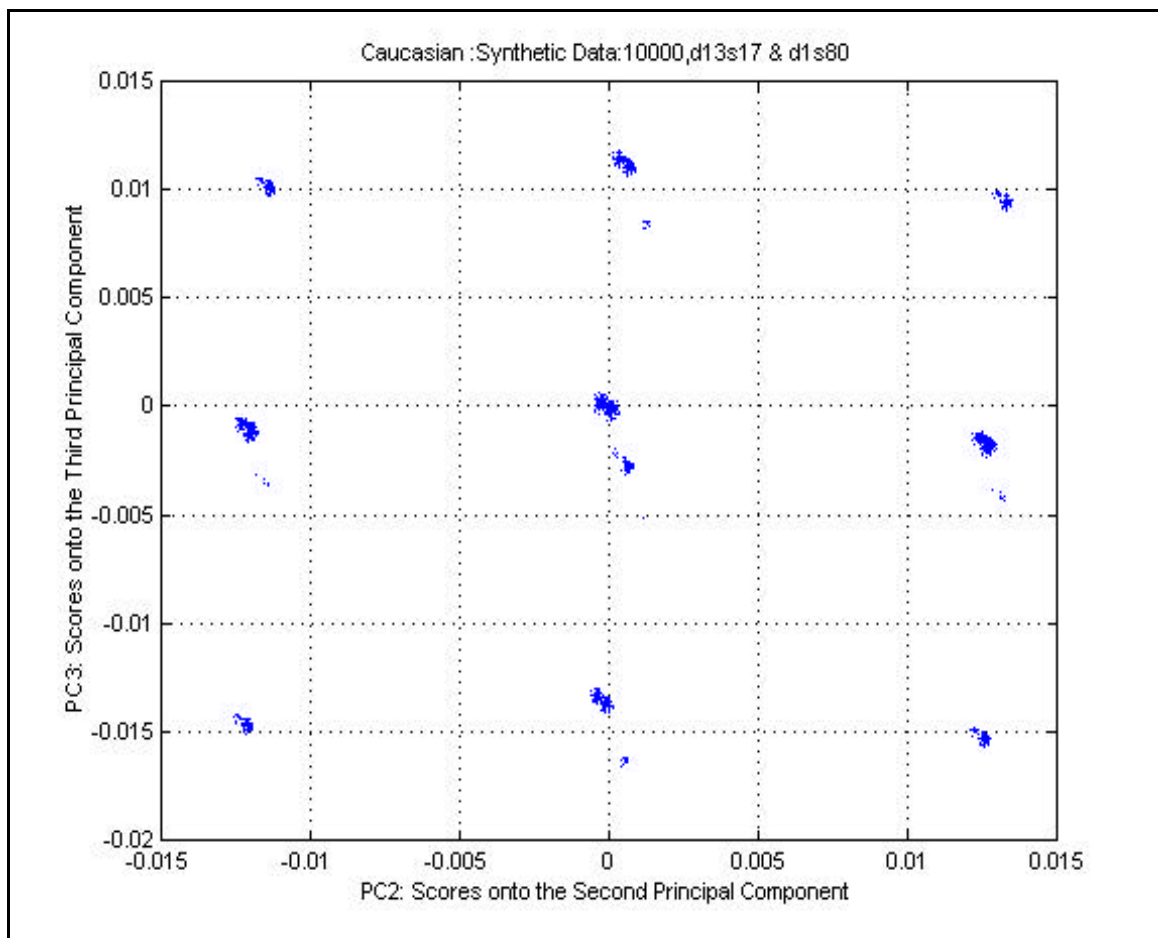


FIGURE 5.2: Scores of 10,000 Caucasian-DNA/STR Profiles onto the 2nd and 3rd Principal Components in d13s317 and d1s80 Loci.

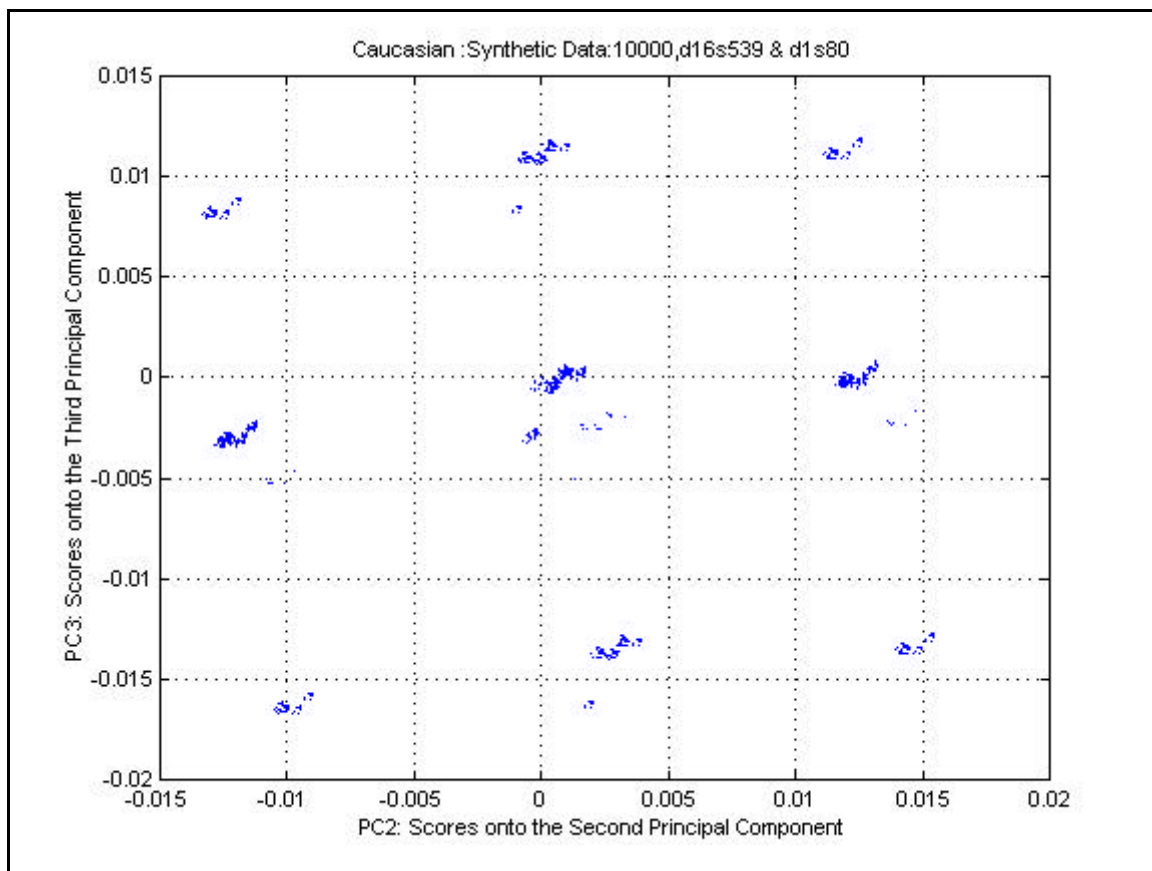


FIGURE 5.3: Scores of 10,000 Caucasian-DNA/STR Profiles onto the 2nd and 3rd Principal Components in d16s539 and d1s80 Loci.

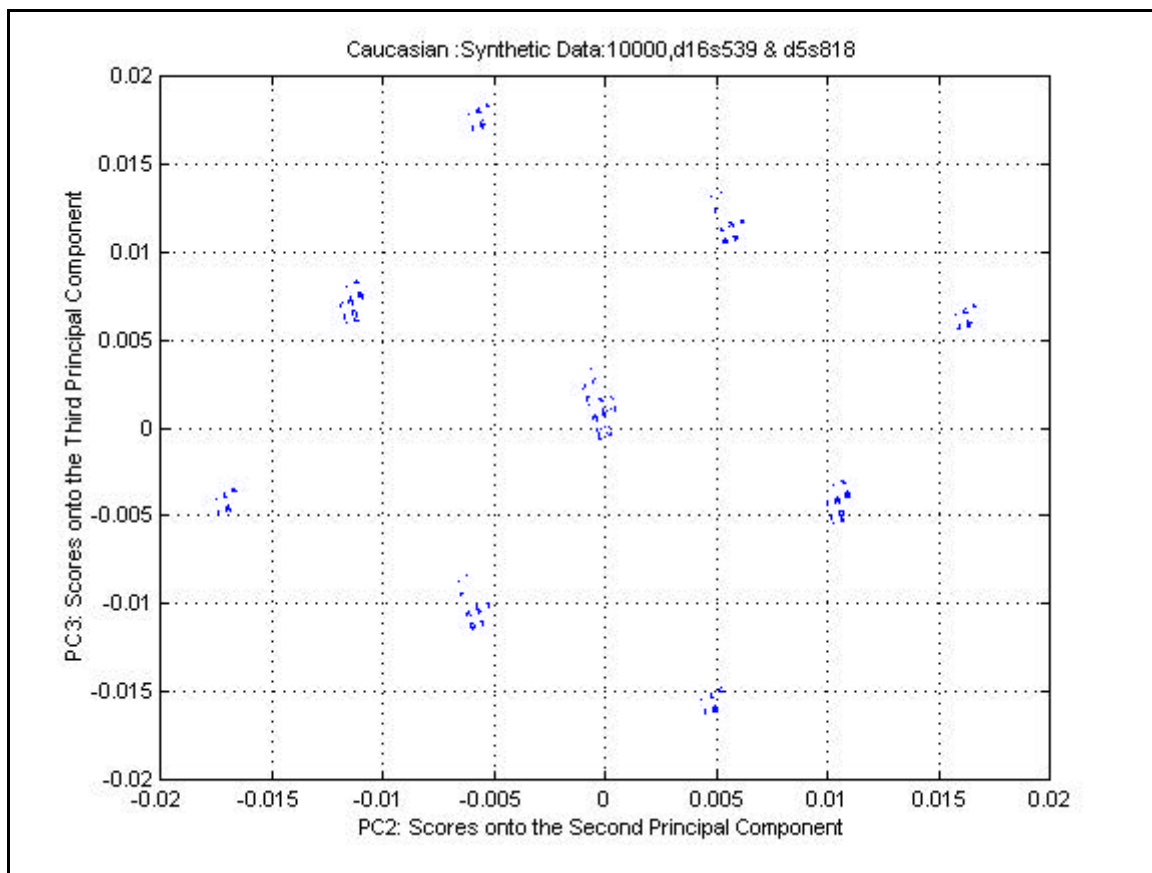


FIGURE 5.4: Scores of 10,000 Caucasian-DNA/STR Profiles onto the 2nd and 3rd Principal Component in d16s539 and d5s818 Loci.

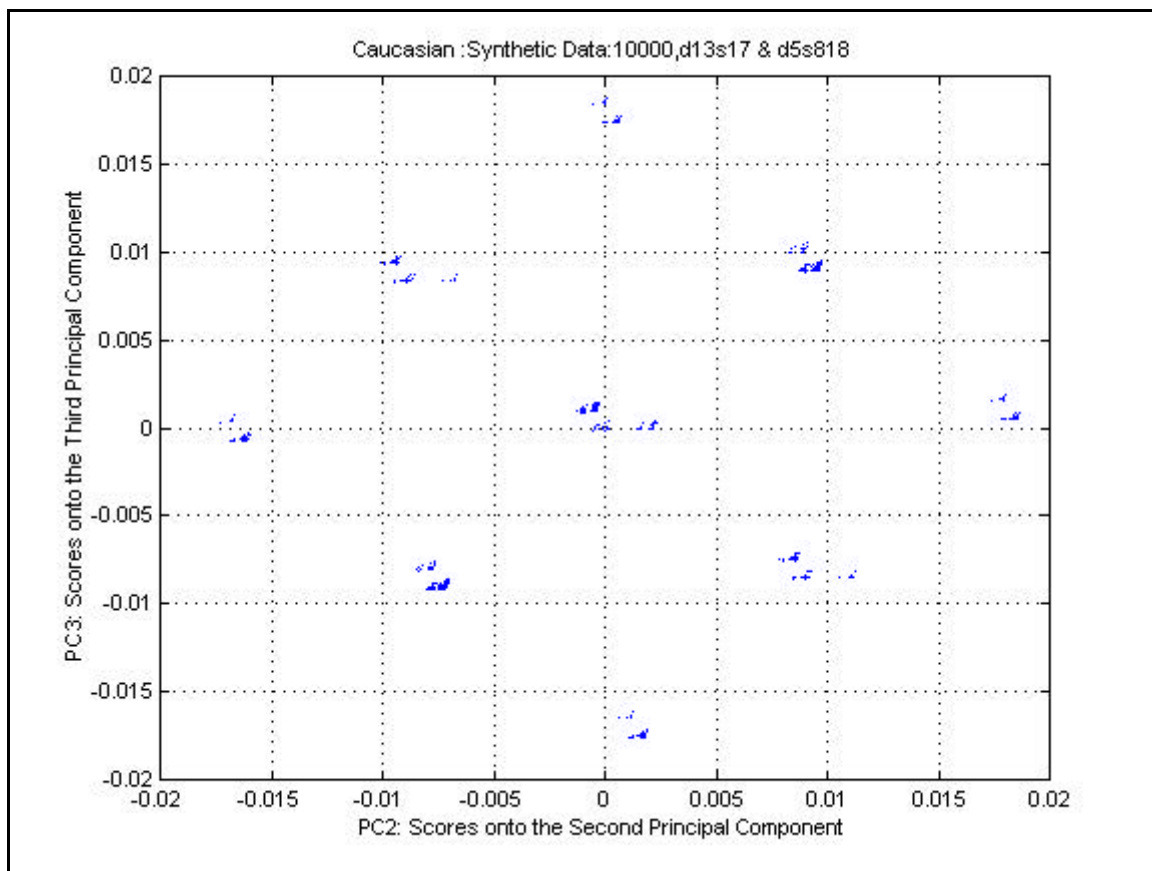


FIGURE 5.5: Scores of 10,000 Caucasian-DNA/STR Profiles onto the 2nd and 3rd Principal Component in d13s317 and d5s818 Loci.

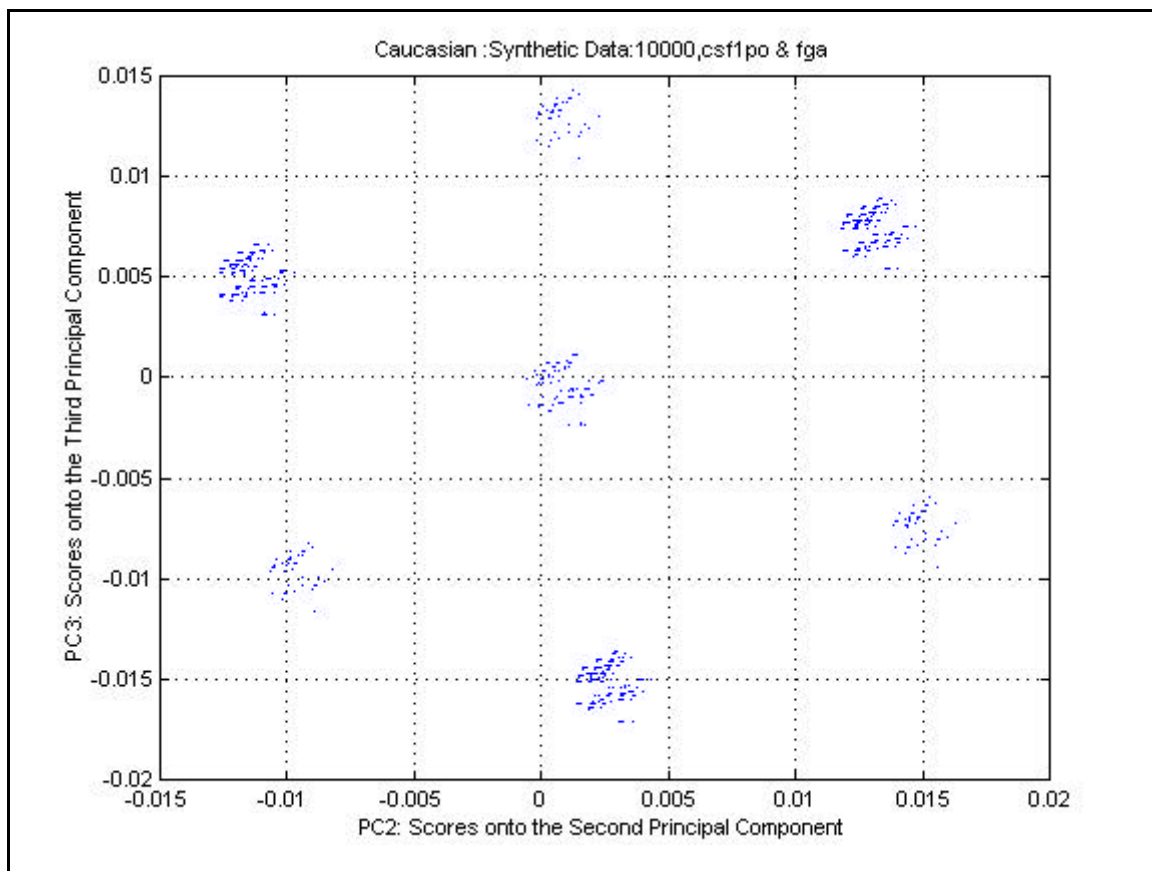


FIGURE 5.6: Scores of 10,000 Caucasian-DNA/STR Profiles onto the 2nd and 3rd Principal Components in csf1po and fga Loci.

5.1.3. STUDY OF THE PCA OF DNA/STR PROFILES WITH INFORMATION AT THE d13s317-d16s539 LOCI-PAIR THAT GIVE GOOD CLUSTERABILITY

The allele frequency distribution for all possible two-loci combinations was studied and analyzed to determine the reason behind the formation of good and distinct clusters. It was observed that loci pair that has allele frequency concentrated at just a few locations gives better clusters in comparison to a loci pair that has a more uniform allele probability distribution pattern. In particular, mostly good and distinct clusters were obtained for loci pair in which only two to three alleles are the major players; that is, in each locus, most allele occurrence probabilities are concentrated at just two to three alleles.

Now consider a particular case of two-loci combination, that of d13s317 and d16s539 for illustration. The relative frequencies of occurrence of the alleles at the d13s317 and d16s539 loci are shown in Tables 5.3 and 5.4 respectively. Figures 5.7 and 5.8 provide graphical representation of the relative allele frequencies of the above loci pair under study. It is observed that alleles 11 and 12 in both loci have a much higher probability of occurrence, as indicated by the shaded rows in Tables 5.3 and 5.4, and the corresponding peaks in Figures 5.7 and 5.8 respectively.

TABLE 5.3: Relative Frequency of Occurrence of Alleles at the d13s317 Locus for Caucasian Population.

Bin	Allele	Fraction
1	7	0.000
2	8	0.115
3	9	0.078
4	10	0.070
5	11	0.313
6	12	0.283
7	13	0.098
8	14	0.043
9	15	0.003
Total Probability $\longrightarrow$		1.000

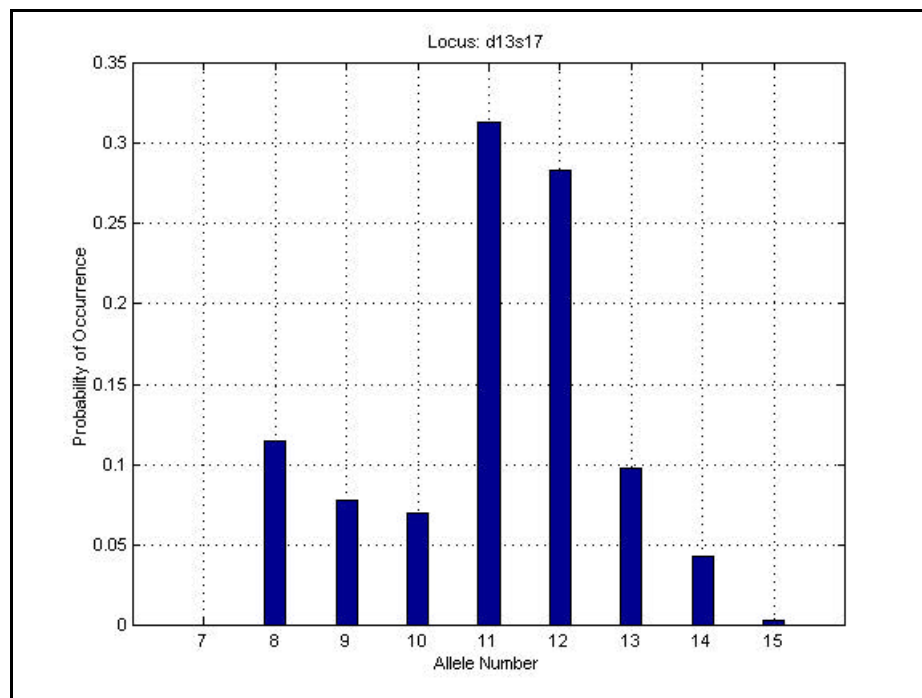


FIGURE 5.7: Allele Frequency Distribution of d13s317 Locus of Caucasian DNA.

TABLE 5.4: Relative Frequency of Occurrence of Alleles at the d16s539 Locus for Caucasian Population.

Bin	Allele	Fraction
1	5	0.000
2	6	0.000
3	7	0.000
4	8	0.026
5	9	0.107
6	10	0.079
7	11	0.319
8	12	0.269
9	13	0.167
10	14	0.031
11	15	0.002
Total Probability $\longrightarrow$		1.000

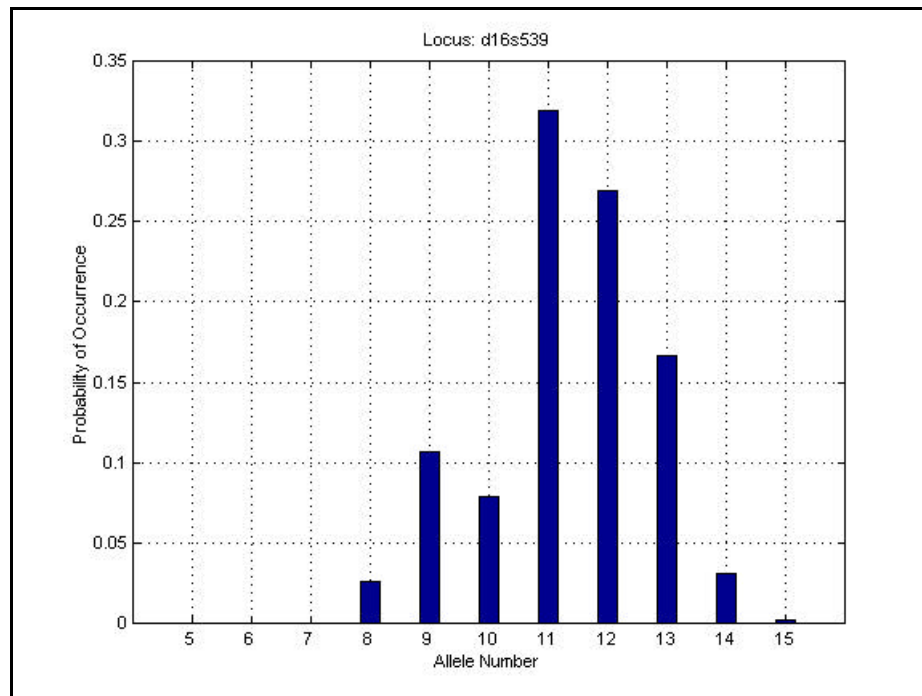


FIGURE 5.8: Allele Frequency Distribution of d16s539 Locus of Caucasian DNA.

Figures 5.9 and 5.10 show the joint probability density of two-allele frequency distribution for the d13s317 and d16s539 locus respectively. As expected, it is seen that only a few of the allele pairs have relatively high probability of occurring; corresponding to allele pairs (11,11), (11,12), (12,11) and (12,12) for both d13s317 and d16s539.

The joint probability distribution of two-allele pairs from two loci, one pair of alleles from d13s317 and the other from d16s539 is shown in Figure 5.11. It is observed that most probability densities are concentrated at a few pairs of two-allele pairs, corresponding to those alleles with relatively high probability of occurring, which are at the 11th and 12th alleles of each locus. With respect to alleles 11 and 12 of each of the two loci, the presence/absence pattern can be described with 1's and 0's as shown in Table 5.5. Since there are four dominant alleles, there will be 16 ($=2^4$) possible allele combinations that define the possible arrangements of the four dominant alleles in this loci-pair.

This distribution pattern is to be contrasted with one where the probability of alleles is more uniformly distributed, such as that of d18s51, as shown in Figure 5.12.

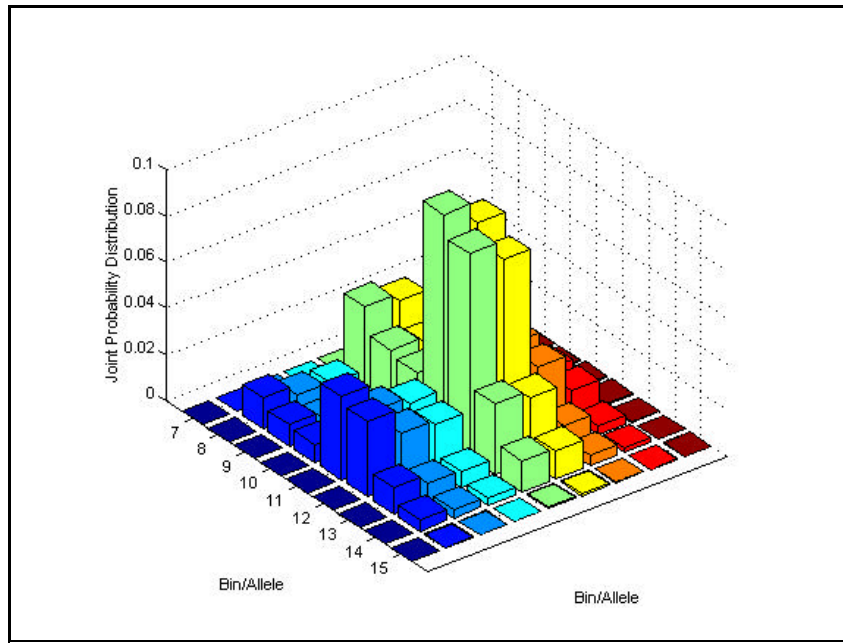


FIGURE 5.9: Joint Probability Density of Two-Allele Distribution of the d13s317 Locus.

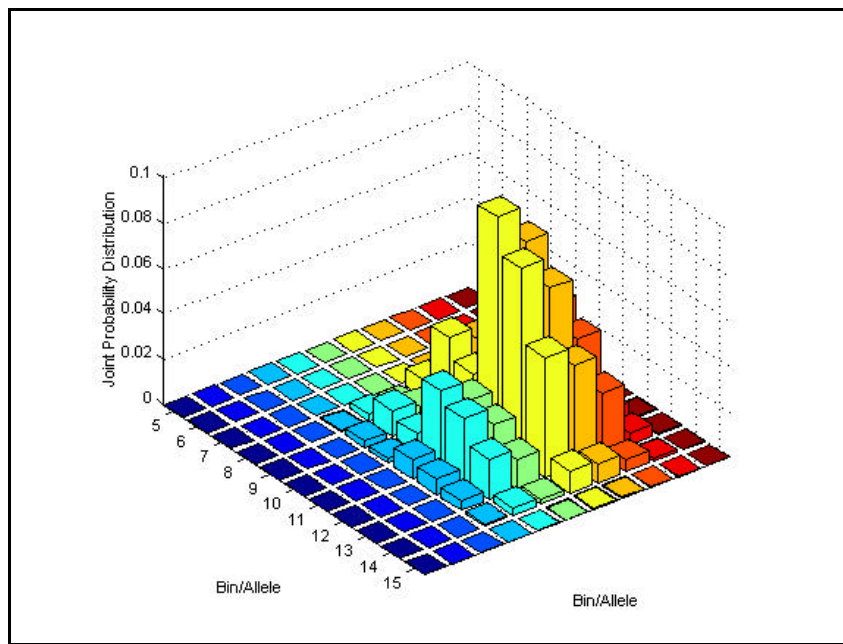


FIGURE 5.10: Joint Probability Density of Two-Allele Distribution of the d16s539 Locus.

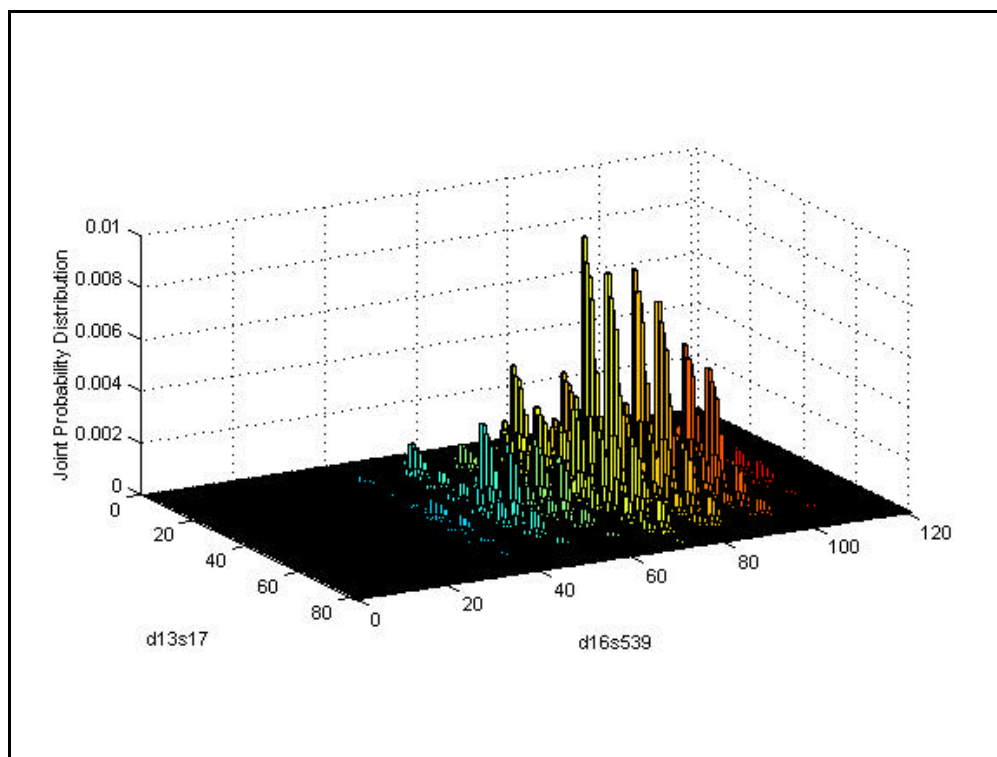


FIGURE 5.11: Joint Probability Distribution of Allele Pairs for d13s317 and d16s539.

TABLE 5.5: Possible Combinations of Presence/Absence of Allele 11 and 12 that Result in Clustering for Loci-Pair d13s317 and d16s539.

d13s317		d16s539	
Allele 11	Allele 12	Allele 11	Allele 12
Yes (1)	Yes (1)	Yes (1)	Yes (1)
Yes (1)	Yes (1)	Yes (1)	No (0)
Yes (1)	Yes (1)	No (0)	Yes (1)
Yes (1)	No (0)	Yes (1)	Yes (1)
No (0)	Yes (1)	Yes (1)	Yes (1)
Yes (1)	Yes (1)	No (0)	No (0)
Yes (1)	No (0)	Yes (1)	No (0)
Yes (1)	No (0)	No (0)	Yes (1)
No (0)	Yes (1)	Yes (1)	No (0)
No (0)	No (0)	Yes (1)	Yes (1)
No (0)	Yes (1)	No (0)	Yes (1)
Yes (1)	No (0)	No (0)	No (0)
No (0)	Yes (1)	No (0)	No (0)
No (0)	No (0)	Yes (1)	No (0)
No (0)	No (0)	No (0)	Yes (1)
No (0)	No (0)	No (0)	No (0)

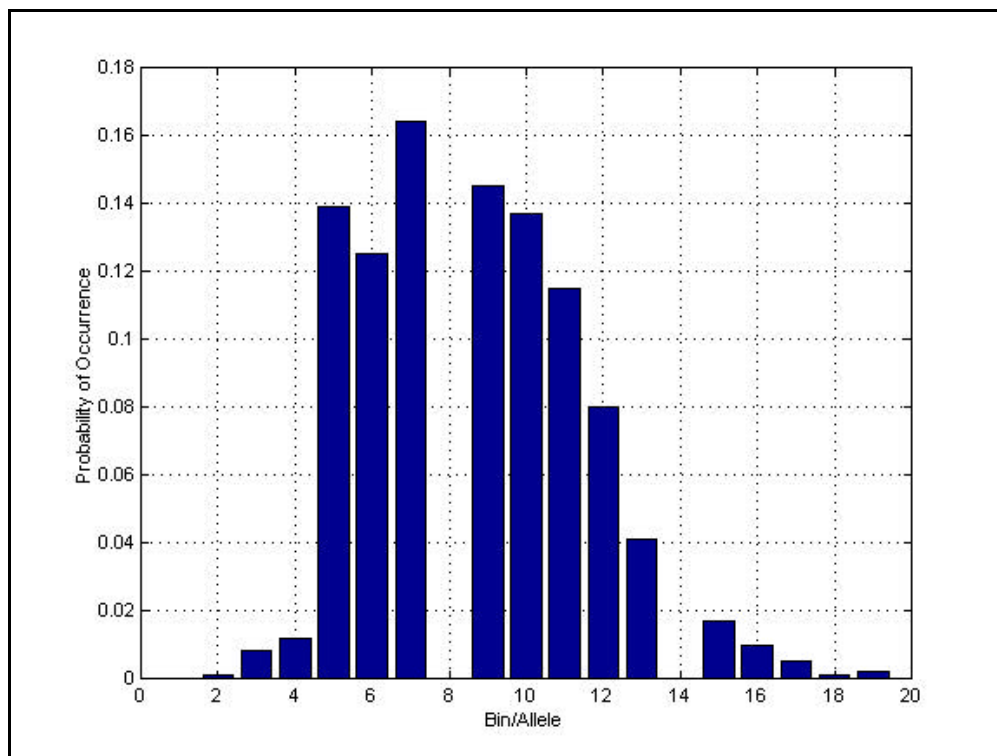


FIGURE 5.12: Allele Frequency Distribution of the d18s51 Locus of Caucasian DNA that Exhibit a More Diffused Probability Density Distribution Pattern and Yield Poor Clusterability in Pair with Another Locus.

It will be shown in Section 5.1.4 that, each of these allele presence/absence patterns corresponds to the formation of one of the nine clusters that are observed in Figure 5.1 for the loci pair of d13s317 and d16s539. Some of these allele patterns will map to the same cluster. The reason for this will be presented later.

Principal component analysis was carried by performing SVD on the binary synthetic data set of 10,000 DNA/STR profiles containing only DNA/STR information at the loci-pair of d13s317 and d16s539. The data submatrix is therefore of size of 10,000-by-20, composed of 1's and 0's. It is noted that the data matrix is not column mean centered first. This implies that the first principal component would only contain the average allele distribution information of all the profiles, instead of representing any unique and potentially differentiating information of the profiles. It is known that the SVD of a matrix decomposes a data into its mutually orthogonal components. The right singular vectors, represented by the V matrix in Equation 4.1 are referred to as the principal component vectors. Recall that the singular values are all positive and are arranged in a decreasing order. The first few principal component axes are the ones onto which the original data rows project the most, and therefore best represent the orientation and information contained in the original data rows. The relative variance carried by the l^{th} principal component vector is given by the square of the l^{th} singular value.

From the allele distribution pattern it is expected that the majority of profiles in the

data set contain alleles that have high probability of occurrence for each of the loci and are thus the most dominant for that particular locus. If the number of dominant alleles in each locus of the loci pair should be few (2 or 3) then the principal components would generate a high level of variance among profiles associated with this locus pair, i.e. either they have or they don't have the 2 or 3 dominant alleles, and would ultimately lead to the formation of good and distinct clusters. For example, considering the d13s317 and the d16s539 loci, alleles 11 and 12, (as shown in Tables 5.3 and 5.4) have high probability of occurrence in each of the two loci, and thus these dominant alleles would determine the clustering of profiles associated with loci pair d13s317-d16s539. The dominating alleles present in different DNA/STR profiles differentiate the assignment of profiles to a particular cluster depending upon the particular presence/absence pattern of each profile at these dominant alleles as will be discussed later in the following sections. The variance of profiles associated with a locus having only two or three dominant alleles, will be higher than for those where a large number of alleles having comparable but equally low probability densities. The locus that has a more uniform but low probability of occurrence of its alleles would not yield distinct clusters because the contribution of each of the original allele variables to the principal component would be comparable and the projection of profiles onto the first few principal components (the scores plot) would more or less be a continuum along the principal component axes, and no distinguishable cluster would result, as will be seen in later sections.

When carrying out principal component analysis, the importance of each principal

component in representing the original data matrix is of interest. A Scree plot was used to aid the analysis, as shown in Figure 5.13. A Scree plot shows the cumulative contribution made by successive principal components, obtained after performing the Singular Value Decomposition (SVD) of the original data matrix. The greater the variance among the profiles along the first few principal components, as exhibited by a steep Scree plot curve, the better it is for the formation of distinct clusters. The large variance is captured by the leading principal components obtained after SVD of the original binary data matrix. The right singular vectors are the principal component vectors.

An important point to note is that, since the data is not column mean centered, the first principal component just gives the average of the overall data, and therefore is not useful in differentiating between the clusters. The second principal component then captures the maximum variance exhibited by the original data matrix. Similarly, the third principal component gives the direction that captures the next maximum variance. The principal components are orthogonal to each other, thus contributing independent information. Table 5.6 shows the singular values obtained and the contribution made by the principal components that resulted in the Scree plot. It is seen from the Scree plot that, for example, that the first three principal components capture around 60% of the total variation exhibited by the original data matrix. Principal components 2 and 3 together captures about 30%

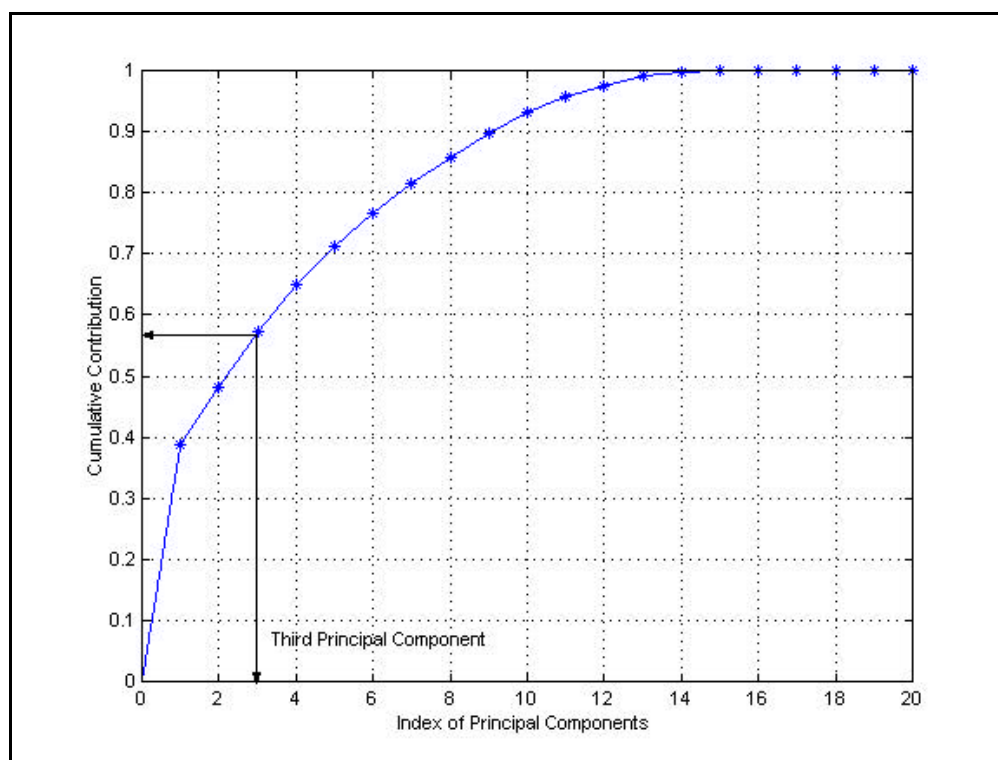


FIGURE 5.13: Scree Plot of the Cumulative Contribution of the Principal Components for 10,000 synthetic profiles for the d13s317 and d16s539 Loci-Pair.

TABLE 5.6: Cumulative Contribution of all Principal Components in the PCA Analysis of the Data Matrix.

Principal	Singular Values (S)	% Contribution	% Cumulative
1	117.6898	38.8731	38.8731
2	57.2250	9.1906	48.0637
3	57.2052	9.1843	57.2480
4	52.0428	7.6014	64.8494
5	47.8330	6.4213	71.2707
6	43.4887	5.3079	76.5786
7	41.8281	4.9103	81.4889
8	38.9614	4.2604	85.7493
9	37.3747	3.9203	89.6696
10	35.8754	3.6122	93.2818
11	29.5364	2.4484	95.7302
12	25.5160	1.8273	97.5575
13	23.2265	1.514	99.0715
14	15.3801	0.6639	99.7354
15	7.9138	0.1758	99.9112
16	5.6262	0.0888	100.0000
17	0.0000	No Contribution	
18	0.0000		
19	0.0000		
20	0.0000		

$\{(9.1906+9.1843)*100/(100-38.8731)\}$ of the unique information contained in the original data matrix. It is also observed that each successive component contributes less and less to the overall data variation. Through trial and error and experimentation, it was found that in general two-dimensional scores plot (projection onto just two principal components) is sufficient to effect good clustering of the profiles.

In the absence of column mean centering, the first few leading principal components after the first are the ones that reflect unique information of the individual profiles. They should be weighted by the dominant alleles that have high allele probability densities. Not surprisingly this is so, as shown in Figures 5.14 and 5.15, which show the make-up of the second and third principal components of the 10,000 synthetic profile data matrix at the d13s317 and the d16s539 loci. Table 5.7 gives the relative contribution of the various alleles in forming the second and third principal components for our loci pair of d13s317 and d16s539. As predicted, the dominant alleles contribute heavily in the definition of the second and third principal components. From Figure 5.14, one can predict from the signs of the contribution by the four dominant alleles that those DNA profiles which do not have allele 11 (Bin 5: - sign) but do have allele 12 (Bin 6: + sign) at the d13s317 locus, and do have allele 11 (+ sign), but do not have allele 12 (- sign) (the 16th and 17th Bin) at the d16s539 locus will score very high on this principal component. The shaded rows in Table 5.7 indicate the alleles that play a major role in the formation of clusters as discussed earlier.

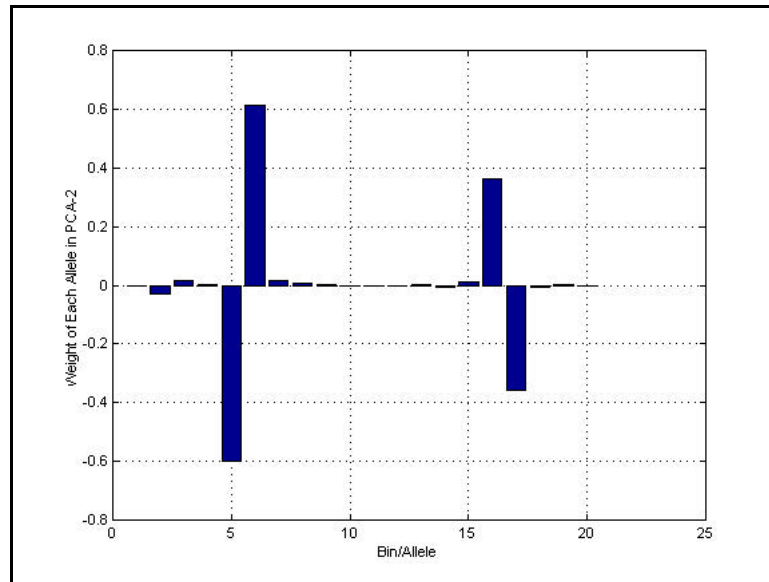


FIGURE 5.14: Make up of the 2nd Principal Component.

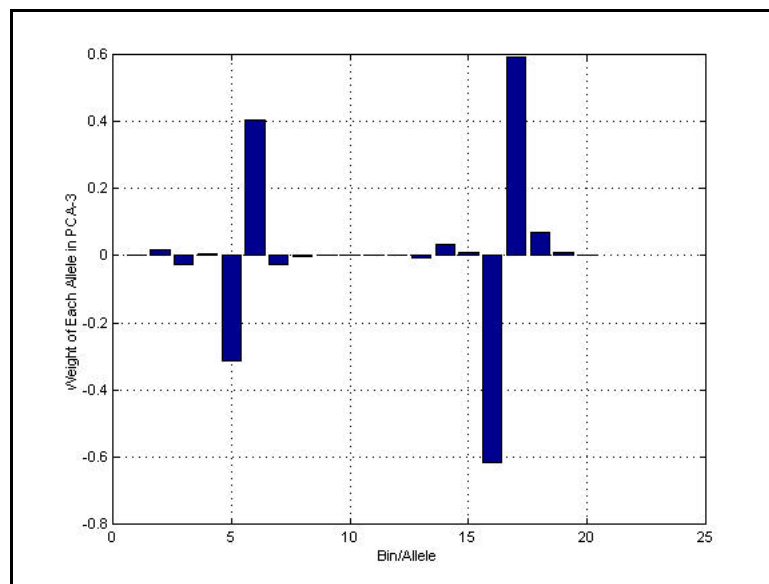


FIGURE 5.15: Make up of the 3rd Principal Component.

TABLE 5.7: The 2nd and 3rd Principal Components of the PCA of 10,000 Synthetic Caucasian DNA Profiles with Alleles in d13s317 and d16s539.

Bin	Locus	Allele	PC-2 (=V2)	PC-3 (=V3)
1	d13s317	7	0.0000	0.0000
2	d13s317	8	-0.0305	0.0157
3	d13s317	9	0.0143	-0.0261
4	d13s317	10	0.0027	0.0063
5	d13s317	11	-0.6016	-0.3124
6	d13s317	12	0.6142	0.4053
7	d13s317	13	0.0151	-0.0271
8	d13s317	14	0.0082	-0.0029
9	d13s317	15	0.0004	0.0007
10	d16s539	5	-0.0000	0.0000
11	d16s539	6	0.0000	-0.0000
12	d16s539	7	-0.0000	0.0000
13	d16s539	8	0.0002	-0.0053
14	d16s539	9	-0.0090	0.0351
15	d16s539	10	0.0101	0.0096
16	d16s539	11	0.3639	-0.6157
17	d16s539	12	-0.3560	0.5926
18	d16s539	13	-0.0064	0.0686
19	d16s539	14	0.0025	0.0100
20	d16s539	15	-0.0010	0.0005

From Table 5.7, it is seen that alleles 11 and 12 of both loci d13s317 and d16s539 are the dominating alleles in both the second and third principal components. These are the alleles that have high probability of occurrence and are the ones that would distinguish profiles into different clusters. The contributions of other alleles to these two principal components are negligible. An important aspect is the sign (+/-) associated with the dominant allele entry in the second and third principal components. It is seen that, the dominating alleles in both loci have opposite signs within each principal component for both principal components. The opposite signs for the two dominant alleles of each locus for PC2 imply the absence/presence pattern of these two alleles within each locus among the profiles that would score high on this principal component. For example, if there is a cluster of the scores of DNA/STR profiles that project highly onto the positive direction of the third principal component, then one can deduce that all members within this cluster will have allele 12(+) (Bin 6) of d13s317 as well as allele 12(+) (Bin 17) of d16s539, and they will not have allele 11(-) (Bin 5) of d13s317 nor allele 11(-) (Bin 16) of d16s539. This can be seen from the make-up (Note the opposite signs involved) of the third principal component shown in Figure 5.15. The presence/absence allele pattern for this cluster case is thus (0 1 0 1) (cluster 7 in Figure 5.16) and is one of the possible sixteen combinations (Table 5.5) of allele patterns at the four dominant alleles that will result in the formation of clusters, and will project highly in the positive direction of principal component three. In other words, profiles that have the allele pattern of (0 1 0 1) at

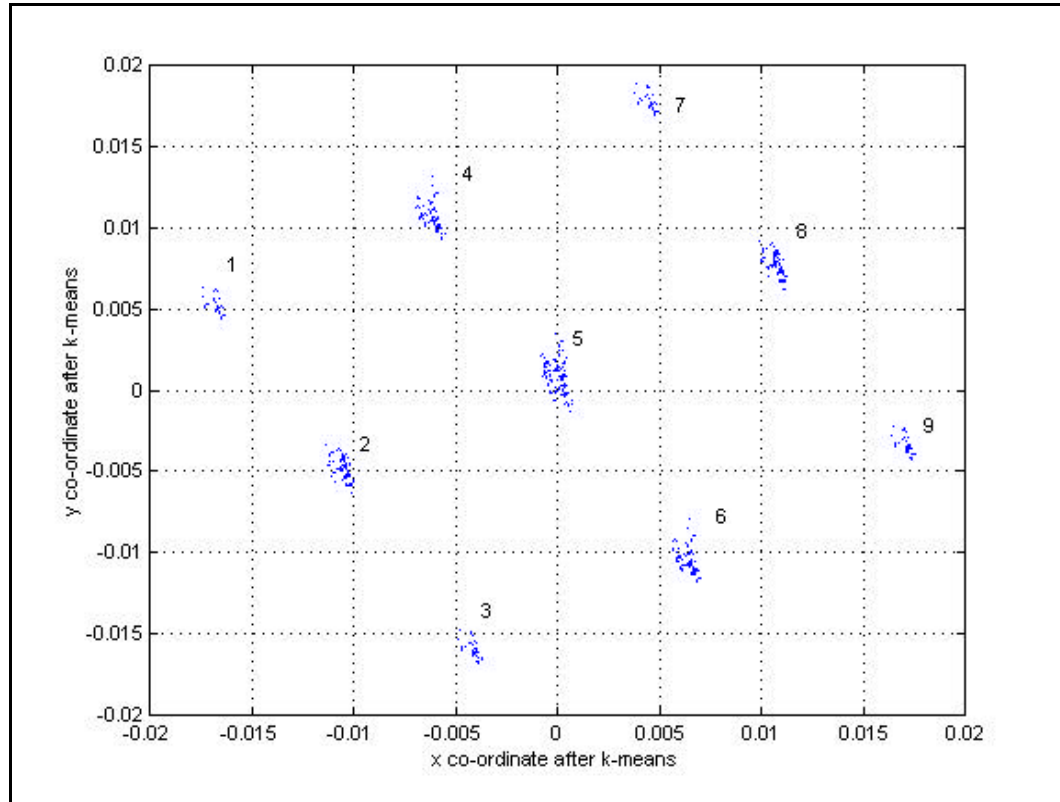


FIGURE 5.16: Using K-Means Algorithm to Identify Clusters of the Scores Projected onto the Second and Third Principal Components of d13s317 and d16s539 Loci Pair for 10,000 Synthetically Generated Sample Profiles.

these four dominant sites would score positively high along this principal component; and those with allele pattern (1 0 1 0) would score negatively high along this principal component, e.g. cluster 3. It will be shown how the allele distribution pattern at these four dominant sites lead to the differentiation of clusters. The profile membership of each cluster is validated by the k-means algorithm, a popular clustering method, to support the arguments presented above in regards to the relationship between allele distribution pattern and cluster assignment.

5.1.4. CLUSTER MEMBERSHIP VALIDATION USING K-MEANS ALGORITHM FOR PROFILES WITH DNA/STR INFORMATION AT THE d13s317-d16s539 LOCI-PAIR

K-means algorithm was used to validate the pattern of partitions of the 10,000 sample profiles into the nine clusters seen earlier from the scores plot (Figure 5.1) after principal component analysis. Figure 5.16 shows the clusters formed by the k-means approach. Assignments of the profiles into the nine clusters given by the principal component analysis and the k-means approach have been compared and are found to be identical. The number of samples was obtained and identified in each cluster. This provides a better understanding of the role the dominant alleles play in the formation of these nine distinct clusters. Table 5.8 gives the size distribution of the nine distinct clusters for our synthetically generated data of 10,000 profiles. Note that the clusters are nearly equal in size. This is good in that when these clusters are placed in a

TABLE 5.8: Number/Percentage of Samples in Each Cluster of the 10,000 Synthetically Generated Profiles for d13s317 and d16s539 Loci-Pair.

Cluster Number	Number of Profiles	Percentage of Profiles
1	1046	10.46
2	1216	12.16
3	1231	12.31
4	1025	10.25
5	1143	11.43
6	1265	12.65
7	918	9.18
8	1077	10.77
9	1079	10.79
Total →	10000	100

decision tree during search, nearly equal partition size leads to a more balanced tree, yielding a faster and more efficient search.

Now, the distribution of alleles among the member profiles within each cluster is explored in order to further elucidate the reason for cluster formation. Figures 5.17-5.25 represent the percent of profiles within each cluster that has each of the alleles indexed along the x-axis, for clusters one to nine respectively. Figure 5.17 shows, for instance, that all members in cluster 1 have allele 11 (Bin 5) of d13s317 and allele 12 (Bin 17) of d16s539, but all do not have allele 12 (Bin 6) and allele 11 (Bin 16) of the two respective loci. The allele distribution pattern at these four dominant sites is thus (1 0 0 1). Likewise, for clusters 3, 7 and 9 the allele patterns at these four sites are (1 0 1 0), (0 1 0 1) and (0 1 1 0) respectively. An interesting situation occurs for clusters 2, 4, 5, 6 and 8, where more than one combination of allele pattern possible for that particular cluster are observed. For cluster 2, the allele distribution pattern could be either (1 0 0 0) or (1 0 1 1). The reason for having two possible combinations is that the presence or the absence of both allele 11 and 12 at locus d16s539 lead to the same sub-score in the overall score in member profiles that constitute cluster 2. Similar interpretations can be rendered for clusters 4, 5, 6 and 8. Table 5.9 shows the combinations of these four dominant alleles for each of the nine clusters, based on the plots shown in Figures 5.17-5.25, above.

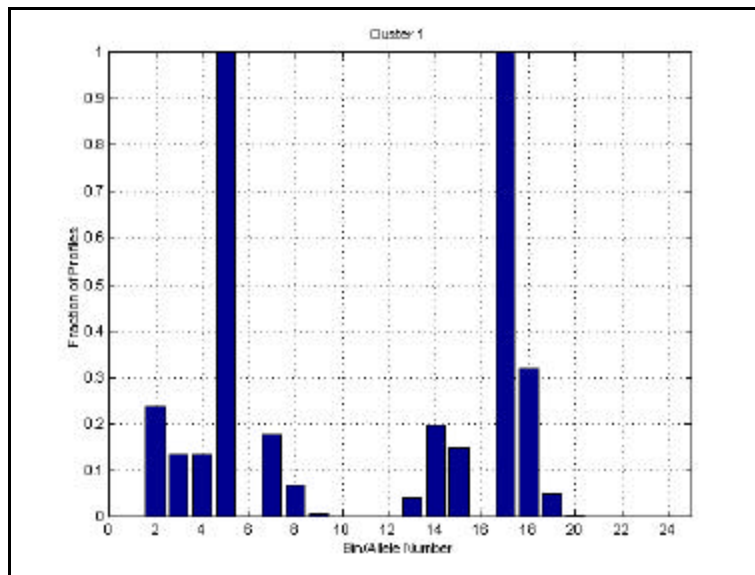


FIGURE 5.17: Percent of Profile Samples in Cluster 1 that has Each of the Alleles in d13s317 and d16s539 Loci.

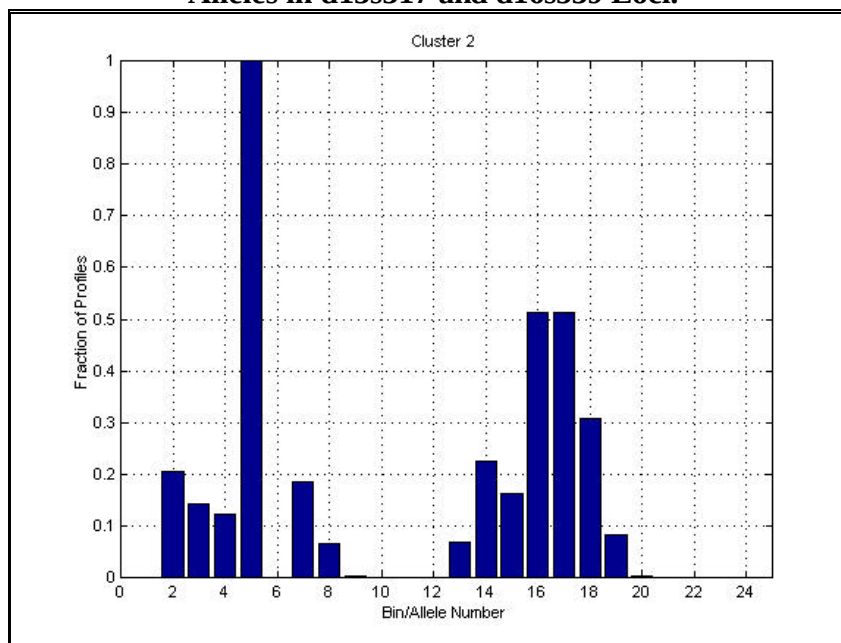


FIGURE 5.18: Percent of Profile Samples in Cluster 2 that has Each of the Alleles in d13s317 and d16s539 Loci.

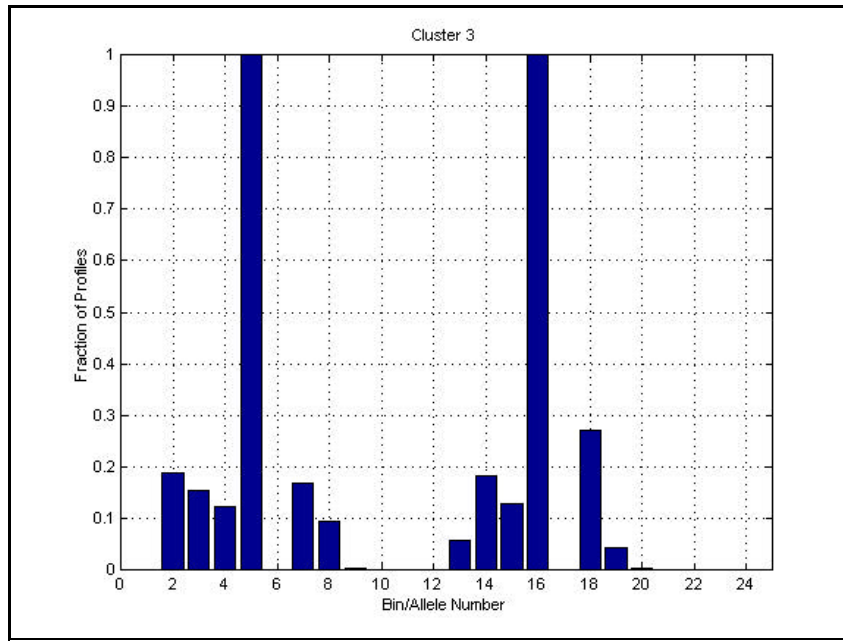


FIGURE 5.19: Percent of Profile Samples in Cluster 3 that has Each of the Alleles in d13s317 and d16s539 Loci.

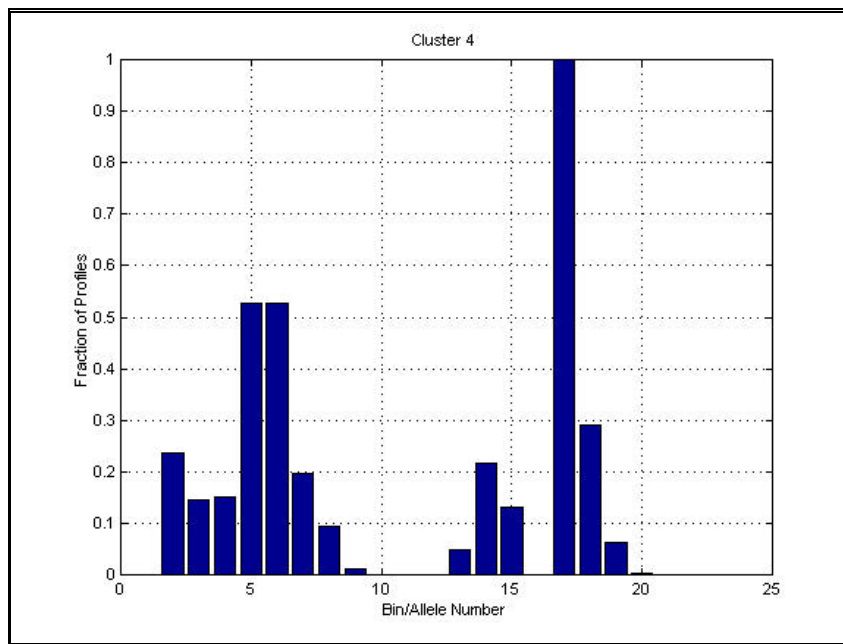


FIGURE 5.20: Percent of Profile Samples in Cluster 4 that has Each of the Alleles in d13s317 and d16s539 Loci.

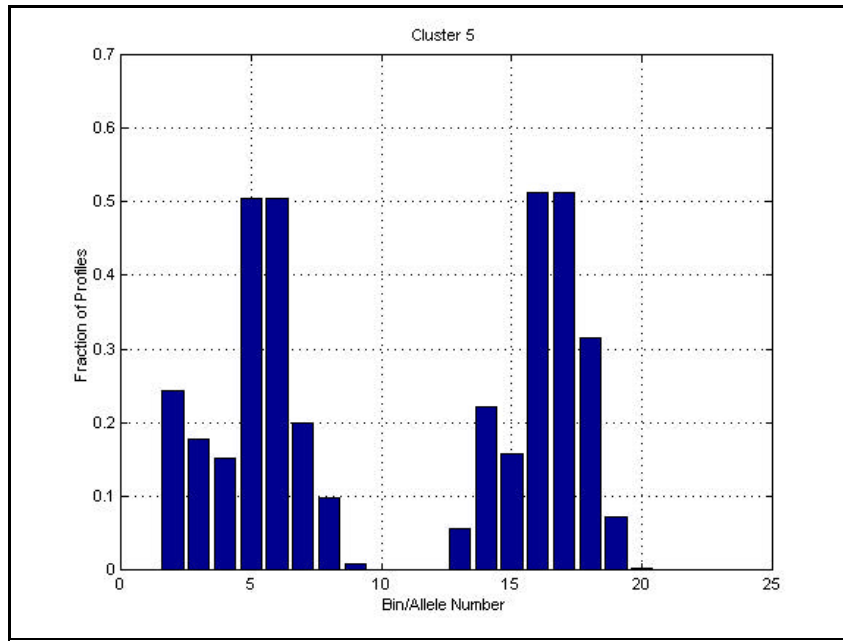


FIGURE 5.21: Percent of Profile Samples in Cluster 5 that has Each of the Alleles in d13s317 and d16s539 Loci.

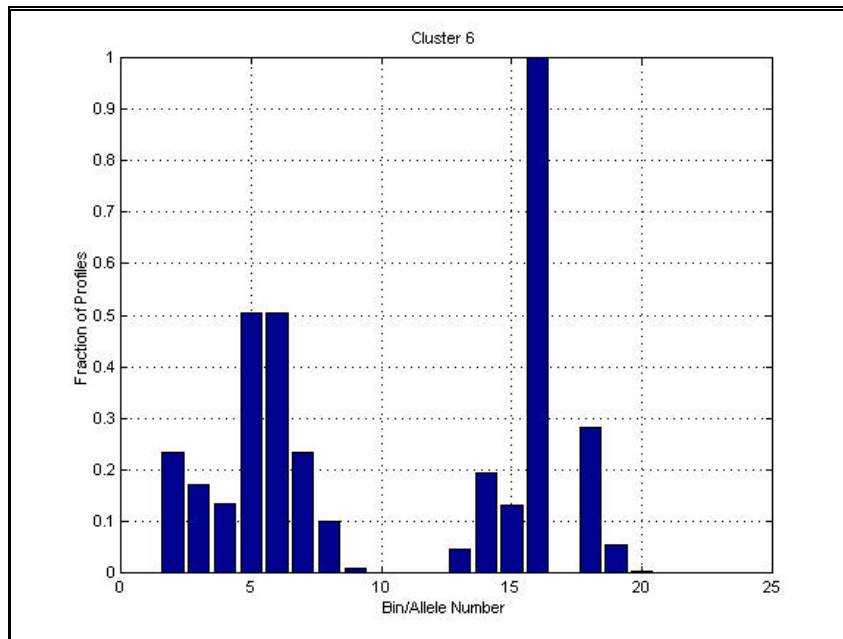


FIGURE 5.22: Percent of Profile Samples in Cluster 6 that has Each of the Alleles in d13s317 and d16s539 Loci.

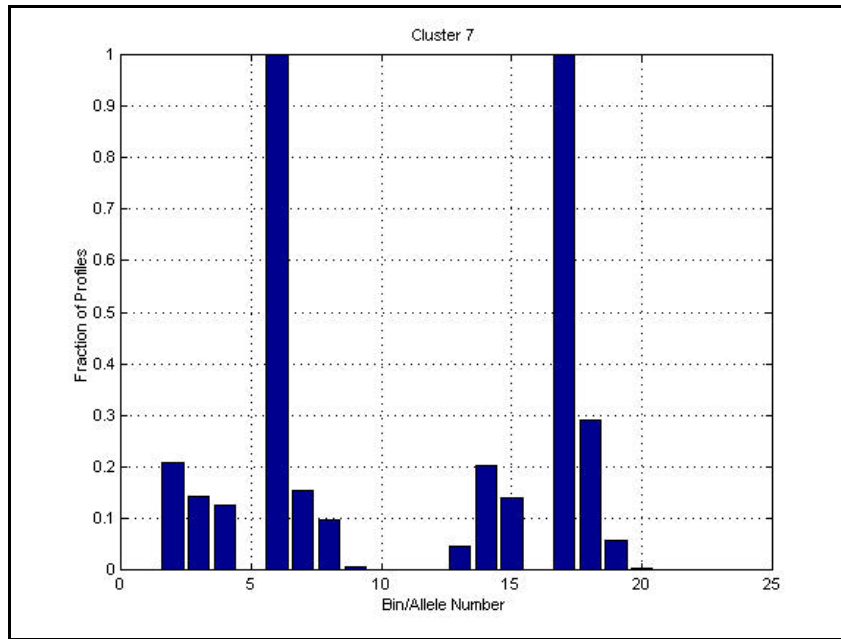


FIGURE 5.23: Percent of Profile Samples in Cluster 7 that has Each of the Alleles in d13s317 and d16s539 Loci.

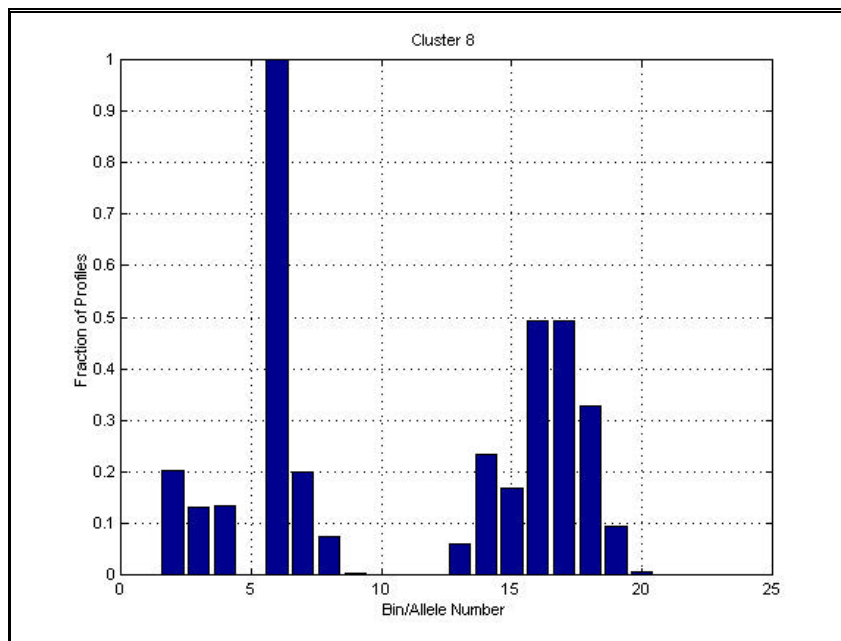


FIGURE 5.24: Percent of Profile Samples in Cluster 8 that has Each of the Alleles in d13s317 and d16s539 Loci.

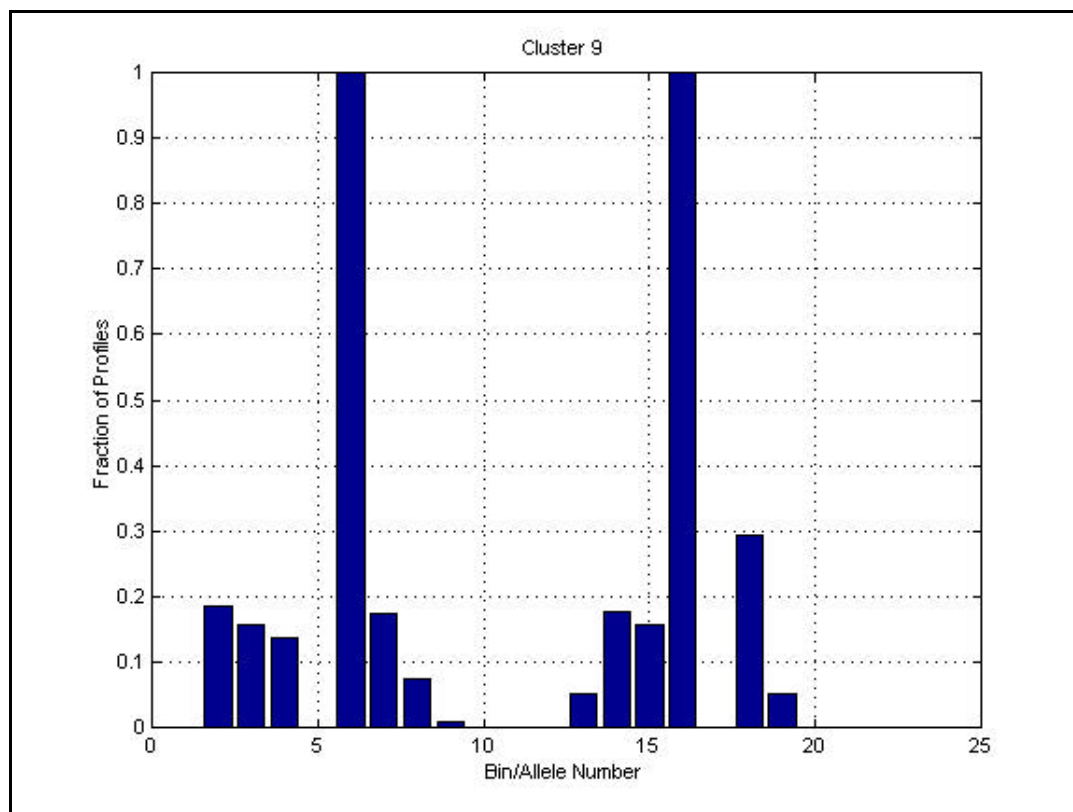


FIGURE 5.25: Percent of Profile Samples in Cluster 9 that has Each of the Alleles in d13s317 and d16s539 Loci.

TABLE 5.9: The Presence (1) or Absence (0) of Alleles 11 and 12 in the d13s317 and d16s539 Loci for each Scores Clusters as shown in Figure 5.1.

Cluster Assignment	d13s317		d16s539	
	Allele 11	Allele 12	Allele 11	Allele 12
	Bin (5)	Bin (6)	Bin (16)	Bin (17)
1	Yes (1)	No (0)	No (0)	Yes (1)
2	Yes (1)	No (0)	No (0)	No (0)
2	Yes (1)	No (0)	Yes (1)	Yes (1)
3	Yes (1)	No (0)	Yes (1)	No (0)
4	Yes (1)	Yes (1)	No (0)	Yes (1)
4	No (0)	No (0)	No (0)	Yes (1)
5	Yes (1)	Yes (1)	Yes (1)	Yes (1)
5	No (0)	No (0)	Yes (1)	Yes (1)
5	Yes (1)	Yes (1)	No (0)	No (0)
5	No (0)	No (0)	No (0)	No (0)
6	Yes (1)	Yes (1)	Yes (1)	No (0)
6	No (0)	No (0)	Yes (1)	No (0)
7	No (0)	Yes (1)	No (0)	Yes (1)
8	No (0)	Yes (1)	Yes (1)	Yes (1)
8	No (0)	Yes (1)	No (0)	No (0)
9	No (0)	Yes (1)	Yes (1)	No (0)

To support the rationale for identifying the allele distribution characteristics of the clusters, allele distribution patterns of twenty randomly selected member profiles from clusters 1-9 respectively are shown in Table 5.10-5.18. From Figure 5.17, it is seen that allele 11 (Bin 5) of locus d13s317, and alleles 12 (Bin 17) of locus d16s539 are present in all the members of cluster 1. Table 5.10 supports this fact, where it is observed that all members in cluster 1 indeed have allele 11 of d13s317 as well as allele 12 of d16s539, giving the allele distribution pattern of (1 0 0 1). Similar observation is made for cluster 2, (Figure 5.18) where all members belonging to this cluster should have allele 11 of d13s317 and some have either both or none of alleles 11 and 12 of d16s539. The breakdown is nearly half for the cluster. This distribution pattern can be seen in Table 5.11. Thus the allele distribution pattern is either (1 0 1 1) or (1 0 0 0). Table 5.12 gives the distribution pattern of the four dominant alleles for cluster 3 where the pattern is (1 0 1 0). Similar interpretations can be extended to the allele distribution patterns at the four dominant sites to other clusters.

The position of each of the clusters in the 2-D scores plot in association with the respective allele distribution pattern will now be discussed. Cluster 9 and 3 will be used as examples for illustration. It was observed from the scores plot of Figure 5.1 that cluster 9 projects highly along the positive direction of the second principal component.

TABLE 5.10: Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 1 for d13s317-16s539 Loci Pair. The Allele Distribution Pattern at the Dominant Sites is (1 0 0 1).

Profile ID	Bin/Alleles for d13s317									Bin/Alleles for d16s539										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	7	8	9	10	11	12	13	14	15	5	6	7	8	9	10	11	12	13	14	15
6	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
19	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
22	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
30	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
52	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
54	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
74	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0
76	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
79	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
105	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
106	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
125	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
133	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
146	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
151	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1	1	0	0
155	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
161	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
174	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
181	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0
186	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0

TABLE 5.11: Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 2 for d13s317-16s539 Loci Pair. The Allele Distribution at the Four Dominant Sites is either (1 0 1 1) or (1 0 1 1).

Profile ID	Bin/Alleles for d13s317									Bin/Alleles for d16s539										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	7	8	9	10	11	12	13	14	15	5	6	7	8	9	10	11	12	13	14	15
10	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
45	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
47	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
48	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
51	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
55	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0
104	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
110	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
119	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
121	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
126	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0
131	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
134	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0
140	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
150	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
153	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0
158	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
177	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
184	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
187	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0

TABLE 5.12: Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 3 for d13s317-16s539 Loci Pair. The Allele Distribution at the Four Dominant Sites is (1 0 1 0).

Profile ID	Bin/Alleles for d13s317									Bin/Alleles for d16s539										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	7	8	9	10	11	12	13	14	15	5	6	7	8	9	10	11	12	13	14	15
1	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
3	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
13	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0
23	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
37	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
38	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
40	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
46	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
50	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0
56	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0
62	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
68	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
73	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
75	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0
82	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
97	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
107	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0
109	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
123	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
145	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0

TABLE 5.13: Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 4 for d13s317-16s539 Loci Pair. The Allele Distribution at the Four Dominant Sites is either (1 1 1 0) or (0 0 0 1).

Profile ID	Bin/Alleles for d13s317									Bin/Alleles for d16s539										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	7	8	9	10	11	12	13	14	15	5	6	7	8	9	10	11	12	13	14	15
24	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0
92	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0
155	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0
124	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
31	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0
162	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0
163	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0
8	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
42	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
153	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
159	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
152	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
193	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
120	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
166	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
21	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0
118	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
99	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
194	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0
171	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0

TABLE 5.14: Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 5 for d13s317-16s539 Loci Pair. The Allele Distribution at the Four Dominant Sites is either (1 1 1 1), (0 0 1 1), (1 1 0 0) or (0 0 0 0).

Profile ID	Bin/Alleles for d13s317									Bin/Alleles for d16s539										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	7	8	9	10	11	12	13	14	15	5	6	7	8	9	10	11	12	13	14	15
40	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
52	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0
20	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0
160	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
8	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
42	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
105	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
130	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0
100	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
146	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
186	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
124	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
114	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
147	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0
164	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
136	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
91	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
7	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0
21	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
94	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0

TABLE 5.15: Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 6 for d13s317-16s539 Loci Pair. The Allele Distribution at the Four Dominant Sites is either (1 1 1 0) or (0 0 1 0).

Profile ID	Bin/Alleles for d13s317									Bin/Alleles for d16s539										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	7	8	9	10	11	12	13	14	15	5	6	7	8	9	10	11	12	13	14	15
134	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
126	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
37	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
9	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	0	0
101	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0
105	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0
191	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
193	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
167	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
80	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
51	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
72	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0
89	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
86	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
133	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
70	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
141	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
132	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0
63	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
168	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0

TABLE 5.16: Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 7 for d13s317-16s539 Loci Pair. The Allele Distribution at the Four Dominant Sites is (0 1 0 1).

Profile ID	Bin/Alleles for d13s317									Bin/Alleles for d16s539										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	7	8	9	10	11	12	13	14	15	5	6	7	8	9	10	11	12	13	14	15
126	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
57	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
128	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
171	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0
33	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
89	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0
196	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
113	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
133	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
80	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0
22	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0
28	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0
65	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0
118	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
25	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0
85	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0
61	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
164	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0
125	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
24	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0

TABLE 5.17: Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 8 for d13s317-16s539 Loci Pair. The Allele Distribution at the Four Dominant Sites is either (0 1 1 1) or (0 1 0 0).

Profile ID	Bin/Alleles for d13s317									Bin/Alleles for d16s539										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	7	8	9	10	11	12	13	14	15	5	6	7	8	9	10	11	12	13	14	15
23	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0
144	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0
70	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
26	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
124	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
187	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0
161	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
111	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
78	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0
15	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0
99	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0
45	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
84	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0
194	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
104	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0
74	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0
142	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1	0	0
189	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
3	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0
81	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0

TABLE 5.18: Allele Distribution Pattern of a Few Randomly Selected Members from Cluster 9 for d13s317-16s539 Loci Pair. The Allele Distribution at the Four Dominant Sites is (0 1 1 0).

Profile ID	Bin/Alleles for d13s317									Bin/Alleles for d16s539										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	7	8	9	10	11	12	13	14	15	5	6	7	8	9	10	11	12	13	14	15
24	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0
72	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	0	0	0
19	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	1	0	0	0	0
176	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
81	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
142	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0
156	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
84	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
159	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0
124	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0
95	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
101	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0
31	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0
115	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
35	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0
163	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
137	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
36	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	1	0	0	0	0
1	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0
76	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0

From Figure 5.14, one sees that the make-up of the second principal component reflects the allele distribution pattern of (0 1 1 0) at the four dominant sites. A '0' is present at the first and fourth dominant alleles and the sign associated with their contribution to this principal component is negative, indicating the absence of; whereas the sign associated with the second and third dominant alleles (represented by '1') is positive, indicating the presence of. Thus, profiles exhibiting this allele pattern should project highly, positively, onto this principal component. Figure 5.25 supports this argument, in which it is evident that in cluster 9, all members have the distribution pattern of (0 1 1 0) at the four dominant allele sites.

From Figure 5.1, it is also observed that cluster 3 projects highly along the negative direction of the third principal component. If one looks at the make-up of the third principal component, as shown in Figure 5.15, one sees that a profile not possessing allele 11 (negative make-up), but possessing allele 12 (positive make-up) of d13s317, not possessing allele 11 but possessing 12 of d16s539, (0 1 0 1), would score high positively along this principal component. In order to score high negatively, as in cluster 3, the distribution would have to be the opposite, or that of (1 0 1 0), the presence of allele 11 at both loci and the absence of allele 12 at both loci. This pattern is indeed exhibited by all the members of cluster 3 (Table 5.9).

Further, Figure 5.19 shows that cluster 5 projects almost to the origin of the 2-D scores plot. From Figure 5.21, it is found that members in this cluster have either both alleles

of the two loci, that is the (1 1 1 1) pattern or neither allele for both loci giving the (0 0 0 0) pattern at the dominant allele locations, so that the effects of multiplying them by the elements of the principal components for each locus cancel as will be explained below. In addition, members in cluster 5 may also have allele distribution patterns of (0 0 1 1) and (1 1 0 0) (See Table 5.9). The reason that all forms of these allele distribution patterns map into cluster 5 is explained below.

Recall that Section 4.1.2.1 shows how the scores of a profile onto a principal component axes is calculated. From Table 5.7, the contribution to the principal components made by each of the dominant alleles of the two loci is known. They are represented by the $\hat{V}$ matrix below. These are numbers extracted from the highlighted positions of the original full V matrix shown in Table 5.7. The $\hat{X}$ matrix represents the submatrix (of the larger binary DNA/STR matrix) composed only of allele information at the four dominant allele sites, and representing the four degenerate allele distribution patterns which all map to cluster 5.

$$\hat{X} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \hat{V} = \begin{pmatrix} -0.6016 & -0.3124 \\ +0.6142 & +0.4053 \\ +0.3639 & -0.6157 \\ -0.3560 & +0.5926 \end{pmatrix}$$

$$\hat{\Sigma}^{-1} = \begin{pmatrix} 0.01747 & 0 \\ 0 & 0.01748 \end{pmatrix}$$

Therefore the Mahalanobis score is obtained as shown below.

$$\begin{aligned}
 \text{Mahalanobis Score} &= \hat{X} * \hat{V} * \hat{\Sigma}^{-1} \\
 &= \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} * \begin{pmatrix} -0.6016 & -0.3124 \\ +0.6142 & +0.4053 \\ +0.3639 & -0.6157 \\ -0.3560 & +0.5926 \end{pmatrix} * \begin{pmatrix} 0.01747 & 0 \\ 0 & 0.01748 \end{pmatrix} \\
 &= \begin{pmatrix} 0.0205 & 0.0698 \\ 0.0126 & 0.0929 \\ 0.0079 & -0.0231 \\ 0.0000 & 0.0000 \end{pmatrix} * \begin{pmatrix} 0.01747 & 0 \\ 0 & 0.01748 \end{pmatrix} \\
 &= \begin{pmatrix} 0.000358 & 0.00122 \\ 0.000220 & 0.00162 \\ 0.000138 & -0.00040 \\ 0 & 0 \end{pmatrix} \tag{5.1}
 \end{aligned}$$

It is seen that, because of the comparable magnitudes but of opposing signs of each of the pairs of the allele weights in each column of the $\hat{V}$ matrix a '11' entry in the row of the $\hat{X}$ matrix for either pair will tend to cancel their contribution to the final score onto that corresponding $\hat{V}$ vector. Therefore, all the four allele distribution patterns lead to relatively low scores along the second and third principal components. As a result, all four map to points near the origin in the scores plot. Similar interpretation can be extended to other clusters to correlate the position occupied by a cluster on the 2-D scores plot to its allele distribution patterns at the four dominant sites.

5.1.4.1. TESTING THE THEORY OF CLUSTER FORMATION BY PREDICTING THE APPROXIMATE CENTERS OF THE CLUSTERS

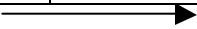
In the preceding sections, it was discussed how the relative locations of the clusters are determined by the allele distribution pattern at the four dominant alleles. Table 5.19 shows the coordinates of the true centers of each of the nine clusters, and Table 5.20 shows the approximate centers of the nine clusters calculated from just the set of all sixteen allele distribution patterns at the four dominant allele sites, following the procedures of Equation 5.1. The 1's and 0's at the four dominant sites of a profile in X (a row in X), when multiplied by the corresponding entries in the V vector will lead to a magnitude of the scores for that profile along the corresponding principal component vector. Presence of alleles at other sites is negligible in their contribution to the final scores along the principal components. Thus, for each suitably chosen loci pair, the approximate centers of the distinct clusters can be calculated from just the allele pattern at the dominant sites. For comparison, the true and approximate cluster centers are plotted as '*' and '•' respectively in Figure 5.26.

From the plot one observes that there are sixteen possible allele distribution patterns to calculate from the predicted cluster centers but they map into only nine cluster centers. The reason is that the manifestations of some patterns are similar, when multiplied by the corresponding entries in the V vectors. It is also observed that the true and the predicted cluster centers are indeed very close to each other, as predicted. The fine differences are due to the relative small contributions made by the

TABLE 5.19: The Coordinates of the True Centers of the Nine Clusters shown in Figure 5.1.

Cluster Number	Coordinates of Cluster Centers	
	X	Y
1	-0.0168	0.0054
2	-0.0062	0.0116
3	0.0045	0.0179
4	-0.0105	-0.0052
5	0.0001	0.0011
6	0.0108	0.0074
7	-0.0042	-0.0159
8	0.0064	-0.0096
9	0.0171	-0.0033

TABLE 5.20: Calculation of the Approximate Cluster Center Coordinates from the Sixteen Possible Allele Distribution Patterns at the Four Dominant Sites.

Cluster Assignment	Allele Combinations	d13s317		d16s539		Coordinates of Approximate Cluster Centers	
		Allele 11	Allele 12	Allele 11	Allele 12	X	Y
1	1	1	0	0	1	-0.01456	0.00444
2	2	1	0	0	0	-0.00915	-0.00496
2	3	1	0	1	1	-0.00903	-0.00532
3	4	1	0	1	0	-0.00362	-0.01472
4	5	0	0	0	1	-0.00541	0.00940
4	6	1	1	0	1	-0.00522	0.01087
5	7	0	0	0	0	0.00000	0.00000
5	8	1	1	0	0	0.00019	0.00147
5	9	0	0	1	1	0.00012	-0.00037
5	10	1	1	1	1	0.00031	0.00111
6	11	0	0	1	0	0.00553	-0.00977
6	12	1	1	1	0	0.00572	-0.00829
7	13	0	1	0	1	0.00392	0.01582
8	14	0	1	0	0	0.00934	0.00642
8	15	0	1	1	1	0.00946	0.00606
9	16	0	1	1	0	0.01487	-0.00334
		Scaling Factor 				0.01747	0.01748

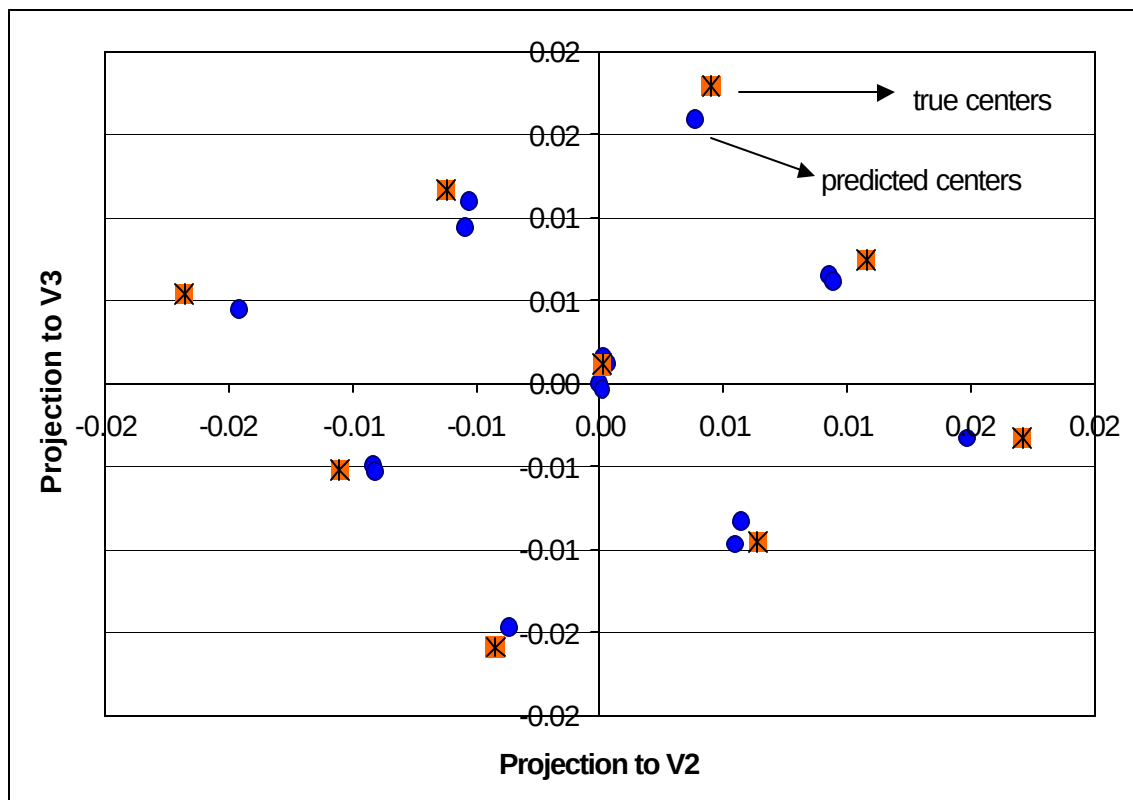


FIGURE 5.26: Comparison of True and Predicted Cluster Centers for Scores of 10,000 Synthetic Profiles for d13s317 and d16s539 Loci –Pair.

presence of other alleles for the members of each cluster. As discussed earlier, all profiles in the original 10,000-row data matrix with identical allele distribution pattern at these four dominant allele positions will map into the same cluster. The non-dominant alleles have negligible contribution towards the determination of cluster centers and assignment of profiles to clusters.

In summary, cluster formation and differentiation is due to the nature of the allele distribution pattern at the dominant alleles of the loci pair. In addition, it is seen that loci with allele probability densities concentrated at just a few (2 or 3) alleles give rise to principal components that are loaded heavily with these dominant alleles, which, in turn, would lead to the formation of good and distinct clusters. Now, the validity of our clustering theory with a real DNA/STR dataset is tested.

5.1.5. CLUSTERING ANALYSIS OF A SMALL REAL DNA/STR DATA SET

All the work presented above has been done with 10,000 synthetic DNA-STR data profiles, generated based on the allele frequency distribution data for Caucasians as given in the CODIS database. Recently, Dr. Budowle of FBI released the STR profiles of six ethnic groups from the Central American regions [49]. This newly released data has a sample size of approximately two hundred each. Testing the developed PCA model on the real DNA data served as a partial validation of the clustering results. The testing was done to see whether PCA clusters from these real data would coincide with the same clusters obtained from the synthetic data, and vice versa. The real DNA data

was projected onto the second and the third principal components obtained from the synthetic dataset of 10,000 profiles at the d13s317 and d16s539 loci. If scores from the real data sample profiles map into existing nine clusters, then it can be concluded that the relative allele frequency distribution at the four dominant alleles are similar to that used for the synthetic data, and that the clusters identified so far can be used to discriminate the profiles in the real dataset. Before the results obtained from the validation test is obtained, the projection of the new DNA dataset onto the existing principal components obtained from the original synthetic data is examined.

To compare the profiles of two different groups or set of data, the DNA profiles of one dataset is projected onto the principal components applicable to the reference dataset. The normalization is to be carried out by the inverse of the singular values derived from the reference dataset, because the normalization factor reflects the variance of the original data matrix, and the variance is directly related to the relative allele frequency distribution data. This step is represented in Equation 5.2.

$$\mathbf{X}_{\text{Scores_Real}} = \mathbf{X}_{\text{Real}} * \mathbf{V}_{\text{Synthetic}} * \mathbf{S}_{\text{Synthetic}}^{-1} \quad (5.2)$$

Where, $\mathbf{X}_{\text{Scores_Real}}$ denotes the scores of the real DNA data set for the 176 Caucasians. $\mathbf{X}_{\text{Real}}$ represents the binary dataset for the real DNA. $\mathbf{V}_{\text{Synthetic}}$ denotes the principal component vectors obtained from the reference set of 10,000 synthetic profiles, and $\mathbf{S}_{\text{Synthetic}}^{-1}$ is the inverse of the singular value matrix for the synthetic set to be used for normalization of the new real data. This step of using the old set of singular values for normalization is important, because the old set of singular values represents the

variance exhibited by the original data set, which is related to the allele distribution of the particular loci pair under consideration. If the real data have allele distribution similar to that for the synthetic dataset, then the old set of singular values should still be applicable for normalization. The 2-D scores plot for the real DNA profiles is obtained by plotting the second and third columns of $X_{\text{scores_real}}$.

The sample set for the real data was first converted to the binary representation format, with 1's and 0's as discussed earlier for the synthetic data set of 10,000 profiles. The allele information for loci d13s317 and d16s539 was extracted from the new DNA data. The companion scores matrix is computed, by projecting the real data set of 176 samples at these two loci onto the second and third principal components derived from the large synthetic data set, as shown in Equation 5.2. In Figure 5.27, the scores from the large data set are drawn overlaid on top of the scores from the small data set. The complete overlap is due to the fact that the real dataset is much smaller compared to the size of the synthetic set. In Figure 5.28 the scores from the small data set is projected onto the same principal components and are plotted over the 10,000 scores of the reference set. It is seen that scores from the real data set do fall completely on top of the corresponding clusters from the reference data set but the most important fact is that clusters are preserved. The dark dots in this graph indicate the scores of the 176 real DNA data samples.

Based on the above findings, it is seen that all Scores in the small real DNA STR dataset (176 profiles) are mapped into the same clusters as those present in the large

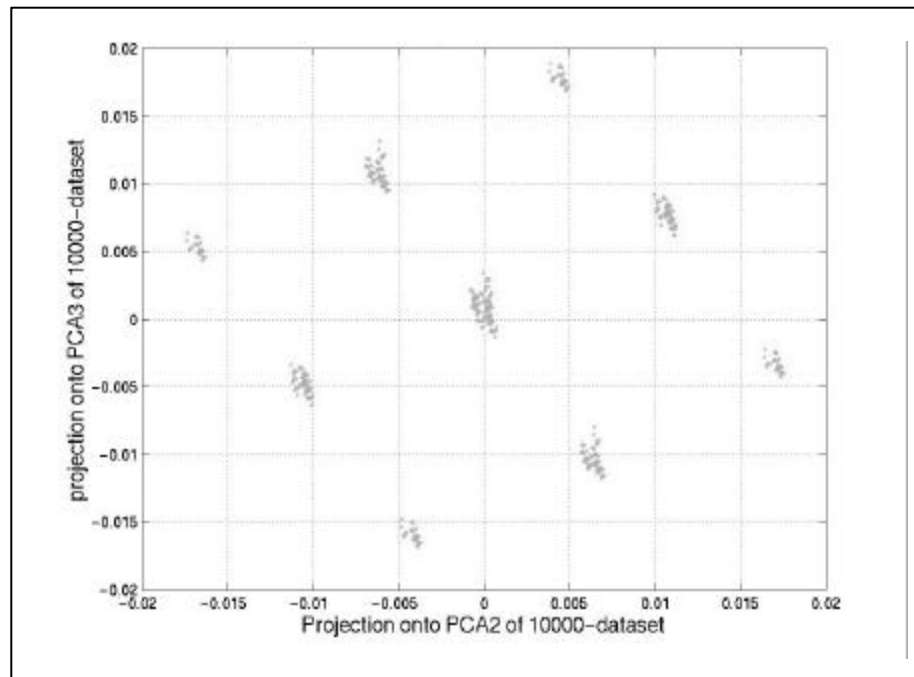


FIGURE 5.27: Scores Clusters of 10000 Synthetic Caucasian DNA Data Plotted over Scores of 176 Real Data Samples in d13s317 and d16s539.

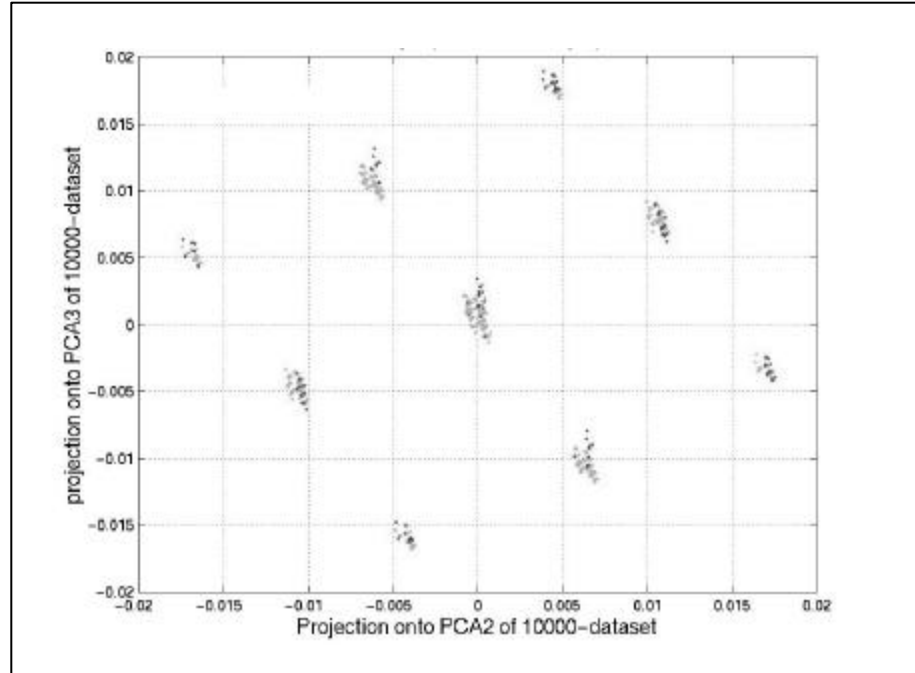


FIGURE 5.28: Scores Clusters of 176 Real Caucasian DNA Data Plotted over Scores of 10000 Synthetic Data in d13s317 and d16s539.

synthetic reference data set (10,000 profiles), and all the allele distribution patterns present in the small real dataset are also present in that of the large dataset.

Now the fraction of the population mapped into each cluster is determined. These fractions reflect the relative allele frequency distribution information of the two loci for these two sample populations. Figure 5.29 shows the comparison between the two data sets. The first bar of each pair of bars represents the fraction of profiles from the small real DNA dataset that belongs to each of the nine clusters, while the second bar represents that of the large synthetic data set. It is observed that the relative fractions of profiles from the two datasets that belong to each of the cluster are 'comparable'. The height of the bars shows similar trend except for the 3rd and the 9th bar where one notices a difference. This may have resulted from the fact that 176 is a relatively small sample size to reflect the true allele distribution frequency. Table 5.21 makes a comparison of the fraction of samples in both the synthetic and real data. But even with some differences observed, the clusters are preserved.

Thus it can be concluded that the real data set and the synthetic data set have comparable allele frequency distributions at these two loci, and that results from clustering studies using synthetic data can be applied directly to real DNA/STR data with the same clustering patterns.

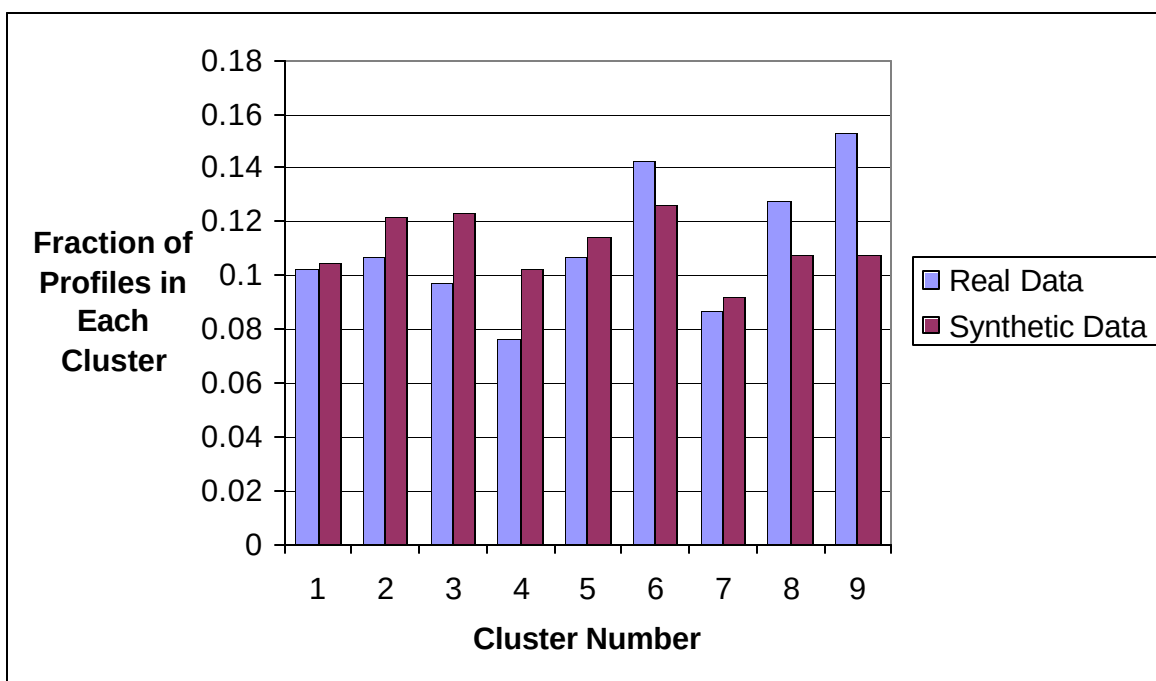


FIGURE 5.29: Comparison of the Percentages of Population within Each Cluster in d13s317 and d16s539 of Real (176 Samples) and Synthetic (10,000 Samples) Data Set.

TABLE 5.21: Comparison of Number/Fraction of Profiles in Each Cluster of the Synthetic and Real Data Set for d13s317 and d16s539 Loci Pair.

Cluster Number	Synthetic Data		Real Data	
	Number of Profiles	Fraction of Profiles	Number of Profiles	Fraction of Profiles
1	1046	0.1046	20	0.1020
2	1216	0.1216	21	0.1071
3	1231	0.1231	19	0.0969
4	1025	0.1025	15	0.0765
5	1143	0.1143	21	0.1071
6	1265	0.1265	28	0.1429
7	918	0.0918	17	0.0867
8	1077	0.1077	25	0.1276
9	1079	0.1079	30	0.1531
Total	10000	1.0000	196	1.0000

5.1.6. EXAMINATION OF LOCI-PAIR THAT GIVE POOR CLUSTER FORMATION

The reason for the occurrence of good distinct clusters in the scores plot has already been examined. Based on the same analysis the unclusterability of DNA samples in certain two-loci combinations can be accounted for. Figure 5.30 shows the 2-D scores plot for 10,000 profiles in d18s51 and d7s820 loci pair. In comparison with examples of good clusters it is seen that this scores plot has elongated, tail-like, loose clusters. Figure 5.31 and 5.32 show poor cluster formation for loci pair d7s820-fga, and fga-th01 respectively.

In Figure 5.30, it might be conclude that there are probably four clusters, but interest lies in the formation of more distinct, and tighter clusters. From visual inspection, it is seen that the clusters are not uniform in size and are not 'tight'. The main reason for poor clusterability is that each of these loci lacks the presence of just a few distinctly dominant alleles. Figure 5.33 shows that that there are around six to seven high occurring alleles, each with probability less than about 0.16. Figure 5.34 shows a similar pattern of the allele frequency distribution for the d7s820 locus.

Having seen the frequency of occurrence of alleles in individual loci, the joint probability of occurrence of 2-allele distribution for the d18s51-d7s820 loci pair is studied, as shown in Figure 5.35. By visual inspection it is seen that lots of peaks of similar heights are present, indicating the lack of just a few dominant allele-pairs for

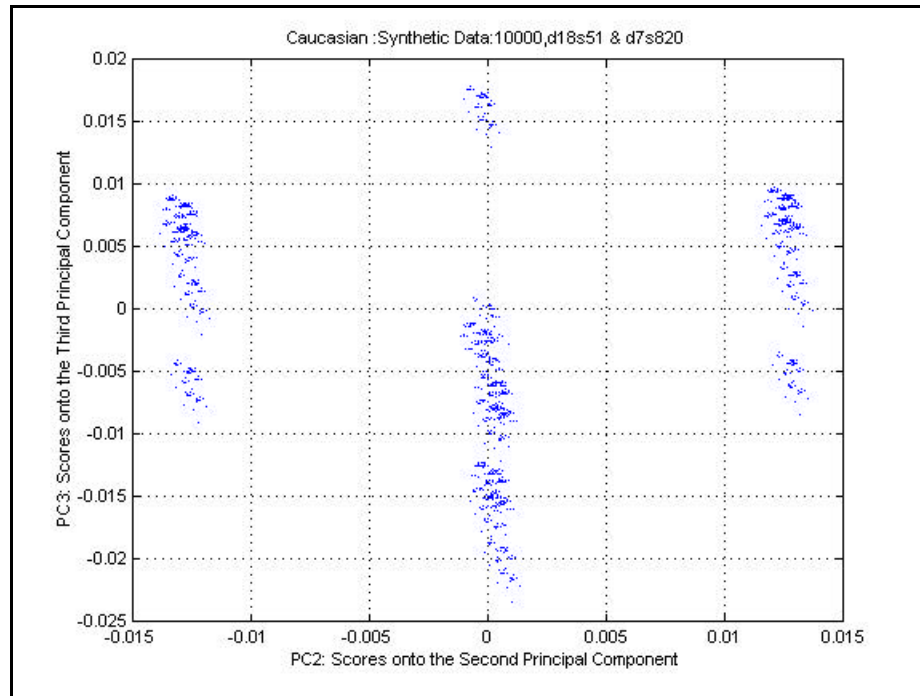


FIGURE 5.30: Poor Cluster Formation by the Scores of 10,000 Synthetic Profiles onto the 2nd and 3rd Principal Components of DNA Samples in Loci d18s51 and d7s820.

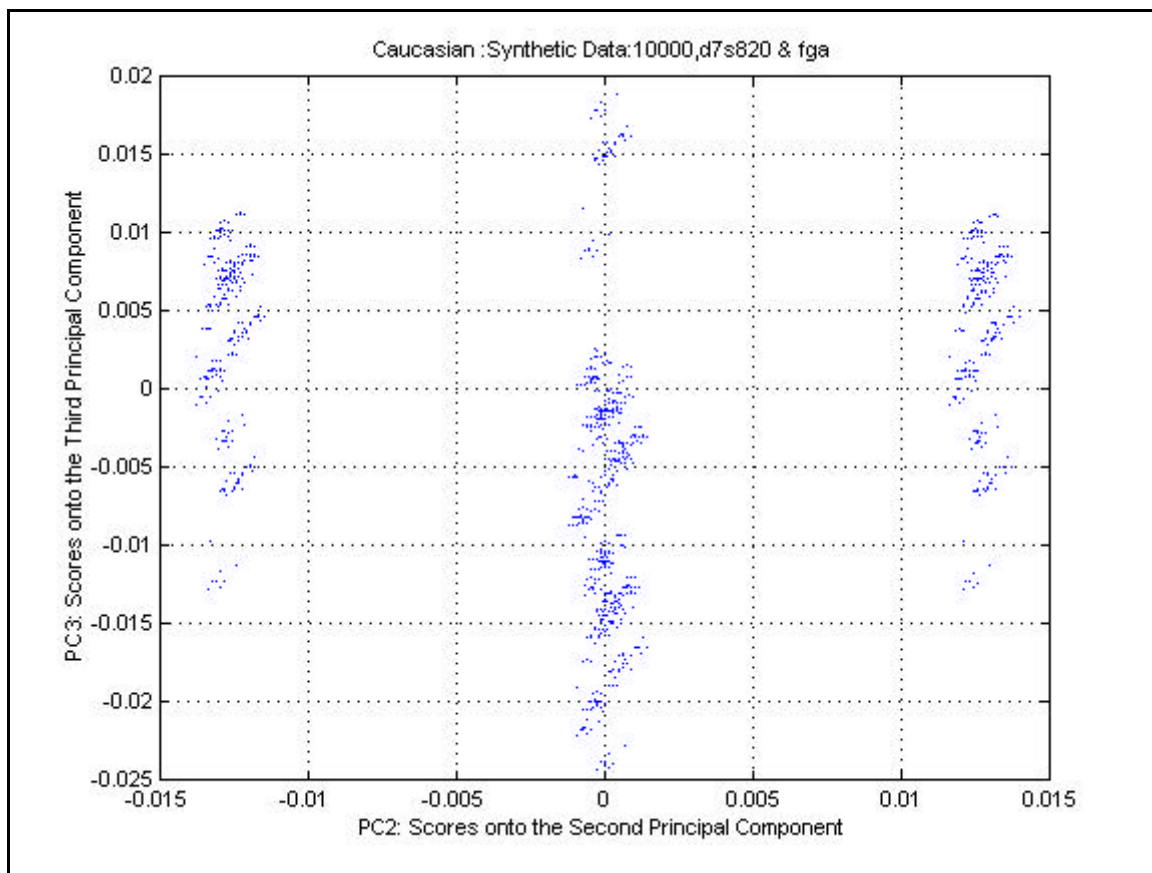


FIGURE 5.31: Poor Cluster Formation by the Scores of 10,000 Synthetic Profiles onto the 2nd and 3rd Principal Components of DNA Samples in Loci d7s820 and fga.

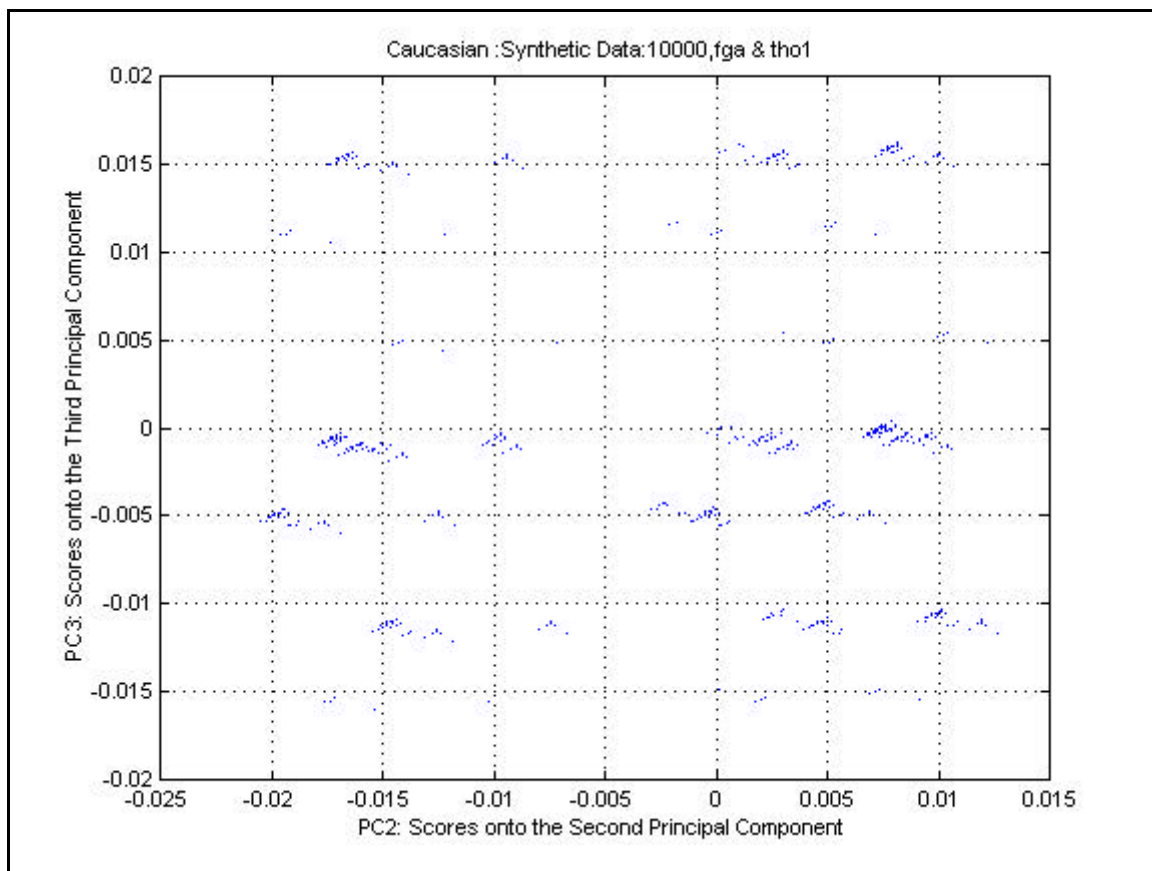


FIGURE 5.32: Poor Cluster Formation by the Scores of 10,000 Synthetic Profiles onto the 2nd and 3rd Principal Components of DNA Samples in Loci fga and tho1.

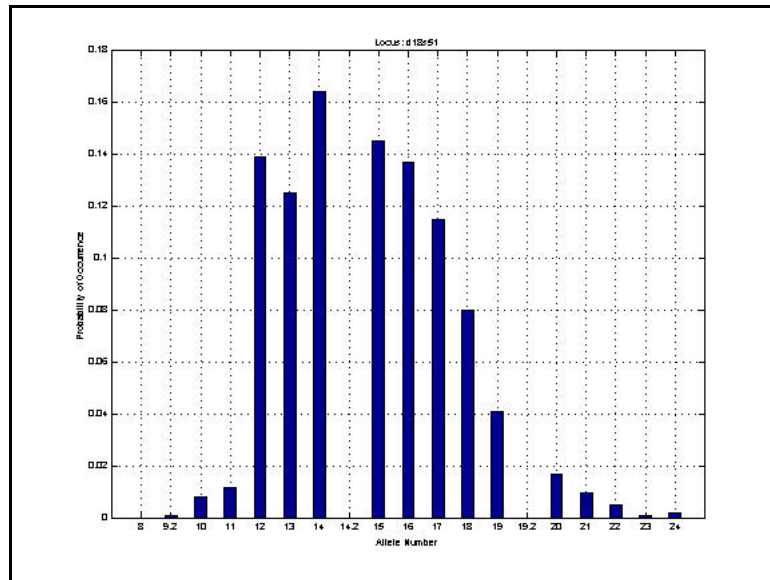


FIGURE 5.33: Allele Frequency Distribution of d18s51.

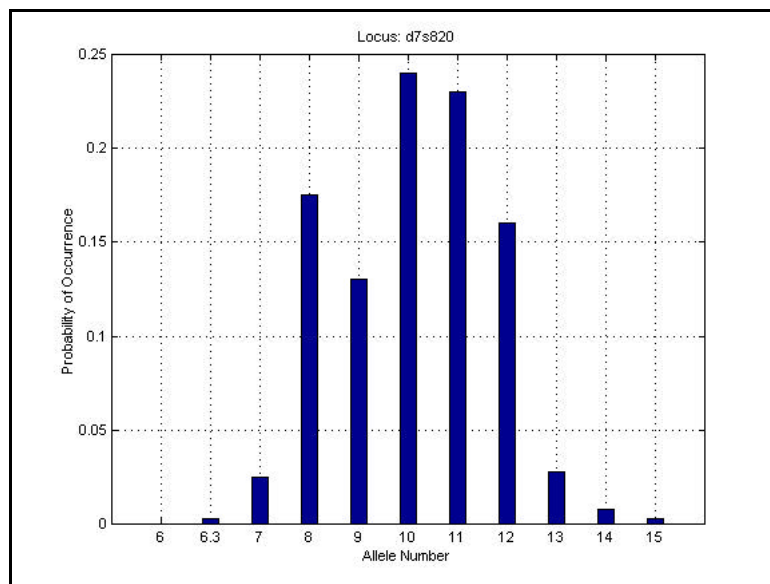


FIGURE 5.34: Allele Frequency Distribution of d7s820.

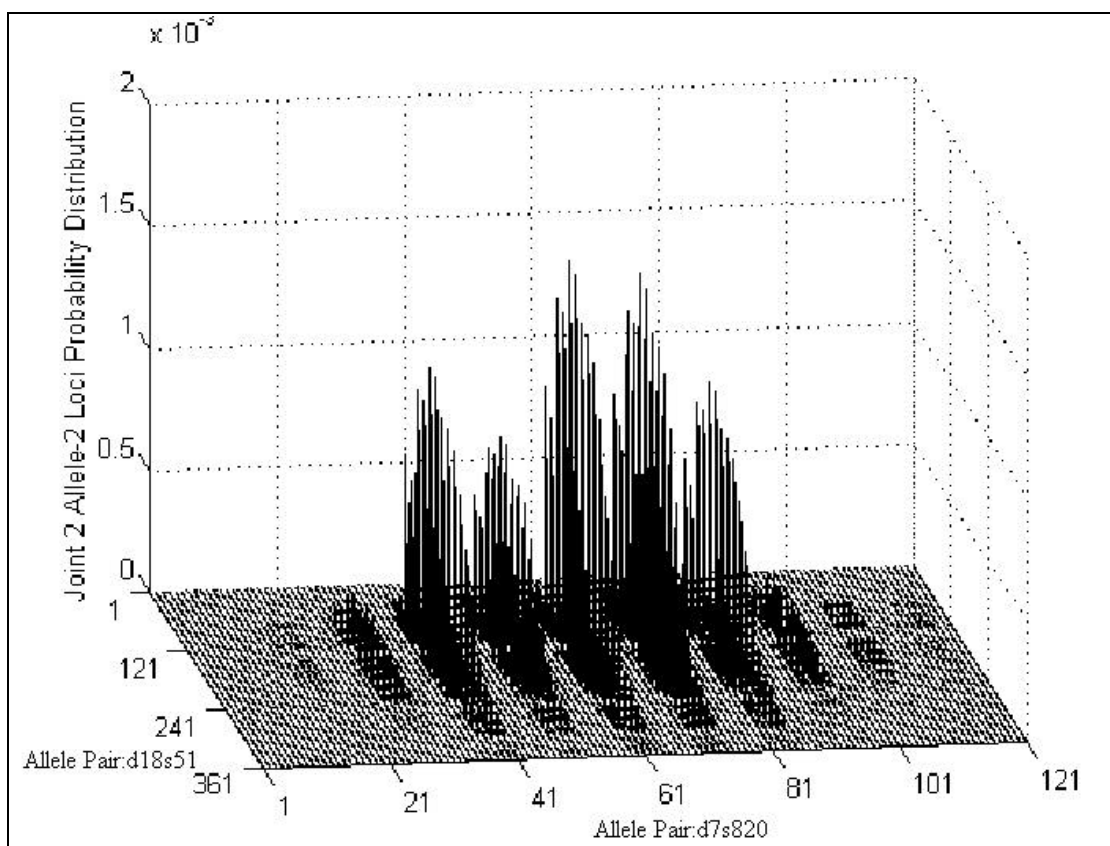


FIGURE 5.35: Joint Probability Density of 2-Allele Pairs for d18s51 and d7s820 Locus.

this loci pair. Figure 5.35 is analogous to similar looking high-rise buildings in a downtown skyline that lacks any dominating building to be distinguished easily. This is to be contrasted with the loci pair that gives good distinct clusters in which just a few high peaks are seen and the contribution by other peaks is negligible. Another example to support our rationale for clusterability is shown in Figures 5.36 and 5.37 for the loci pair d7s820 and fga. They show the allele distribution of the fga locus, in which there are 5 or so alleles that have comparable probability of occurrence, and thus lacking the lumpy nature that renders clustering of sample profiles. Figure 5.37 shows the joint probability of 2-allele distribution of d7s820-fga loci pair, where again a similar downtown trend of joint probability peaks is observed.

It is concluded that the lack of just 2 to 3 distinct high lumpiness in the allele frequency distribution leads to poor cluster formation. Out of the 120 possible 2-loci combination pairs, more than 80 yielded poor clusters.

5.1.7. INCORPORATION OF CLUSTERS INTO THE DECISION TREE FOR SEARCH AND RETRIEVAL OF DNA PROFILES

The main objective of studying clusterability of DNA profiles is to be able to organize the storage of DNA profiles into distinct clusters, so as to improve efficiency of the search engine, when searching for DNA profiles that match the TARGET profile in a forensic application. Figure 5.38 represents a decision tree structure that is employed for performing the search. The top-most node in the decision tree is the root node.

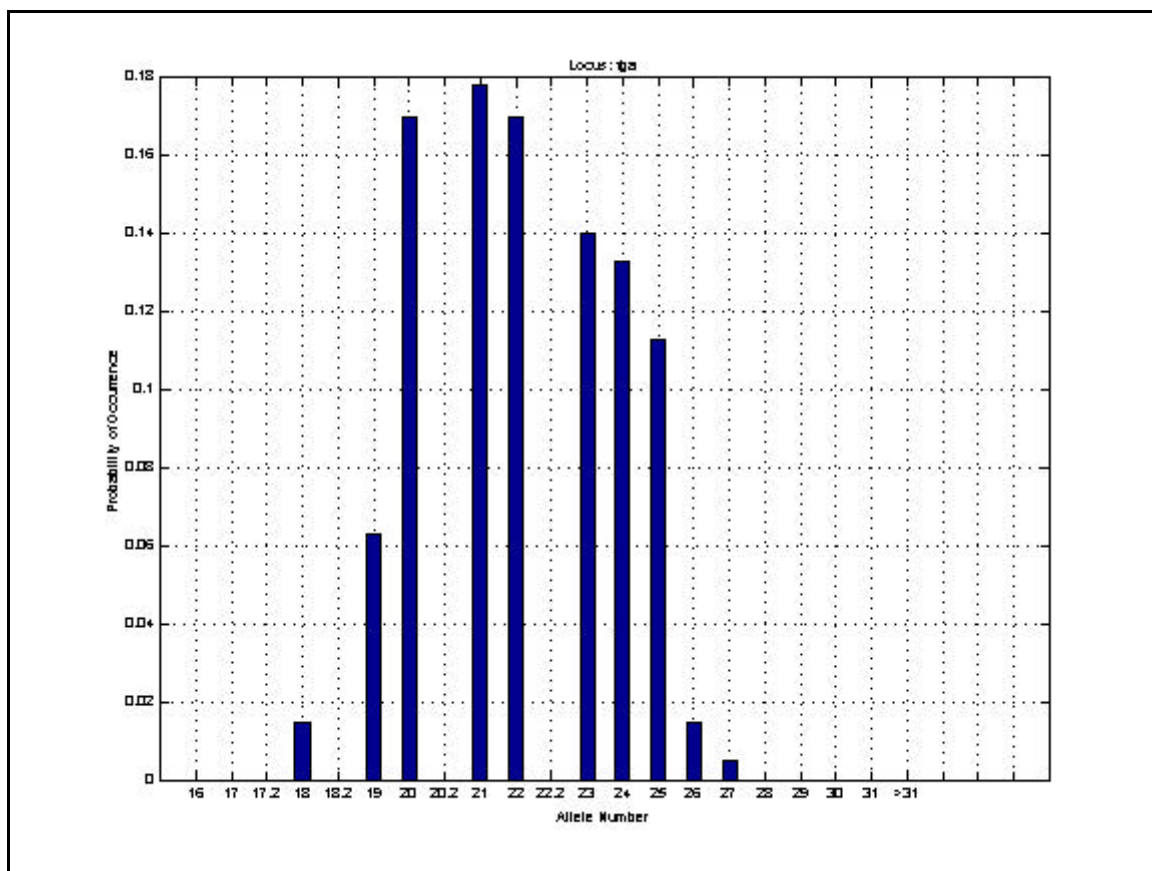


FIGURE 5.36: Allele Frequency Distribution of fga.

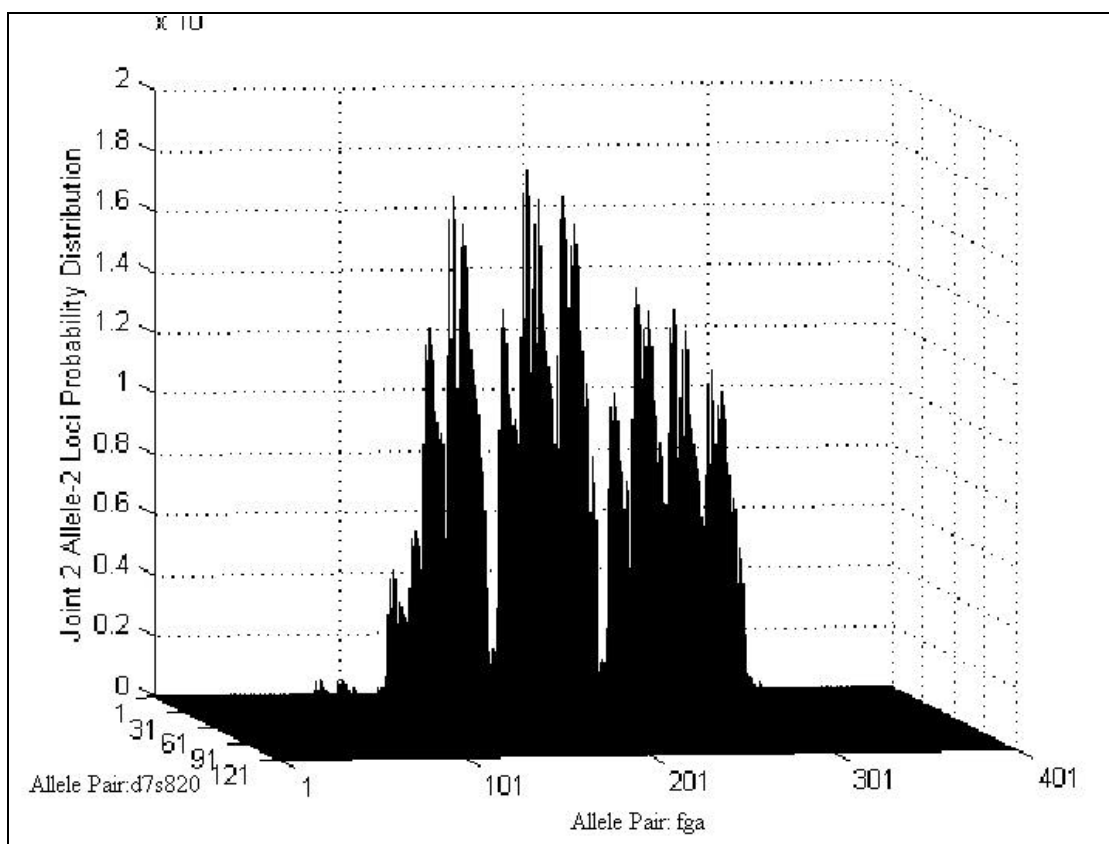


FIGURE 5.37: Joint Probability Density of 2-Allele Pairs for d7s820 and fga Locus

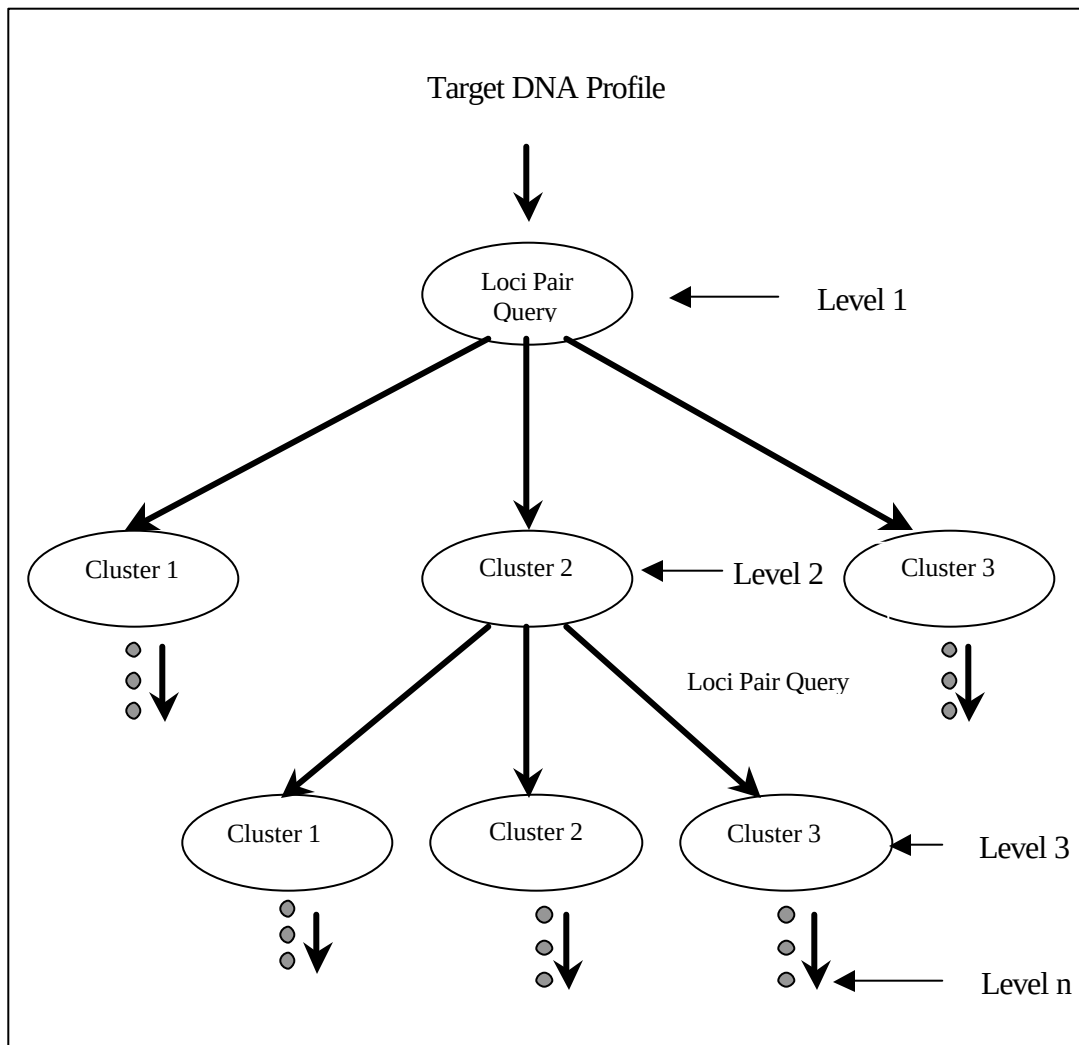


FIGURE 5.38: Decision Tree Structure Based on Clusters Obtained Using Multivariate Statistical Analysis.

The root node further branches out to various non-terminal or leaf nodes. These leaf nodes further branch out to nodes originating at that non-terminal node. Levels 2, 3 and down to Level 'n' show the arrangement of the non-terminal/leaf nodes. The data records of the DNA database are stored at each non-terminal node, and data clustering identified by the multivariate statistical analysis can be performed at each level to obtain partitioning of the DNA database. Each cluster would correspond to a node.

As discussed earlier and from Table 5.9, it is known that the arrangement of clusters that can be utilized to partition the database into groups is of nearly equal size. After establishing good and distinct clusters by using principal component analysis for a particular loci pair, the membership profiles of each cluster are recorded to denote the partitioning of the entire DNA database into these pre-determined clusters. Each DNA profile belongs to only one of these clusters at each successive level. The members of each cluster from the preceding level will be further partitioned using another loci-pair. Assignment of clusters is made based on the DNA allele information contained in the profiles.

Partitioning reduces the search problem by approximately one order of magnitude at each level (if there are nine to ten clusters for each loci pair clustering). The improved search efficiency for the above task is due to the fact, for exact match that at each level, in spite of searching the entire database the search is carried out in only one cluster of the data records in the database where the target may reside. The search

process continues by query of the target's allele distribution at other pairs of DNA loci at each level, resulting in classification to a subsequent PCA cluster and thereby reducing the search by another order of magnitude. Partitions based on PCA may be inserted at the non-terminal nodes of the database tree at each level. As a result, for a database size of 10^8 , after descending about six or seven levels of the decision tree, the number of profiles to be searched over is down to the order of 10's and 100's. Then, sequential search through these profiles for matching profiles can be carried out in a linear fashion.

5.1.8. SUMMARY FOR THE CLUSTERABILITY OF PROFILES BY THE PRINCIPAL COMPONENT ANALYSIS (PCA) APPROACH

Using synthetic data, it has been established that DNA/STR profiles can be partitioned into distinct clusters using the PCA approach. The partition of the database is based on the allele distribution pattern at the two suitably chosen loci. From earlier discussion it is seen that certain two-loci combinations yield better clustering than others. The important factors that determine good clustering and the reason for the clustering have been presented and discussed. Successive partitioning using a different two-loci combination at each node level would reduce the number of profiles that have to be searched through for matching profiles, and thereby improving the search efficiency by about one order of magnitude through each level of the decision tree. The construction of a tree with 'n' levels will reduce the number of profiles to be searched through at the n^{th} level by about n orders of magnitude.

5.2. DEVELOPMENT AND UTILITY OF A SIMILARITY INDEX (SI) PARAMETER FOR RANKING CANDIDATE PROFILES

In this section, results obtained using the developed similarity index, as described in chapter 4, on the synthetic database will be presented. From Figure 5.39, it is seen that the similarity index will be applied after level 'n'.

5.2.1. ROBUST RANKING OF A SET OF SYNTHETIC DATA PROFILES

Two sets of DNA data, one of one hundred and twenty six synthetic profiles, and the other of one thousand and twenty six candidate profiles in six loci were generated to study the feasibility of the developed similarity index in ranking the profiles. The profiles contain loci with either heterozygous or homozygous alleles. Six loci were chosen because it is simpler than handling the entire set of thirteen loci, and it is projected that successful ranking of six-loci profiles would imply successful ranking of thirteen loci profiles. The set also contains mixtures and profiles with unequal number of loci to others. The six loci that were used to create synthetic profiles are as follows: d13s317, d16s539, d5s818, d18s1179, tpox and vWa. These loci were chosen because they have been shown to result in good clusterability by the principal component analysis approach. The development of the Similarity Index has been presented in Section 4.2.1. Table 5.22 and 5.23 shows the characteristics of profiles generated for similarity index study for the small (126) and large (1026) data sets respectively. Twenty six profiles of each of the two datasets are of profiles of ambiguous nature (example: mixtures), therefore are to be flagged, not ranked.

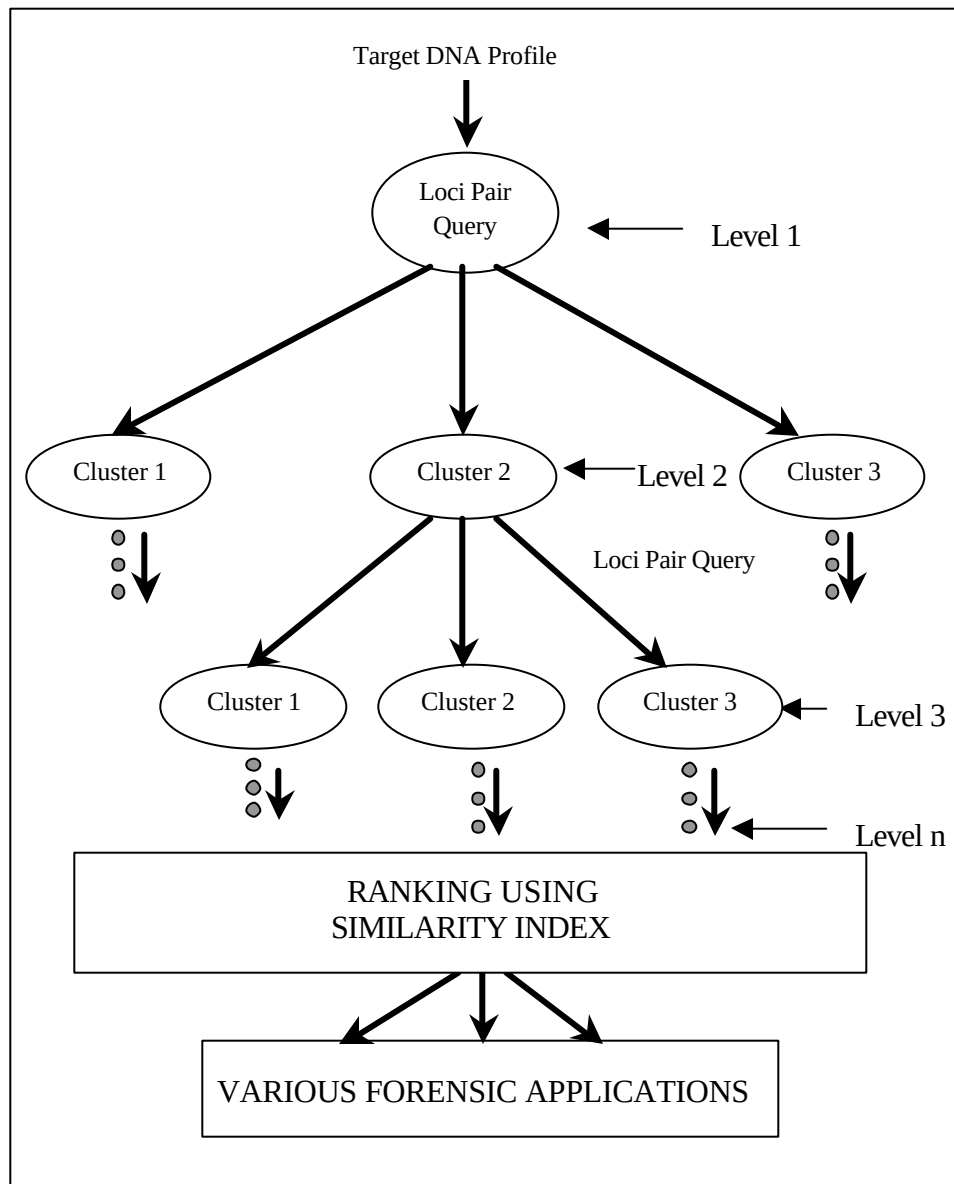


FIGURE 5.39: Utilization of Similarity Index (SI) in the Search Decision Tree.

The similarity index developed, correctly rank-ordered all the unambiguous profiles of a set of one hundred and twenty six DNA/STR synthetic profiles of known characteristics and relative rank ordering. Results obtained using the set of weights ($\alpha_1=0.70$, $\alpha_2=0.20$, $\alpha_3=0.10$) is shown in Tables 5.24 through 5.28. Table 5.24 shows the results of the top twenty ranked candidates and the associated allele distribution patterns, the numerical values of the corresponding PML, PMA, PBPD^c and the overall SI.

From Table 5.24, it is seen that candidate ID 100 is ranked first, as it should, with SI of 1 because it is an exact replica of that of the target. Looking into candidates ID 1 and 8, ranked second and third respectively, both candidates have the same PML value (0.8333, i.e. 10 out of 12) but they differ in PMA, because ID 1 has one-mismatched allele at locus 3 whereas ID 8 has two-mismatched allele at locus 1. Candidate 1 has a higher PMA (0.9167) than candidate 8, which has a PMA value of 0.8333, and therefore candidate 1 is ranked higher as expected. Another example, where PMA of two candidates is the same but their PML's differ is explored. Before looking into the example, it is expected that for the same PMA, the higher the corresponding PML value the more similar that profile would be regarded to the target. Both candidate ID 3 and 7 have the same PMA (0.8333), each with two mismatched alleles; but the two mismatched alleles are distributed in one locus only in ID 3, and into two loci in ID 7. Therefore, the corresponding PML value is 0.8333 for ID 3, and only 0.6667 for ID 7. The ranks for candidate 3 and 7 are 5th and 8th respectively.

Table 5.22: Types of Synthetic Profiles (126) used to Study the Feasibility of Similarity Index.

Total Number of Profiles	126
Target Profile to be Compared With	Heterozygous across all six loci
Total Number of Loci Used	6 (d13s317, d16s539, d18s1179, d5s818, tpox, vWa)
Number of Profiles with Mixtures	9 (ID#: 118-126)
Number of Ambiguous Profiles	12 (ID#: 106-117)
Number of Profiles with 1 mismatched allele to that of the target	1 (ID#: 1)
Number of Profiles with 2 mismatched allele to that of the target	9 (ID#: 2-10)
Number of Profiles with 4 mismatched allele to that of the target	30 (ID#: 11-40)
Number of Profiles with 5 mismatched allele to that of the target	20 (ID#: 41-60)
Number of Profiles with 8 mismatched allele to that of the target	10 (ID#: 61-70)
Number of Profiles with 9 mismatched allele to that of the target	10 (ID#: 71-80)
Number of Profiles with 10 mismatched allele to that of the target	5 (ID#: 81-85)
Number of Profiles with 11 mismatched allele to that of the target	5 (ID#: 86-90)
Number of Profiles with 12 mismatched allele to that of the target	9 (ID#: 91-99)
Number of Profiles that are exact replica of the target	1 (ID#: 100)
Number of Profiles with an extra 7th Locus	5 (ID#: 101-105)

Table 5.23: Types of Synthetic Profiles (1026) used to Study the Feasibility of Similarity Index.

Total Number of Profiles	1026
Target Profile to be Compared With	Heterozygous across all six loci
Total Number of Loci Used	6 (d13s317, d16s539, d18s1179, d5s818, tpox, vWa)
Number of Profiles with Mixtures	9 (ID#: 1018-1026)
Number of Ambiguous Profiles	12 (ID#: 1006-1017)
Number of Profiles with 1 mismatched allele to that of the target	1 (ID#: 1)
Number of Profiles with 2 mismatched allele to that of the target	9 (ID#: 2-100)
Number of Profiles with 4 mismatched allele to that of the target	30 (ID#: 101-400)
Number of Profiles with 5 mismatched allele to that of the target	20 (ID#: 401-600)
Number of Profiles with 8 mismatched allele to that of the target	10 (ID#: 601-700)
Number of Profiles with 9 mismatched allele to that of the target	10 (ID#: 701-800)
Number of Profiles with 10 mismatched allele to that of the target	5 (ID#: 801-850)
Number of Profiles with 11 mismatched allele to that of the target	5 (ID#: 851-900)
Number of Profiles with 12 mismatched allele to that of the target	9 (ID#: 901-999)
Number of Profiles that are exact replica of the target	1 (ID#: 1000)
Number of Profiles with an extra 7th Locus	5 (ID#: 1001-1005)

TABLE 5.24: Results obtained by Applying the Similarity Index on the Synthetic Data Set of 126 Profiles for Six Loci and Weights ($a_1=0.70$, $a_2=0.20$, $a_3=0.10$), Showing Ranks from 1-20.

LOCUS	CANDIDATE ID	100	1	8	9	3	10	2	7	6	5	4	21	17	18	20	19	35	27	16	29
	RANK	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	TARGET ALLELES	CANDIDATE ALLELES																			
1	8	8	8	9	8	8	8	8	8	10	8	8	8	9	8	8	10	8	8	8	8
	11	11	11	12	11	11	11	11	12	11	11	11	11	12	11	11	13	11	11	11	12
2	8	8	8	8	8	8	8	8	8	8	8	10	9	9	8	8	8	9	10	9	9
	11	11	11	11	11	11	11	12	11	11	11	11	12	12	11	11	11	12	11	13	13
3	9	9	10	9	11	11	9	11	9	9	9	11	9	9	11	9	9	9	9	9	9
	13	13	13	13	14	14	13	13	13	14	13	13	13	13	14	13	13	14	14	14	13
4	10	10	10	10	10	10	12	10	10	10	10	10	11	10	10	11	10	10	10	10	10
	14	14	14	14	14	14	16	14	16	14	14	14	15	14	14	15	14	14	14	14	14
5	8	8	8	8	8	8	8	8	8	8	8	8	8	8	9	10	8	8	8	8	9
	11	11	11	11	11	11	11	11	11	11	13	11	11	11	12	12	11	11	11	11	11
6	15	15	15	15	15	15	15	15	15	15	16	15	15	15	15	15	16	15	16	16	15
	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	21	20	19	18	18
TOTAL NUMBER OF MATCHING LOCI		6	5	5	5	5	5	4	4	4	4	4	4	4	4	4	4	3	3	3	3
TOTAL NUMBER OF MATCHING ALLELES		12	11	10	10	10	10	10	10	10	10	10	8	8	8	8	8	8	8	8	8
TOTAL NUMBER OF MISMATCHED ALLELES		0	1	2	2	2	2	2	2	2	2	2	4	4	4	4	4	4	4	4	4
PML		1.0000	0.8333	0.8333	0.8333	0.8333	0.8333	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.5000	0.5000	0.5000	0.5000
PMA		1.0000	0.9167	0.8333	0.8333	0.8333	0.8333	0.8333	0.8333	0.8333	0.8333	0.8333	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667
PBPD		0.0000	0.0185	0.0417	0.0556	0.0556	0.0667	0.0537	0.0542	0.0602	0.0643	0.0704	0.0667	0.0750	0.1032	0.1048	0.1500	0.0852	0.0852	0.0852	0.0946
PBPD ^c = 1-PBPD		1.0000	0.9815	0.9583	0.9444	0.9444	0.9333	0.9463	0.9458	0.9398	0.9357	0.9296	0.9333	0.9250	0.8968	0.8952	0.8500	0.9148	0.9148	0.9148	0.9054
SIMILARITY INDEX		1.0000	0.8648	0.8458	0.8444	0.8444	0.8433	0.7280	0.7279	0.7273	0.7269	0.7263	0.6933	0.6925	0.6897	0.6895	0.6850	0.5748	0.5748	0.5748	0.5739

TABLE 5.25: Results obtained by Applying the Similarity Index on the Synthetic Data Set of 126 Profiles for Six Loci and Weights ($a_1=0.70$, $a_2=0.20$, $a_3=0.10$), Showing Ranks from 21-40.

LOCUS	CANDIDATE ID	32	37	33	40	24	14	15	38	23	28	39	46	59	45	34	31	26	25	11	12
	RANK	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
	TARGET ALLELES	CANDIDATE ALLELES																			
1	8	9	8	8	8	8	9	8	8	8	8	8	8	9	9	9	9	8	8	9	8
	11	13	11	11	11	11	12	11	13	11	12	11	11	11	11	11	11	11	12	11	12
2	8	8	8	8	8	8	9	8	8	8	10	8	9	9	9	8	8	8	8	9	9
	11	11	12	13	12	11	11	11	11	11	12	13	11	12	12	12	12	12	11	11	11
3	9	9	9	9	10	9	9	10	9	9	9	9	11	9	11	9	10	10	9	9	9
	13	14	13	15	13	13	13	15	13	14	13	13	15	13	15	13	13	13	14	15	13
4	10	11	10	10	10	12	10	10	10	10	10	12	10	10	10	11	10	10	10	12	10
	14	14	16	14	14	16	17	14	14	14	14	15	14	14	14	14	14	14	14	14	14
5	8	8	8	8	10	8	8	9	9	8	10	8	8	10	8	8	8	8	10	8	9
	11	11	11	11	12	12	11	11	11	13	11	13	11	12	11	12	12	13	11	11	11
6	15	15	17	16	15	15	15	15	16	16	15	15	16	15	15	15	15	15	16	15	15
	18	18	19	19	18	19	18	20	20	20	18	18	19	18	18	18	18	19	18	18	21
TOTAL NUMBER OF MATCHING LOCI		3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2
TOTAL NUMBER OF MATCHING ALLELES		8	8	8	8	8	8	8	8	8	8	8	7	7	7	8	8	8	8	8	8
TOTAL NUMBER OF MISMATCHED ALLELES		4	4	4	4	4	4	4	4	4	4	4	5	5	5	4	4	4	4	4	4
PML		0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333
PMA		0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667	0.5833	0.5833	0.5833	0.6667	0.6667	0.6667	0.6667	0.6667	0.6667
PBPD		0.0977	0.1000	0.1037	0.1066	0.1071	0.1083	0.1127	0.1155	0.1161	0.1185	0.1310	0.1241	0.1256	0.1282	0.0780	0.0798	0.0995	0.1036	0.1079	0.1113
PBPD ^c = 1-PBPD		0.9023	0.9000	0.8963	0.8934	0.8929	0.8917	0.8873	0.8845	0.8839	0.8815	0.8690	0.8759	0.8744	0.8718	0.9220	0.9202	0.9005	0.8964	0.8921	0.8887
SIMILARITY INDEX		0.5736	0.5733	0.5730	0.5727	0.5726	0.5725	0.5721	0.5718	0.5717	0.5715	0.5702	0.5543	0.5541	0.5538	0.4589	0.4587	0.4567	0.4563	0.4559	0.4555

TABLE 5.26: Results obtained by Applying the Similarity Index on the Synthetic Data Set of 126 Profiles for Six Loci and Weights ($a_1=0.70$, $a_2=0.20$, $a_3=0.10$), Showing Ranks from 41-60.

LOCUS	CANDIDATE ID	22	30	13	36	54	43	44	52	58	60	57	53	50	56	47	41	49	42	55	51
	RANK	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
	TARGET ALLELES	CANDIDATE ALLELES																			
1	8	8	8	8	10	9	9	9	9	8	9	8	8	10	8	8	9	8	8	9	9
	11	13	13	11	11	11	11	13	12	11	11	12	11	12	13	11	11	11	13	11	11
2	8	8	8	8	10	9	9	8	9	9	10	8	8	8	8	8	9	8	9	9	8
	11	12	11	11	11	12	11	11	11	11	12	11	12	13	11	13	11	12	11	11	13
3	9	9	10	11	9	9	9	9	9	11	9	9	10	9	9	9	10	9	10	10	11
	13	13	13	13	15	13	15	14	13	15	13	15	13	15	13	13	13	14	13	13	13
4	10	10	10	13	10	11	11	10	10	11	10	12	10	11	12	12	11	10	10	10	11
	14	14	16	14	14	14	15	15	14	14	14	14	14	14	14	16	14	16	14	14	14
5	8	9	8	9	8	8	8	8	10	9	10	9	10	8	8	8	9	9	8	10	8
	11	11	12	11	11	11	11	11	11	11	11	12	13	11	13	13	11	11	12	11	11
6	15	17	15	16	15	16	15	16	16	15	16	15	16	15	16	16	15	15	17	17	15
	18	18	18	18	19	18	18	18	18	18	18	18	18	18	19	18	18	20	18	18	20
TOTAL NUMBER OF MATCHING LOCI		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1
TOTAL NUMBER OF MATCHING ALLELES		8	8	8	8	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
TOTAL NUMBER OF MISMATCHED ALLELES		4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
PML		0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.1667	0.1667	0.1667	0.1667	0.1667
PMA		0.6667	0.6667	0.6667	0.6667	0.5833	0.5833	0.5833	0.5833	0.5833	0.5833	0.5833	0.5833	0.5833	0.5833	0.5833	0.5833	0.5833	0.5833	0.5833	0.5833
PBPD		0.1155	0.1173	0.1275	0.1287	0.0875	0.1079	0.1144	0.1226	0.1312	0.1351	0.1388	0.1471	0.1495	0.1560	0.1643	0.0965	0.1257	0.1340	0.1370	0.1412
PBPD ^c = 1-PBPD		0.8845	0.8827	0.8725	0.8713	0.9125	0.8921	0.8856	0.8774	0.8688	0.8649	0.8612	0.8529	0.8505	0.8440	0.8357	0.9035	0.8743	0.8660	0.8630	0.8588
SIMILARITY INDEX		0.4551	0.4549	0.4539	0.4538	0.4413	0.4392	0.4386	0.4377	0.4369	0.4365	0.4361	0.4353	0.4350	0.4344	0.4336	0.3237	0.3208	0.3199	0.3196	0.3192

TABLE 5.27: Results obtained by Applying the Similarity Index on the Synthetic Data Set of 126 Profiles for Six Loci and Weights ($a_1=0.70$, $a_2=0.20$, $a_3=0.10$), Showing Ranks from 61-80.

LOCUS	CANDIDATE ID	48	63	68	70	69	61	67	77	76	65	66	62	64	71	74	72	79	75	73	78
	RANK	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
	TARGET ALLELES	CANDIDATE ALLELES																			
1	8	9	8	8	10	8	10	8	9	8	8	9	8	10	8	9	9	8	10	10	9
	11	11	11	11	12	12	11	13	12	11	12	11	12	13	12	11	13	13	12	12	13
2	8	8	9	9	8	10	9	10	8	9	8	9	10	10	9	9	9	10	9	10	10
	11	13	12	11	12	13	12	13	11	11	13	11	13	11	12	12	11	13	11	12	11
3	9	9	10	11	9	10	11	10	10	11	9	9	9	11	10	9	10	10	11	9	10
	13	13	13	15	14	15	15	13	15	15	14	15	14	13	15	15	15	14	13	15	15
4	10	12	11	10	12	10	11	10	11	12	11	12	10	12	11	11	11	10	11	10	10
	14	14	16	15	15	14	14	14	16	15	16	14	16	15	14	14	14	16	14	16	16
5	8	10	10	9	9	10	8	10	8	9	10	9	10	8	10	10	10	10	10	8	10
	11	11	11	12	13	12	11	13	12	12	11	12	12	13	12	13	13	12	12	13	11
6	15	15	16	16	15	15	17	16	16	17	16	17	15	15	15	16	16	15	17	16	17
	18	19	19	20	18	19	20	20	20	20	20	20	20	19	19	20	18	19	20	20	20
TOTAL NUMBER OF MATCHING LOCI		1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
TOTAL NUMBER OF MATCHING ALLELES		7	4	4	4	4	4	4	3	3	4	4	4	4	3	3	3	3	3	3	3
TOTAL NUMBER OF MISMATCHED ALLELES		5	8	8	8	8	8	8	9	9	8	8	8	8	9	9	9	9	9	9	9
PML		0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PMA		0.5833	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.2500	0.2500	0.3333	0.3333	0.3333	0.3333	0.2500	0.2500	0.2500	0.2500	0.2500	0.2500	0.2500
PBPD		0.1518	0.1828	0.2050	0.2191	0.2312	0.2324	0.2721	0.2210	0.2550	0.2203	0.2222	0.2441	0.2680	0.2145	0.2531	0.2633	0.2668	0.2710	0.2805	0.2990
PBPD ^c = 1-PBPD		0.8482	0.8172	0.7950	0.7809	0.7688	0.7676	0.7279	0.7790	0.7450	0.7797	0.7778	0.7559	0.7320	0.7855	0.7469	0.7367	0.7332	0.7290	0.7195	0.7010
SIMILARITY INDEX		0.3182	0.2651	0.2628	0.2614	0.2602	0.2601	0.2561	0.2446	0.2412	0.1446	0.1445	0.1423	0.1399	0.1286	0.1247	0.1237	0.1233	0.1229	0.1220	0.1201

TABLE 5.28: Results obtained by Applying the Similarity Index on the Synthetic Data Set of 126 Profiles for Six Loci and Weights ($a_1=0.70$, $a_2=0.20$, $a_3=0.10$), Showing Ranks from 81-100.

LOCUS	CANDIDATE ID	80	85	83	84	81	82	86	87	88	89	90	99	98	92	97	93	95	94	91	96
	RANK	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
	TARGET ALLELES	CANDIDATE ALLELES																			
1	8	10	9	8	10	8	8	9	9	10	10	10	9	9	9	10	9	9	9	10	10
	11	13	12	12	13	13	12	13	13	13	13	13	13	12	12	12	12	13	12	12	13
2	8	10	9	10	8	10	10	9	10	10	9	10	9	9	9	9	9	10	10	10	10
	11	11	11	13	13	13	12	12	12	12	13	12	12	13	13	12	12	12	12	13	13
3	9	9	11	9	11	11	11	10	11	10	9	10	11	10	10	10	11	11	11	11	11
	13	15	14	15	14	14	15	14	13	15	14	15	14	15	15	15	15	15	14	14	14
4	10	12	12	11	10	11	11	11	12	11	12	12	11	12	11	11	12	11	12	12	12
	14	16	16	15	15	16	16	15	15	16	16	16	16	15	15	15	15	16	16	16	15
5	8	10	8	10	9	9	9	9	9	10	9	9	9	9	9	10	10	10	10	10	10
	11	11	12	12	12	11	13	12	12	12	12	13	12	13	13	13	13	12	13	13	13
6	15	16	17	16	17	17	17	15	17	16	17	17	16	16	16	16	17	16	17	17	17
	18	20	20	20	19	19	18	19	20	18	20	18	20	19	20	20	19	20	20	20	19
TOTAL NUMBER OF MATCHING LOCI		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TOTAL NUMBER OF MATCHING ALLELES		3	2	2	2	2	2	1	1	1	1	1	0	0	0	0	0	0	0	0	0
TOTAL NUMBER OF MISMATCHED ALLELES		9	10	10	10	10	10	11	11	11	11	11	12	12	12	12	12	12	12	12	12
PML		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PMA		0.2500	0.1667	0.1667	0.1667	0.1667	0.1667	0.0833	0.0833	0.0833	0.0833	0.0833	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
PBPD		0.3180	0.2710	0.2793	0.2865	0.2877	0.2997	0.2305	0.3138	0.3270	0.3328	0.3603	0.2990	0.3020	0.3020	0.3300	0.3443	0.3580	0.3758	0.3800	0.4008
PBPD ^c = 1-PBPD		0.6820	0.7290	0.7207	0.7135	0.7123	0.7003	0.7695	0.6862	0.6730	0.6672	0.6397	0.7010	0.6980	0.6980	0.6700	0.6557	0.6420	0.6242	0.6200	0.5992
SIMILARITY INDEX		0.1182	0.1062	0.1054	0.1047	0.1046	0.1034	0.0936	0.0853	0.0840	0.0834	0.0806	0.0701	0.0698	0.0698	0.0670	0.0656	0.0642	0.0624	0.0620	0.0599

Thus, it is seen that for the same PMA value, the higher the corresponding PML value, the closer the candidate resembles the target profile because it has a higher number of loci that match that of the target and the weight applicable to PML is usually chosen to be the highest of all weights. Consider another interesting case where PML and PMA for two candidates are the same but they differ in $PBPD^c$. In this example, both candidates have four-mismatched alleles. It is seen from Table 5.24 that candidates 17 and 20 have the same PML (0.6667) and PMA (0.6667) but they differ in $PBPD$ (0.0750 v/s 0.1048). Remember that what goes into SI is $(1-PBPD)$ or $PBPD^c$, not $PBPD$ directly for reasons discussed earlier. It is expected that candidate 17 is ranked higher than candidate 20, as is the case (13th vs 15th). One would expect this ranking because if the percentage of matching loci and the percentage of matching alleles are the same for two candidates, then less the base-pair difference between the unmatched alleles the closer that profile is to the target. From experience, it is known that $PBPD^c$ acts as a fine-tuning parameter in the similarity index. That is, it does and should have some impact on the final profile ranking but not overwhelmingly so. This can be seen by studying another interesting example. Consider candidate profiles 4 and 21. Candidate 4 has two-mismatched alleles distributed in two loci (one each at locus 1 and 3), whereas candidate 21 has four-mismatched alleles but also distributed in two loci (two each at locus 2 and 4). Therefore both candidates have the same PML (0.6667). From Table 5.24, it is observed that candidate 4 has a PMA value of 0.8333 and a $PBPD$ value of 0.0704 (i.e. $PBPD^c=1-PBPD = 0.9296$), whereas candidate 21 has a PMA of 0.6667 and a $PBPD$ value of 0.0667 (i.e. $PBPD^c=1-PBPD = 0.9333$).

Looking closely, it is seen that candidate 4 has a higher PMA but a lower PBPD^c than candidate 21, and is ranked 11th above that of candidate 21 (ranked 12th) with $\alpha_1=0.7$, $\alpha_2=0.2$ and $\alpha_3=0.1$. Thus from our examples it is seen that PML is the most important parameter for ranking followed by PMA and PBPD/PBPD^c in that order.

The following example will illustrate the importance of PML in determine ranking. It shows that candidates with a higher number of mismatched alleles may sometimes be ranked higher than candidates with a less number of mismatched alleles. Consider candidate 46 with five-mismatched alleles, and candidate 25 with four-mismatched alleles as another interesting example. Candidate 46/25 has a PML, PMA and PBPD values of 0.5000/0.3333, 0.5833/0.6667 and 0.1241/0.1036 respectively. It is seen that candidate 46 has a higher PML value even though it has more mismatched alleles than that of candidate 25. The reason ID 46 has higher PML is that the five-mismatched alleles in ID 46 are distributed across three loci (i.e. locus 2, 3 and 6) whereas the four-mismatched alleles of ID 25 are distributed across four loci (i.e. locus 1, 3, 5 and 6). Candidate 46 and 25 are therefore ranked 32nd and 38th respectively justifying the fact that PML is the most important parameter for determine rankings. Similar observations can be deduced for various other scenarios as seen in Table 5.25, where candidates 77 and 76 with nine-mismatched alleles are ranked higher than candidates with eight-mismatched alleles (i.e. candidate ID 65, 66, 62 and 64).

Now we explore how sensitive rankings are to changes in the weights used for the

three subparameters of the Similarity Index. In order for SI to have a meaningful application utility, it is desirable for the rankings to remain robust over a range of weight values for the PML, PMA and PBPD^c parameters of the Similarity Index. Otherwise, the ranking becomes arbitrary and would be open to subjective interpretation and thus lose its utility and credibility.

Table 5.29 shows the ranking of the top twenty-five profiles for the same synthetic data set of one hundred and twenty six profiles, resulting from using twenty different weight sets shown in Table 5.30. The first nine sets of weights differ in PML and PMA weights, while keeping the weight for PBPD^c low but the same. This is to see how weights for PML and PMA affect the ranking, because in our previous discussion we concluded that PBPD^c should not have a significant impact on rankings. The ranking results when weights for PBPD^c are also changing will also be shown and discussed. In a second set (Cases 10-20, Table 5.30), the effect of changing weights for PBPD^c on profile rankings is discussed.

Table 5.29 shows that, irrespective of the choice of weights, candidate ID 100 always occupies the topmost rank '1' because it is the exact replica of the target. Now we examine the profile rankings for the first nine weight sets where the value of PBPD^c is kept at 0.1 but PML values vary from 0.9 down to 0.1. It is seen that the rankings of the top 20 candidates are preserved and they do not shuffle except in Case 1, where ranks 11 and 12 are interchanged in comparison to rankings obtained for Cases 2-10.

TABLE 5.29: Top Twenty Five Ranked Profiles obtained from Similarity Index for Different Set of Weights (Table 5.30) on a Data Set of 126 Synthetic Candidate Profiles.

CASE	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
RANK	CANDIDATE ID																			
1	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	10	1	1
3	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	9	10	8
4	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	8	9	2
5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	8	7
6	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	2	1	7	9
7	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	7	21	6	3
8	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	10	20	5	6
9	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	19	4	5
10	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	18	3	21
11	21	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	17	2	10
12	4	21	21	21	21	21	27	21	21	21	21	21	21	21	21	21	21	7	40	4
13	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	6	39	17
14	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	35	5	38	34
15	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	27	4	37	31
16	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	16	2	36	35
17	35	35	35	35	35	35	35	35	35	35	35	35	35	35	35	35	34	59	35	27
18	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	31	46	34	16
19	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	18	45	33	54
20	29	29	29	29	29	29	29	29	29	29	29	29	29	29	29	29	29	40	32	29
21	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32	20	39	31	41
22	37	37	37	37	37	37	37	37	37	37	37	37	37	37	37	37	32	38	30	32
23	33	33	33	33	33	33	33	33	33	33	33	33	33	33	33	33	37	37	29	26
24	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	54	35	28	37
25	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	24	33	33	27	18

TABLE 5.30: Different Set of Weights for PML, PMA and PBPD used to Rank 126 Candidate Profiles using Similarity Index Parameter.

WEIGHTS (PARAMETERS)	CASES																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
PML	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.85	0.80	0.70	0.50	0.50	0.30	0.10	0.05	1.00	0	0
PMA	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.10	0.15	0.25	0.45	0.25	0.60	0.85	0.05	0.00	1.00	0.00
PBPD^c	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.05	0.05	0.05	0.05	0.25	0.10	0.05	0.90	0.00	0.00	1.00

The reason for the slight shuffle is accounted for by the fact that the weight for PMA in Case 1 is zero therefore any difference in PMA is ignored. Further look into the allele distribution of candidates 21 and 4 (Table 5.31) to explore how the value of zero weight for PMA affected the minor shuffling in the rank. From Table 5.31 it is seen that candidate 4 and 21 both have the same number of matching loci (i.e. PML is the same for both profiles) but candidate 4 has a higher PMA value with two-mismatched alleles in comparison to that of candidate 21 with four-mismatched alleles. It is also seen that PBPD/PBPD^c is the same for both candidates. Therefore the value of the similarity index is the same since the weight for PMA is zero, resulting in both candidates having the same rank. But if the weight of PMA was non-zero, then candidate 4 will have a higher rank, as observed for case 2-9, than candidate 21 because of its higher PMA value.

An important observation to be made is that, top twenty-five ranked profiles are the same for the sets of weights for Cases 2-16 where the weights for PBPD^c remains low. Then sudden changes in ranks occur thereafter, when the value of PML weight is chosen to be very small, as in Cases 17, 19 and 20, where PML is 0.05, 0.00 and 0.00 respectively. The small value for the PML weight downplays the contribution to similarity of the matching loci leading to an erroneous conclusion. This supports earlier findings that PML is the most important parameter and should have a significant weight attached to it when ranking profiles. Having seen that similarity index measure does provide a way to rank profiles, an important issue

TABLE 5.31: Comparison of Allele Distribution of Candidate 4 and 21 for the Set of Weights of $a_1=0.90$, $a_2=0.00$, and $a_3=0.10$ (i.e. Case 1, Table 5.30).

ALLELE DISTRIBUTION			
LOCUS	TARGET	CANDIDATE 4	CANDIDATE 21
1	8	8	8
	11	11	11
2	8	10	9
	11	11	12
3	9	11	9
	13	13	13
4	10	10	11
	14	14	15
5	8	8	8
	11	11	11
6	15	15	15
	18	18	18

now is to formally establish the extent of the weight region in which the rankings would remain stable despite changes in weight assigned to the parameters. The following section will discuss results obtained from a robustness study of ranking over a wide range of weights. A graphical approach is developed to depict the region of weights that give stable ranking.

5.2.2. SENSITIVITY OF RANKING TO THE CHOICE OF WEIGHTS FOR SIMILARITY INDEX SUBPARAMETERS

The ranking of profiles using the nominal weight set of $(1/3, 1/3, 1/3)$ was compared to the rankings of the corresponding profiles in each of the test weight cases. The shuffling index was computed to reflect the sum total of relative shift in rankings when different weight sets were used. The ternary contour curves obtained by plotting the shuffling index for different weight sets give insight into the impact of the selection of weights on ranking and the range for which rankings remain relatively stable. The shuffling index for the top ten, twenty and all hundred ranked profiles for both the small (126, Table 5.22) and large (1026, Table 5.23) datasets were computed, when different reference weight sets were used to see if the choice of reference weight makes any difference. For each reference weight, different weight sets of the three subparameters varying between 0 and 1 were generated. The size of the weight matrix is 4182-by-3 (Section 4.3.1). Shuffling index contour plots are shown in this section to show the extent of the stable region.

Figure 5.40 shows the contour curves of the shuffling index for the top ten ranked candidate profiles. Using the reference weight set of $[\alpha_1=1/3, \alpha_2=1/3, \alpha_3=1/3]$. The vertical color bar on the right shows the value of the shuffling index as a function of the corresponding color level. It is observed that there is no shuffling (blank space) over most of the weight set region, indicating that the ranks are mostly preserved for this test run. At the outermost region (corresponding to near 0 PML value), the shuffling reached about 20, indicating that on the average, each of the top ten ranked profiles shifted by two positions in comparison to that of reference weight. Figure 5.41 represents the percent of shuffling (Equation 4.15) for the same test case. The figure shows that the percentage shuffling for the top ten ranked candidates is very small, up to about 2% of possible shuffling. From now on contour plots will be presented as percentage shuffling to make appropriate normalized comparison for different weight sets and for different number of top ranked profiles based on which the shuffling is calculated. The contour plot for the top twenty ranked profiles is shown in Figure 5.42. It is still observed that a wide-open region of weight combinations produced no change in ranking. The shuffling at extreme PML value is around 8%. The ternary plot is now presented for the shuffling of the entire set of one hundred profiles over the entire weight sets (26 of the 126 profiles were not ranked because of their ambiguity status). Figure 5.43 shows that the area in which no shuffling occurred is smaller now. However, most of the shuffling is well within 6%. At the extremely low PML weight values, the shuffling increases up to about 12%.

Would the stability in ranking be different if the reference weight set is changed from that of $[\alpha_1=1/3, \alpha_2=1/3, \alpha_3=1/3]$ to others? Does that impact the degree of shuffling? Figures 5.44, 5.45 and 5.46 show the contour plots for the top ten ranked candidate profiles using $[\alpha_1=0.8, \alpha_2=0.1, \alpha_3=0.1]$, $[\alpha_1=0.1, \alpha_2=0.8, \alpha_3=0.1]$, and $[\alpha_1=0.1, \alpha_2=0.1, \alpha_3=0.8]$ as the reference weight set, respectively. These figures show very little changes in ranking across the spectrum. Figures 5.53-5.55 show similar trend for the large dataset of 1026 candidate profiles, where shuffling in the stable region is up to only 2%. This shows that the top ten ranked profiles are insensitive to the choice of reference weight set across a large set of weights chosen. Similar results are obtained when comparisons for the top twenty and top hundred (entire dataset) ranked profiles using the set of new reference weight set are made. Results are shown in Figures 5.47-5.52.

With respect to the top 10% ranked profiles, ranking remains stable over about the entire weight set region. Meaning, α_1 and α_2 can be changed as much as 0.9 without affecting the relative rankings of the top ten to twenty profiles. Robustness in ranking increases the validity of applying the similarity index measuring in the ranking of profiles.

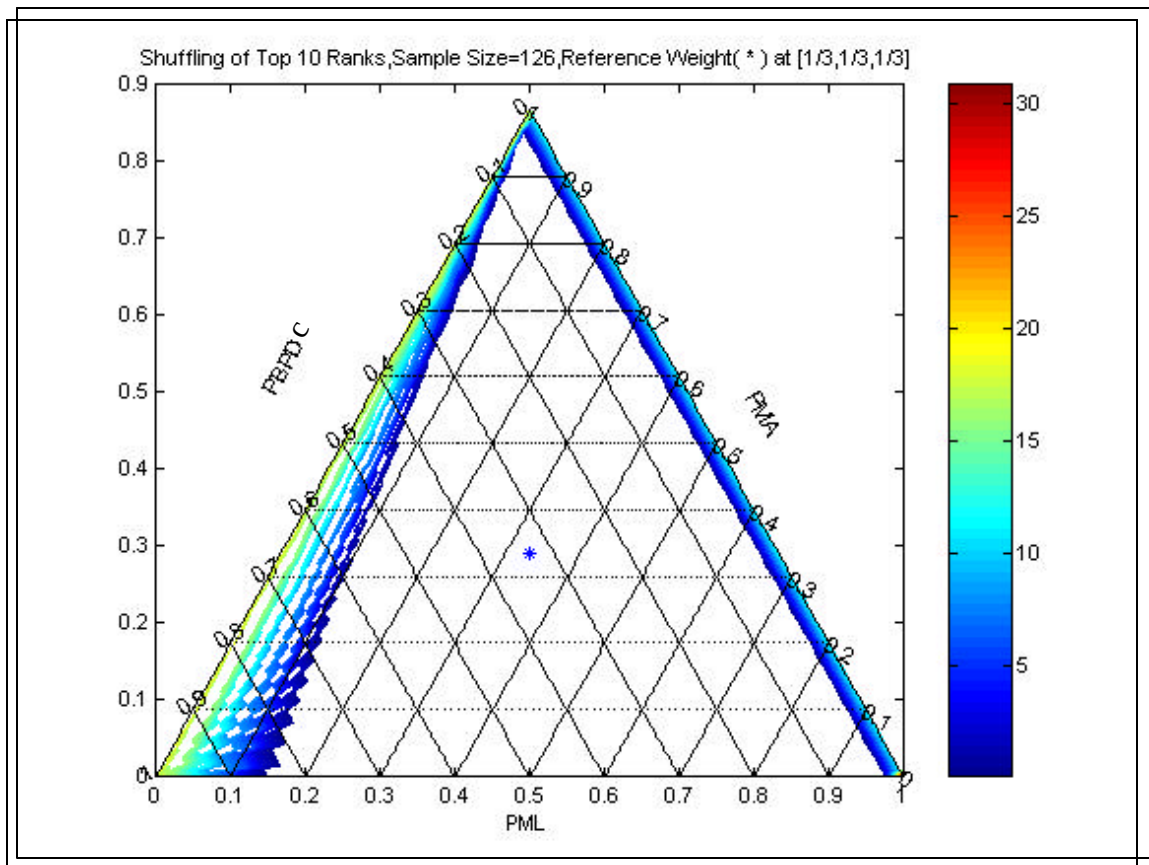


FIGURE 5.40: Contour Plot of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the Top 10 Ranked Candidate Profiles. Reference Weight: $[a_1=1/3, a_2=1/3, a_3=1/3]$, Sample Size=126.

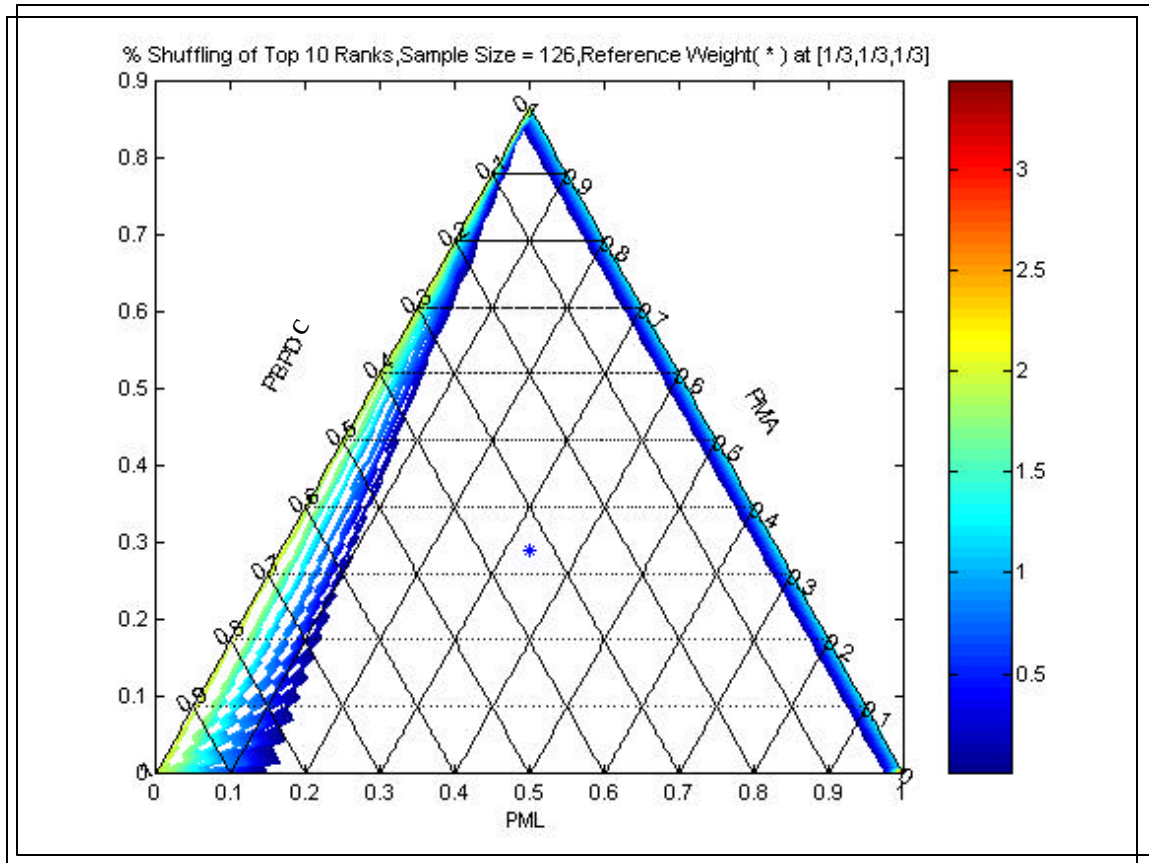


FIGURE 5.41: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the Top 10 Ranked Candidate Profiles. Reference Weight: $[a_1=1/3, a_2=1/3, a_3=1/3]$, Sample Size=126.

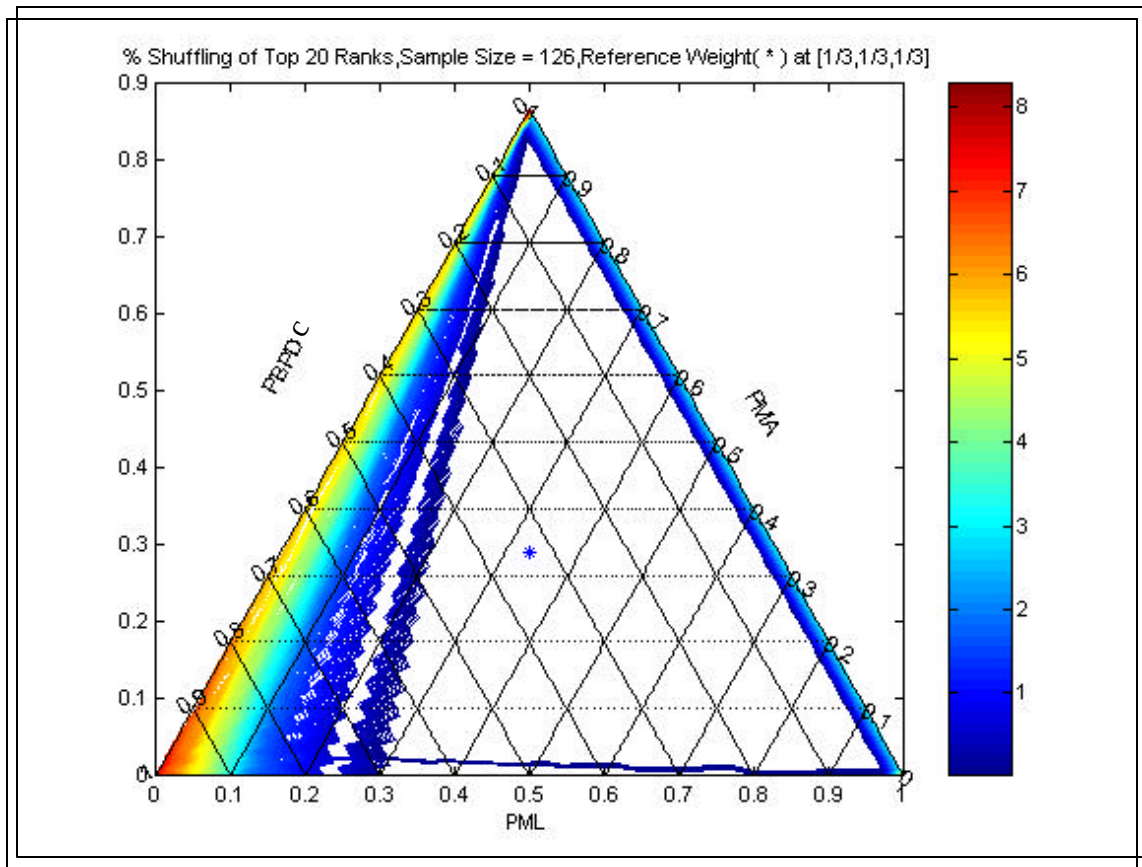


FIGURE 5.42: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the Top 20 Ranked Candidate Profiles. Reference Weight: $[a_1=1/3, a_2=1/3, a_3=1/3]$, Sample Size=126.

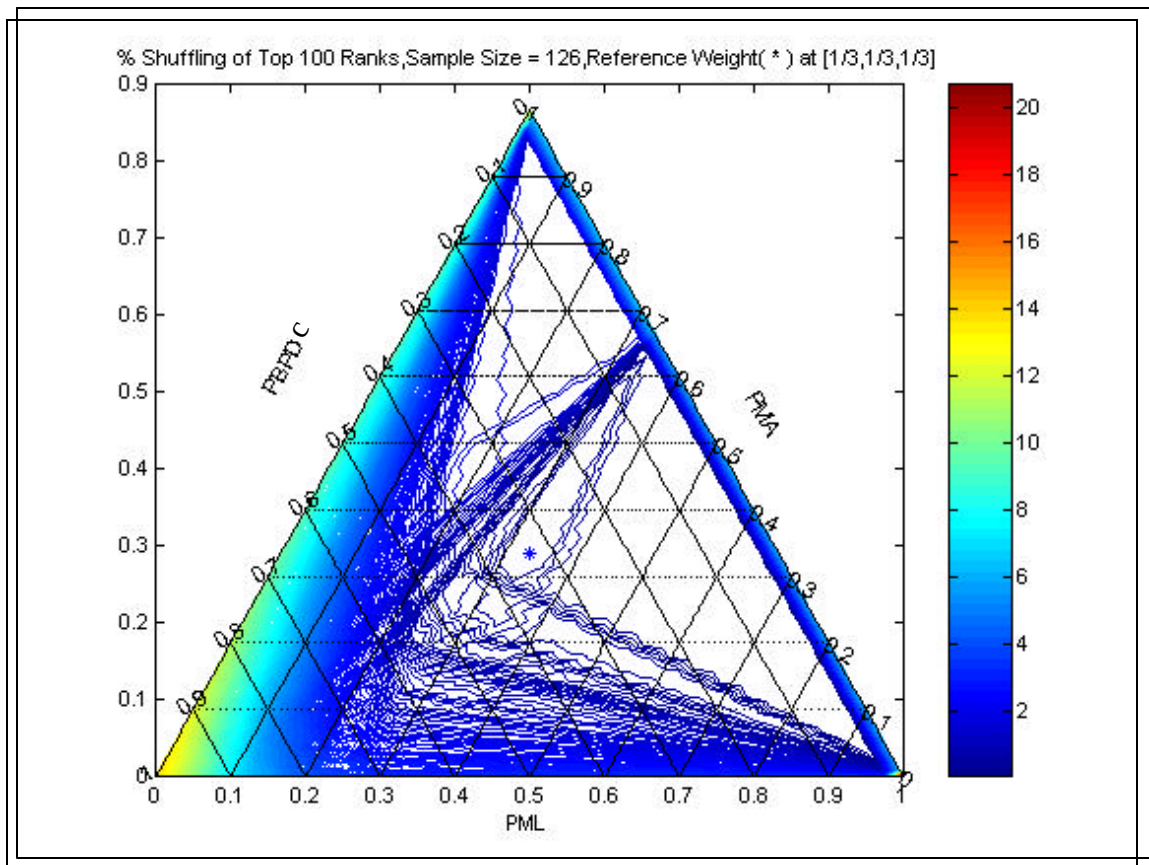


FIGURE 5.43: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the Top 100 Ranked Candidate Profiles. Reference Weight: $[a_1=1/3, a_2=1/3, a_3=1/3]$, Sample Size=126.

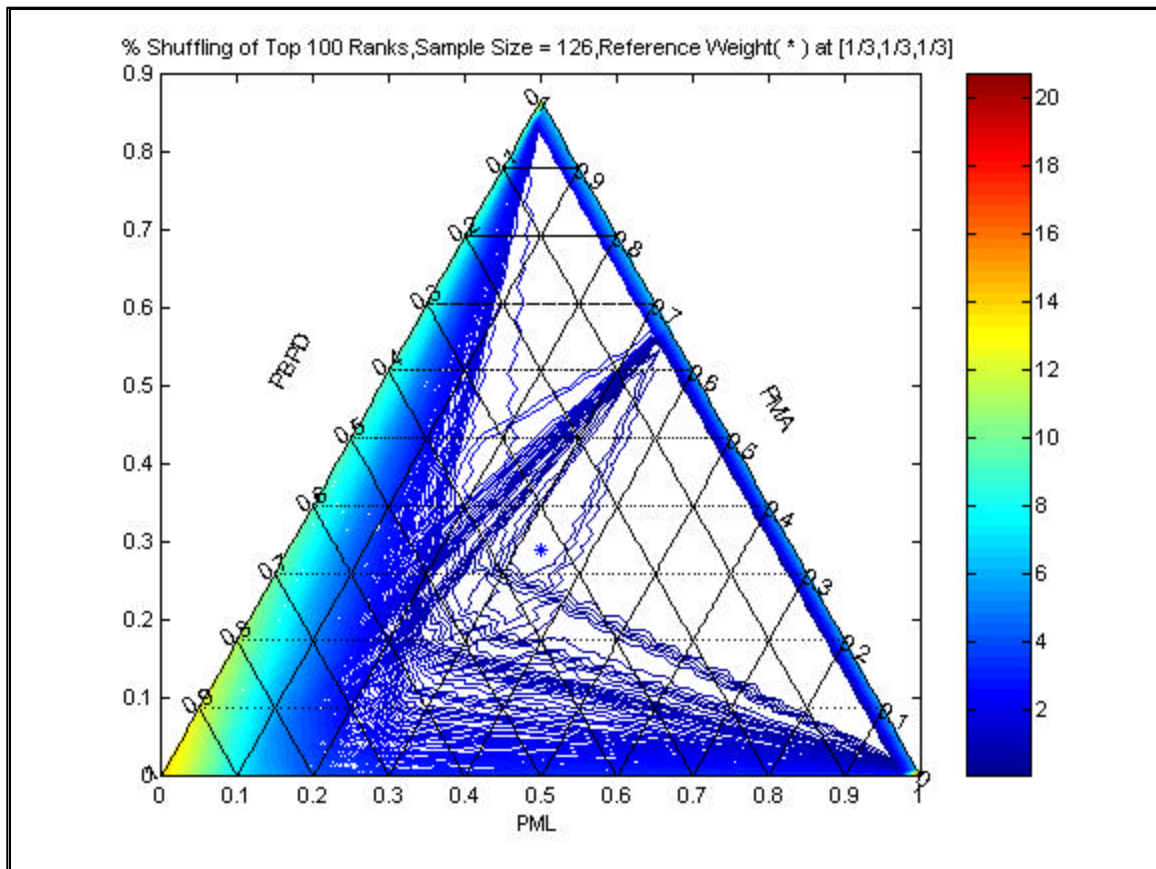


FIGURE 5.44: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the Top 10 Ranked Candidate Profiles. Reference Weight: $[a_1=0.8, a_2=0.1, a_3=0.1]$, Sample Size=126.

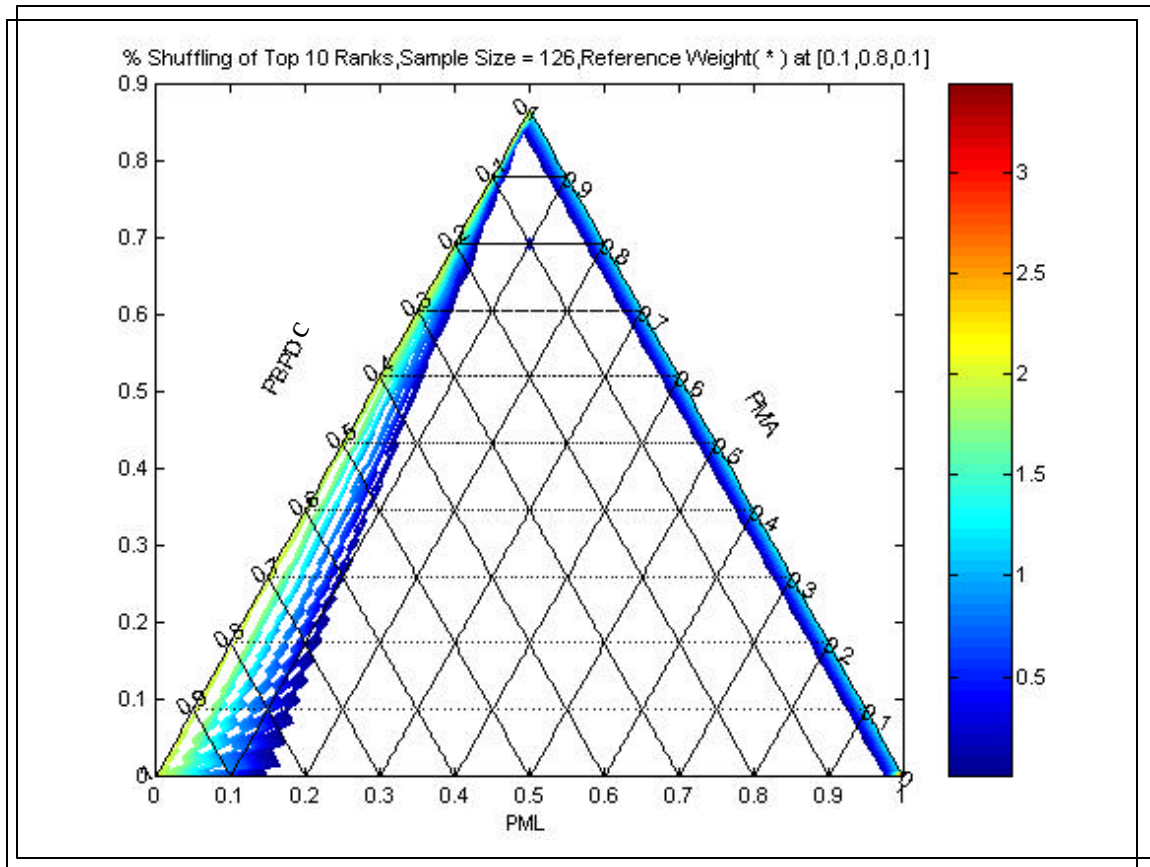


FIGURE 5.45: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the Top 10 Ranked Candidate Profiles. Reference Weight: [$a_1=0.1$, $a_2=0.8$, $a_3=0.1$], Sample Size=126.

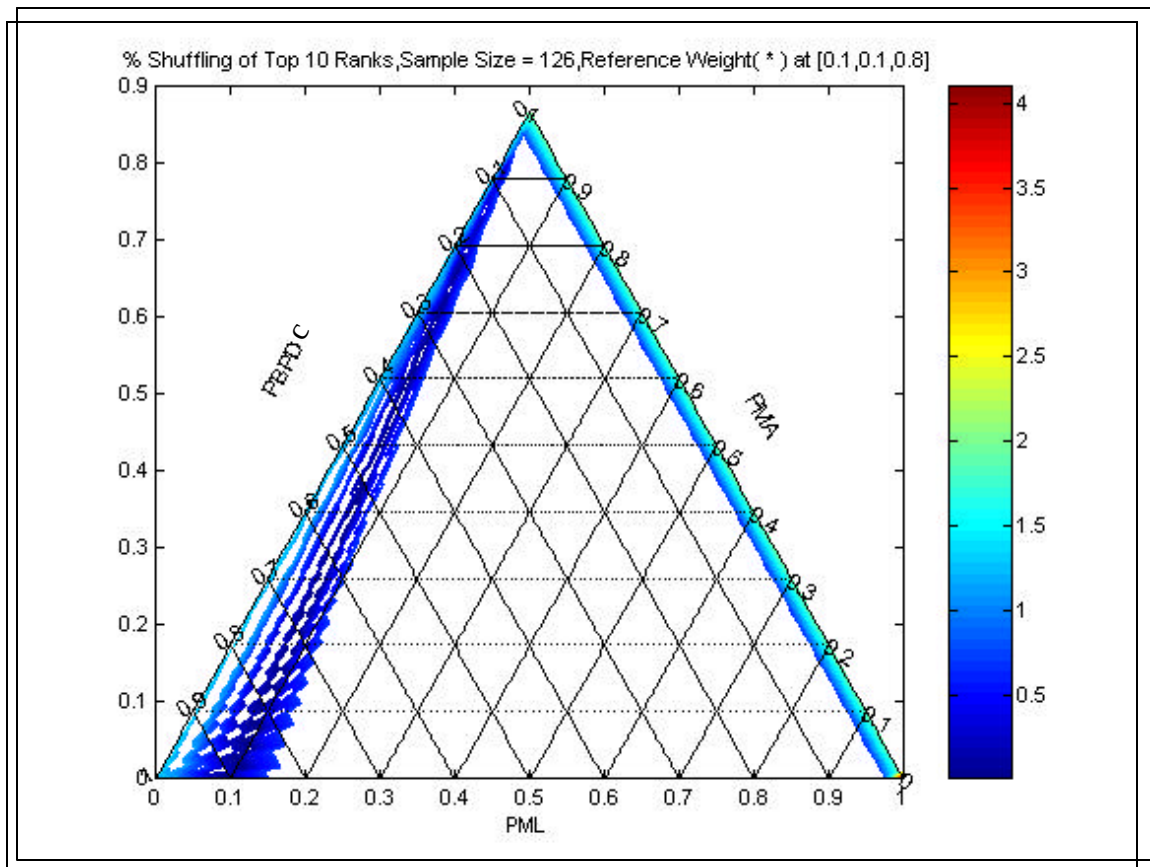


FIGURE 5.46: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the Top 10 Ranked Candidate Profiles. Reference Weight: [$a_1=0.1$, $a_2=0.1$, $a_3=0.8$], Sample Size=126.

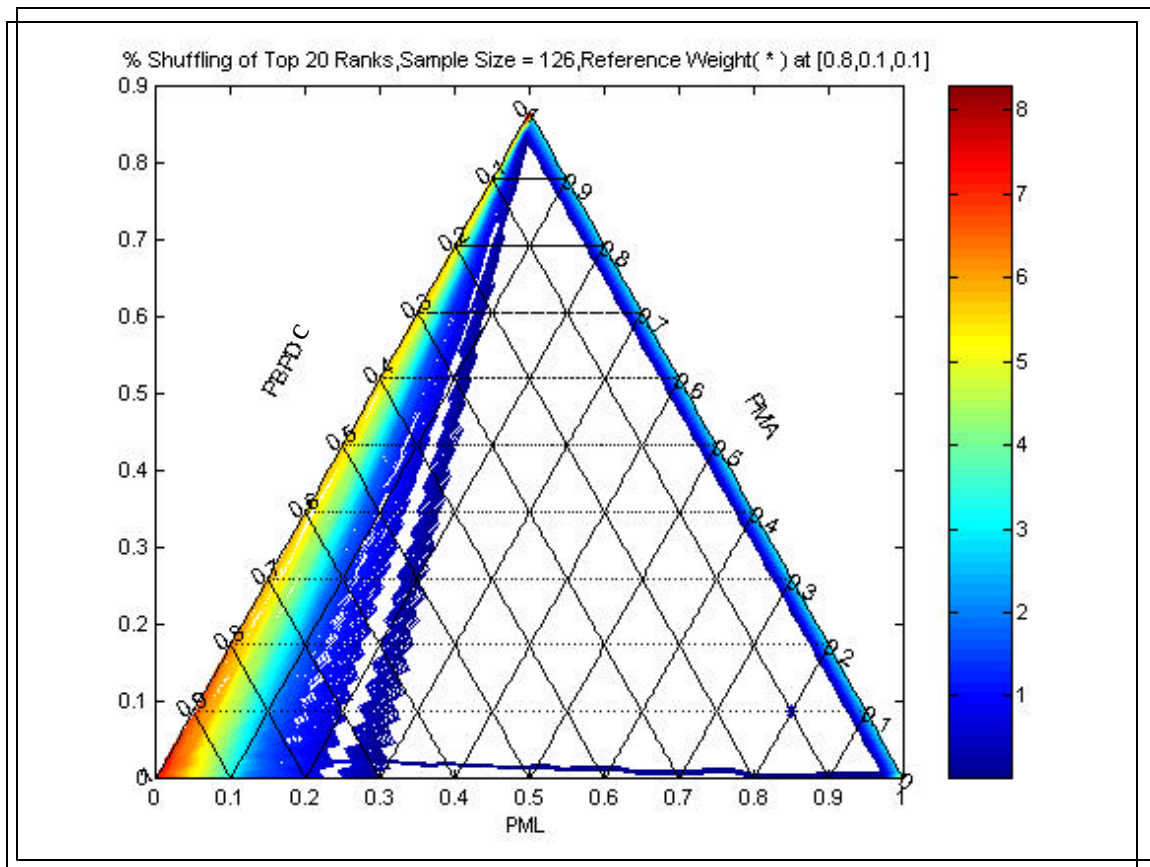


FIGURE 5.47: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the Top 20 Ranked Candidate Profiles. Reference Weight: [$a_1=0.8$, $a_2=0.1$, $a_3=0.1$], Sample Size=126.

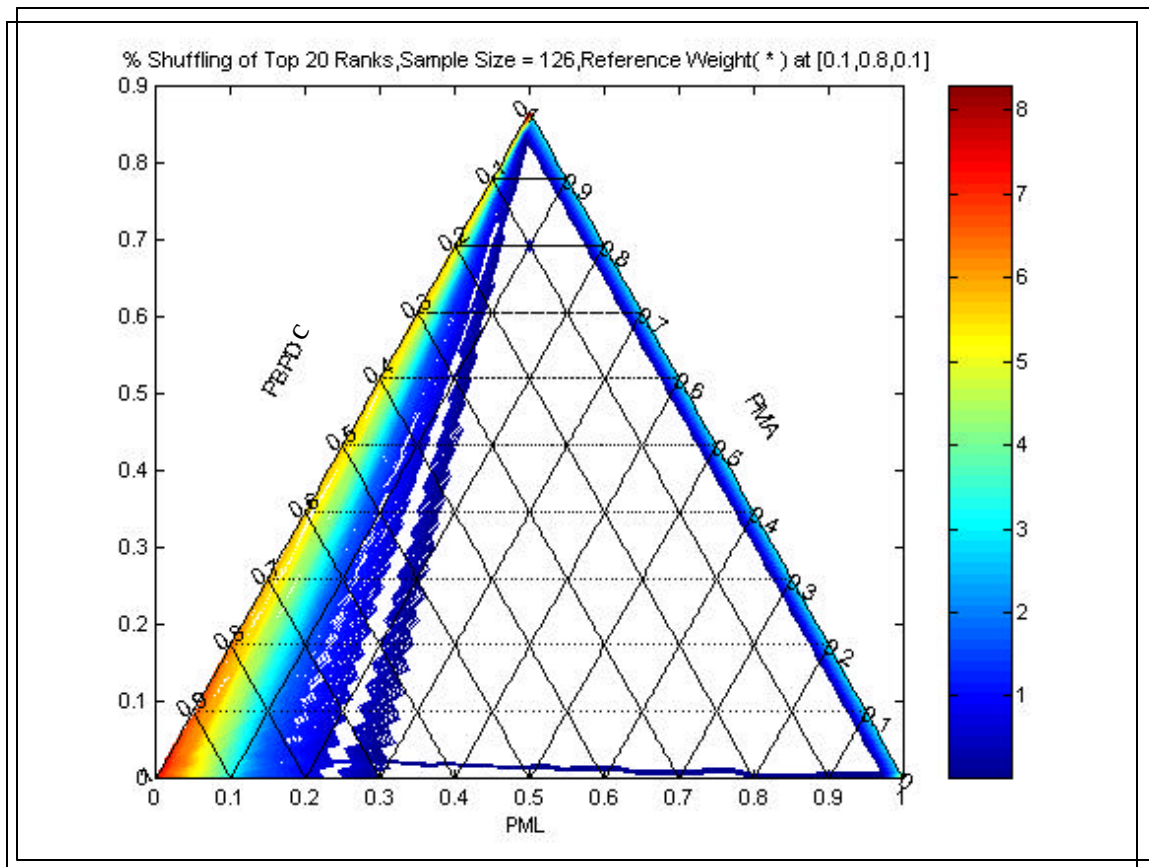


FIGURE 5.48: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the Top 20 Ranked Candidate Profiles. Reference Weight: [$a_1=0.1$, $a_2=0.8$, $a_3=0.1$], Sample Size=126.

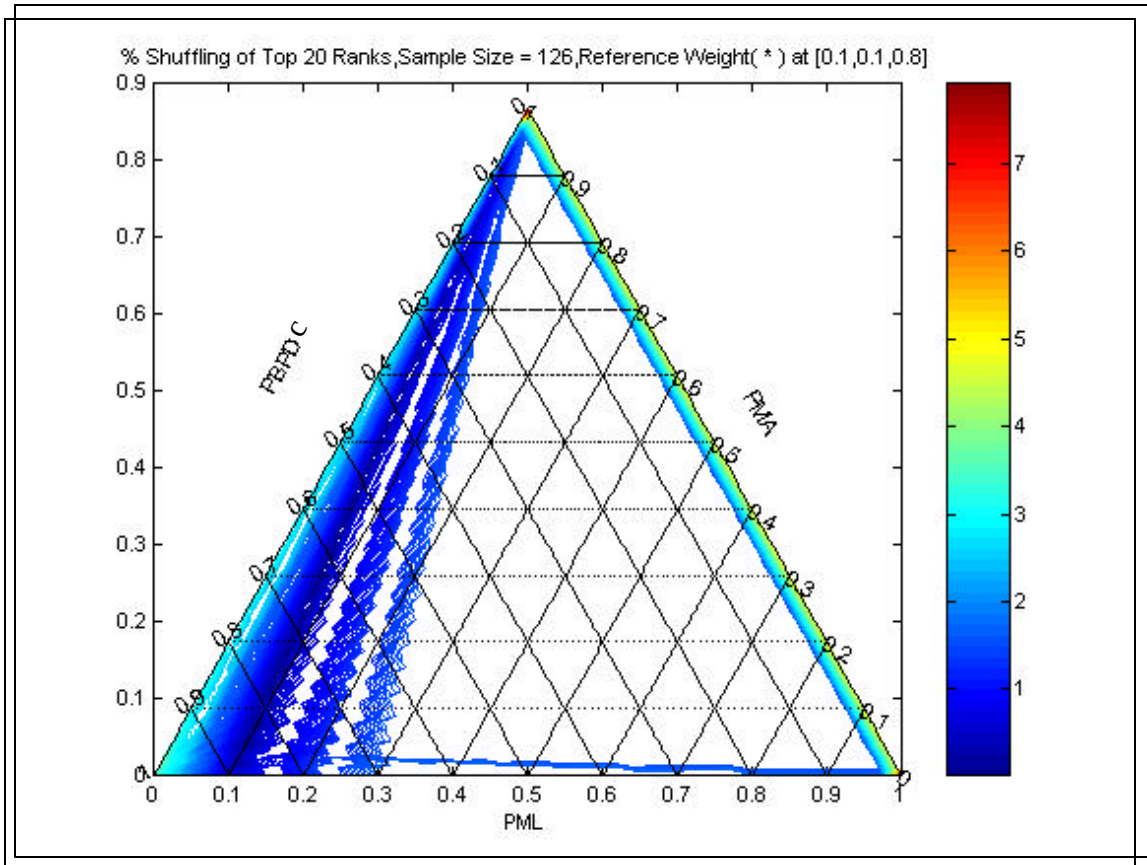


FIGURE 5.49: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the Top 20 Ranked Candidate Profiles. Reference Weight: [$a_1=0.1$, $a_2=0.1$, $a_3=0.8$], Sample Size=126.

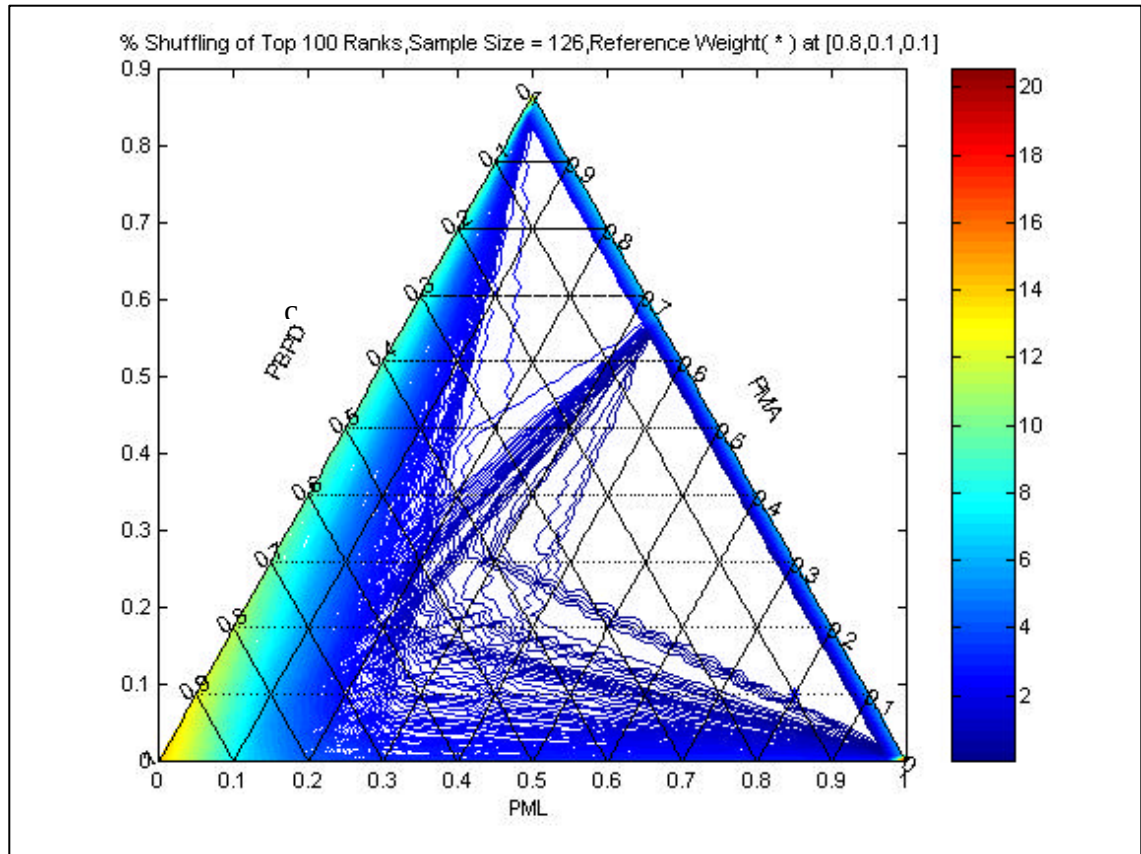


FIGURE 5.50: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the top 100 Ranked Candidate Profiles. Reference Weight: [$a_1=0.8$, $a_2=0.1$, $a_3=0.1$], Sample Size=126.

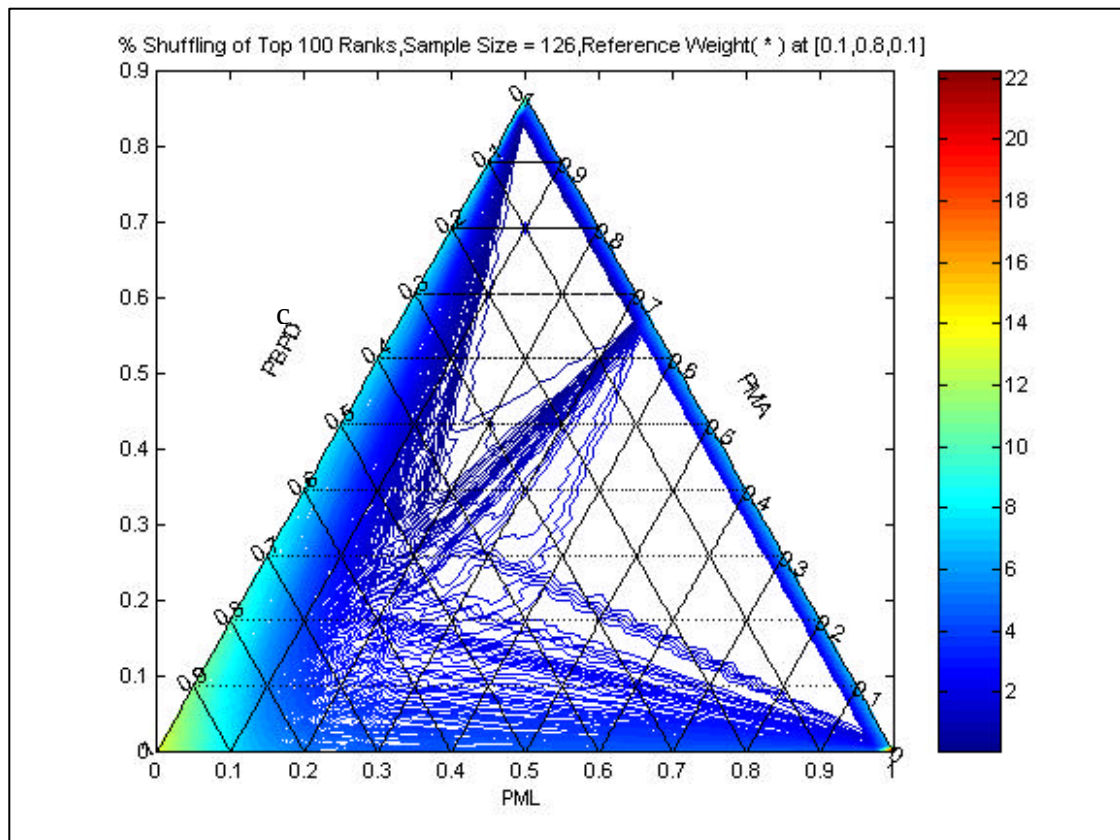


FIGURE 5.51: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the top 100 Ranked Candidate Profiles. Reference Weight: [$a_1=0.1$, $a_2=0.8$, $a_3=0.1$], Sample Size=126.

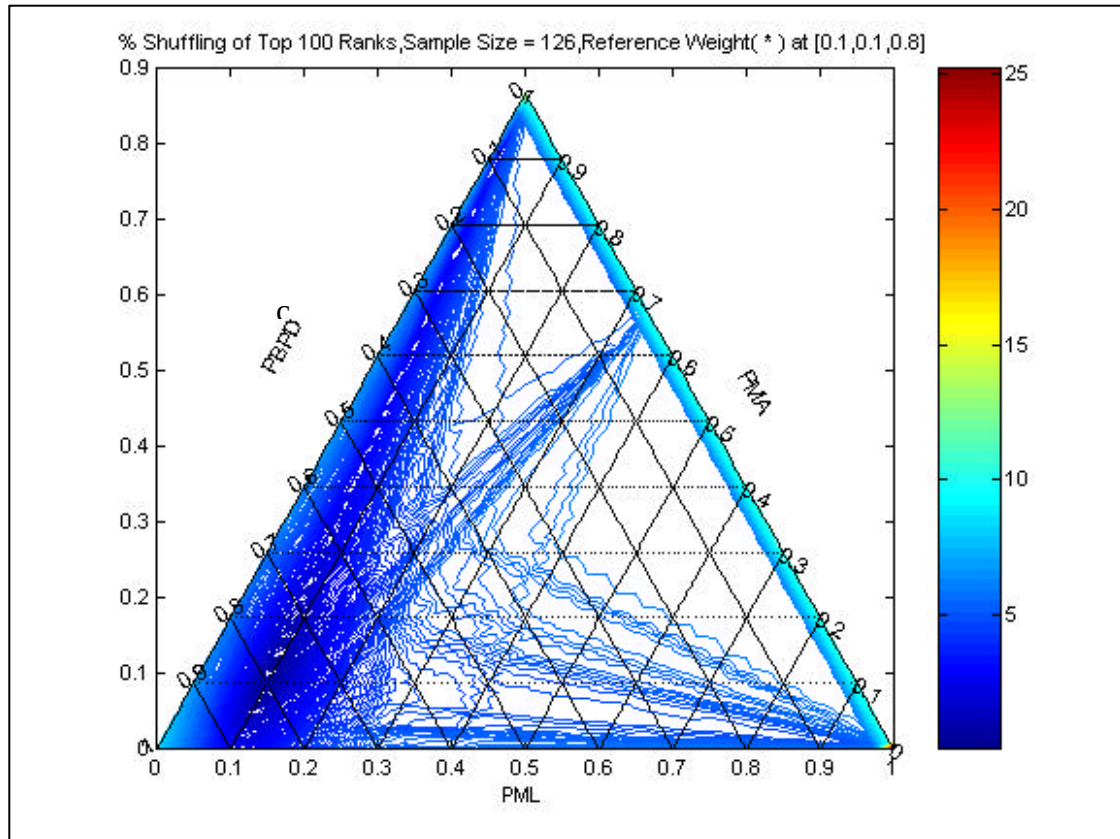


FIGURE 5.52: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the top 100 Ranked Candidate Profiles. Reference Weight: [$a_1=0.1$, $a_2=0.1$, $a_3=0.8$], Sample Size=126.

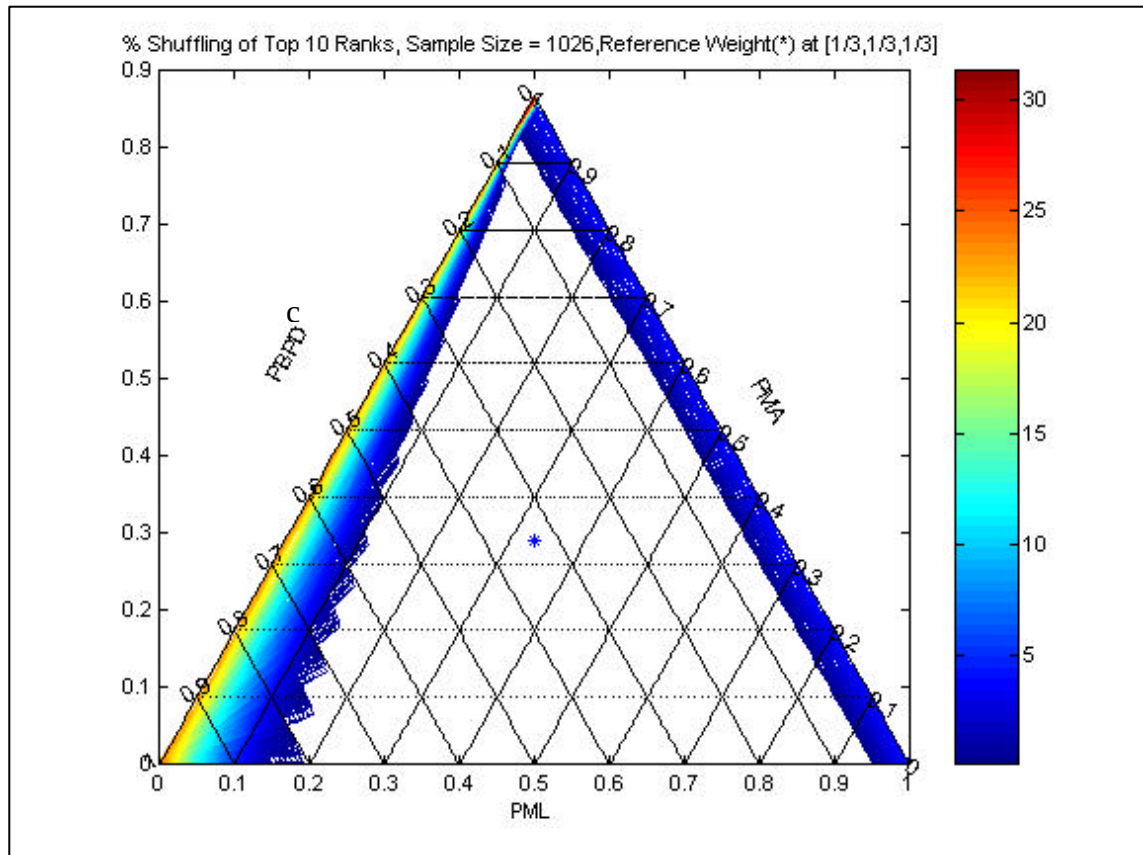


FIGURE 5.53: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the top 10 Ranked Candidate Profiles. Reference Weight: [$a_1=1/3$, $a_2=1/3$, $a_3=1/3$], Sample Size=1026.

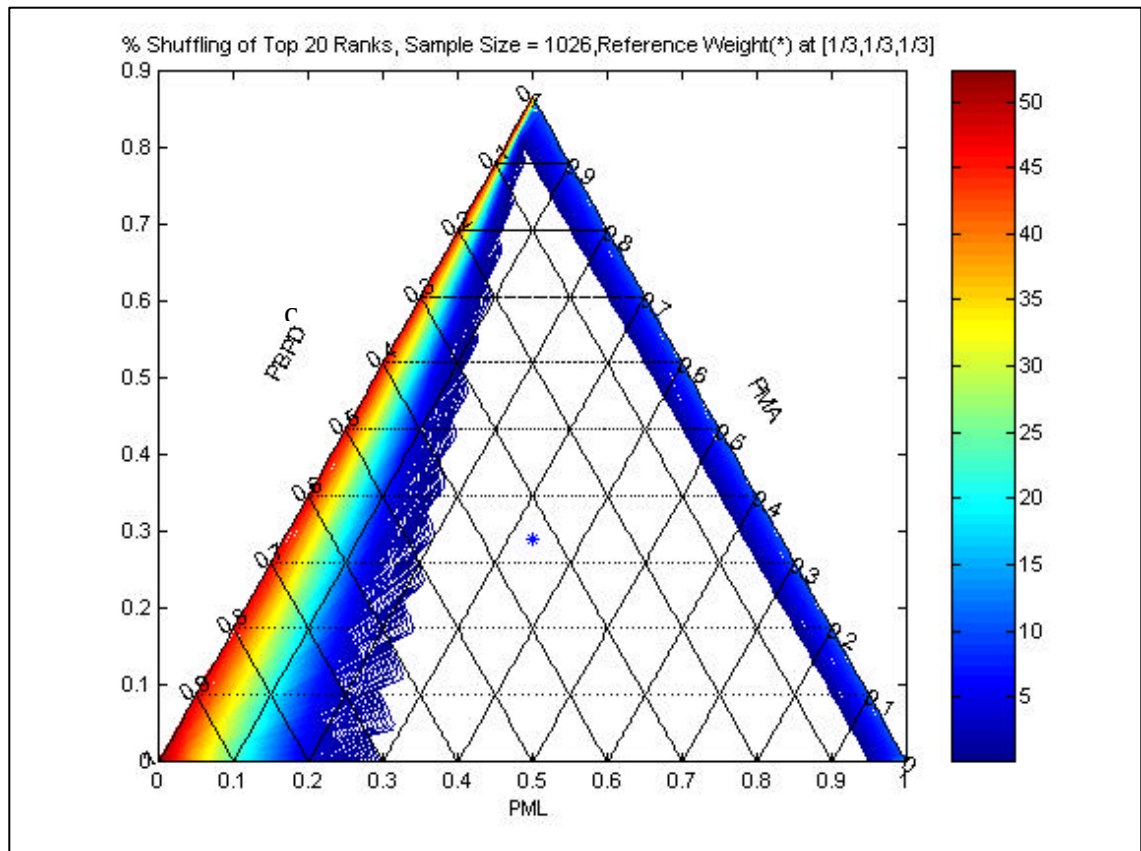


FIGURE 5.54: Contour Plot of Percent of Shuffling Index in Response to Changes in Weights of the Similarity Index Sub Parameters, of the top 20 Ranked Candidate Profiles. Reference Weight: $[a_1=1/3, a_2=1/3, a_3=1/3]$, Sample Size=1026.

CHAPTER 6

SUMMARY, CONCLUSION AND RECOMMENDATIONS FOR FUTURE RESEARCH

A new method has been developed to create and maintain a tree-structured index to multidimensional data using naturally occurring patterns and clusters within the data, and thereby allows the implementation of efficient search and retrieval strategies in a database. A set of 10,000 DNA/STR profiles based on the STR allele probability distribution density for the Caucasians was generated and analyzed using Multivariate Statistical analysis, in specific; the Principal Component Analysis (PCA) to detect clustering patterns among the profiles for the sixteen loci. The analysis revealed that with the choice of some loci-pairs (such as d13s17 and d16s539) good and distinct clusters were obtainable. A total of 120 ($C_2^{16} = 16 \cdot 15 / 2 = 120$) two-loci combinations were analyzed using PCA to decipher the clusterability of the profiles under consideration.

Members within each distinct cluster were further studied to determine the attributes that made them distinct from all members of other clusters. K-means clustering algorithm was used to identify and validate the memberships of each cluster formed by the DNA information contained in a loci pair using PCA.

Cluster formation and differentiation is due to the nature of the allele distribution

pattern at the dominant alleles of the loci pair. In addition, it is seen that loci with allele probability densities concentrated at just a few (2 or 3) alleles give rise to principal components that are loaded heavily with these dominant alleles, which, in turn, would lead to the formation of good and distinct clusters.

The PCA analysis results with a real DNA/STR dataset also showed similar clustering patterns. Testing the developed PCA model on the real DNA data served as a partial validation of the clustering results. The testing was done to see whether PCA clusters from these real data would coincide with the same clusters obtained from the synthetic data, and vice versa. Thus it can be concluded that the real data set and the synthetic data set have comparable allele frequency distributions at these two loci, and that results from clustering studies using synthetic data can be applied directly to real DNA/STR data with the same clustering patterns.

Using synthetic data, it has been established that DNA/STR profiles can be partitioned into distinct clusters using the PCA approach. Successive partitioning using a different two-loci combination at each node level would reduce the number of profiles that have to be searched through for matching profiles, and thereby improving the search efficiency by about one order of magnitude through each level of the decision tree. The construction of a tree with 'n' levels will reduce the number of profiles to be searched through at the n^{th} level by about n orders of magnitude.

A new Similarity Index (SI) parameter has been developed that reflects the degree of similarity of a candidate's profile to that of the target. The SI was successfully tested on a small and a large datasets. Further, a Shuffling Index was developed to study the sensitivity analysis, using contour plots of the SI to the selection of weights for the similarity index sub-parameters.

The incorporation of multivariate statistical clusters to partition the database at each level of the search decision tree would be the natural next step in this research. Efficiency of using the mode of partitioning on the search will be compared to that currently employed in the new UTK search engine. It is envisioned that a combination of the two methods of partitioning will be employed in the final implementation.

In addition, a modified SI would be developed to rank order DNA profiles with unequal number of loci to that of target profile, a common occurrence in the database.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] Hillis, D. M., Craig Moritz, and Barbara K. Mable, *Molecular Systematics*, second edition, Sinauer Association, 1996.
- [2] Chakraborty, R., and Li Jin., "Determination of Relatedness Between Individuals Using DNA- fingerprinting," *Human Biology*, 1993, 65:875-895.
- [3] Papiha, S.S., Chahal, S.M.S., Roberts, D.F., Murty K.J.R., Gupta, R.L., and Sidhu, L.S., "Genetic Differentiation and Population Structure in Kinnar District, Himachal Pradesh, India," *Human Biology*, 1984, 56,2: 231-257.
- [4] Lynch, M., "Analysis of population genetic structure by DNA fingerprinting", 1991, pp. 113-126. In T. Burke, G. Dolf, A. J. Jeffreys, and R. Wolff (eds.) *DNA Fingerprinting: Approaches and Applications*. Birkhäuser Verlag, Basel.
- [5] Chakraborty, R., Ranjan Deka, Li Jin, and Robert E. Ferrell, "Allele Sharing at Six VNTR Loci and Genetic Distances Among Three Ethnically Defined Human Populations," *Human Biology*, 1996, 4:387-397.
- [6] Ferak, V., Sivakova, D., and Kroupova, Z, "Genetic Distances, Geographic Distances and Migration between Four Villages of a Single Region in Slovakia," *Journal of Human Evolution*, 1980, 9:573-581.
- [7] Edwards, A.W.F, "Distances between Populations on the Basis of Gene Frequencies". *Biometrics*, 1971, 27:873-881.
- [8] Balakrishnan, V. and Sanghvi, L.D., "Distance between Populations on the Basis of Attribute Data," *Biometrics*, 1968, 24:859-865.

- [9] Cavalli-Sforza, L.L. and Edwards, A.W.F., "Phylogenetic Analysis: Models and Estimation Procedures," *Evolution*, 1967, 21:550-570.
- [10] Chakraborty, R., Stivers, D.N., Budowle, B., and Zhong, Y., "The Utility of Short Tandem Repeat Loci beyond Human Identification: Implications for Development of New DNA Typing Systems," *Electrophoresis*, 1999, 20:1682-1696.
- [11] Robinson, S., Gutowski, S., and Mitchell, J., "Genetic Diversity among Selected Ethnic Sub-populations of Australia: Evidence from Three Highly Polymorphic DNA Loci," *Human Biology*, 1996, 68,4: 489-508.
- [12] Nei, M., "Identity of Genes and Genetic Distance between Populations," *Genetics*, 1971, 68:s47.
- [13] Lynch, M., "Estimation of Relatedness by DNA Fingerprinting," *Molecular Biology and Evolution*, 1988, 5(5) c 584-599.
- [14] National Research Council, "The Evaluation of Forensic DNA Evidence," An update, 1996.
- [15] Robertson, J., Ross, A.M., and Burgoyne, L.A., *DNA in Forensic Science, theory techniques and applications*, Ellis Horwood Series in Forensic Science, 1997.
- [16] Inman, K., Rudin, N, *An Introduction to Forensic DNA Analysis*, CRC Press, 1997.
- [17] Budowle, B., Giusti, A.M., Waye, J.S., Baechtel, F.S., Fourney, R.M., Adams, D.E., Presley, L.A., Deadman, H.A., and Monson, K.L., "Fixed-Bin Analysis for Statistical Evaluation of Continuous Distributions of Allelic Data from VNTR Loci, for use in Forensic Comparisons," *Am. J. Hum. Genet.*, 1991, 48:841-855.

- [18] Budowle, B., Baechtel, F.S., Smerick, J.B., Presley, K.W., Giusti, A.M., Parsons, G., Alevy, M.C., and Chakraborty, R., "DIS80 Population Data in African Americans, Caucasians, Southeastern Hispanics, Southwestern Hispanics, and Orientals," *Journal of Forensic Science*, JFSCA, 1995, 40,1: 38-44.
- [19] Budowle, B., "The Effects of Inbreeding on DNA Profile Frequency Estimates using PCR- based Loci," *Genetica*, 1995, 96:21-25.
- [20] Chakraborty, R., and Weiss, K.M., "Genetic Variation of the Mitochondrial DNA Genome in American Indians Is at Mutation-Drift Equilibrium," *American Journal of Physical Anthropology*, 1991, 86:497-506.
- [21] Chakraborty, R., Srinivasan, M.R., and Daiger, S.P., "Evaluation of Standard Error and Confidence Interval of Estimated Multilocus Genotype Probabilities, and Their Implications in DNA Forensics," *Am. J. Hum. Genet.*, 1993, 52:60-70.
- [22] Lynch, M., "The Similarity Index and DNA Fingerprinting," *Molecular Biology and Evolution*, 1990, 7,5:478-484.
- [23] Lynch, M., "The Analysis of Population Survey Data on DNA Sequence," *Molecular Biology and Evolution*, 1990, 7,4:377-394.
- [24] Lynch, M., "Methods for the Analysis of Comparative Data in Evolutionary Biology," *Evolution*, 1991, 45,5:1065-1080.
- [25] Lynch, M., and Jarrell, P.E., "A method to Calibrating Molecular Clocks and its Applications to Animal Mitochondrial DNA," *Genetics*, 1993, 135:1197-1208.
- [26] Gill, P., and Evett, I., "Population Genetics of Short Tandem Repeat (STR) Loci," *Genetica*, 1995, 96:69-87.

- [27] Donnelly, P., "Match Probability Calculations for Multi-locus DNA Profile," *Genetica*, 1995, 96:55-67.
- [28] Masatoshi, Nei, "Estimation of Average Heterozygosity and Genetic Distance of Individuals," *Genetics*, 1978, 89:583-590.
- [29] Al-Daoud, M.B., and Roberts, S.A., "New Methods for the Initialization of Clusters," *ELSEVIER / Pattern Recognition Letters*, 1996, 17:451-455.
- [30] Feucht, D., "Pattern Recognition: Basic Concepts and Implementations," *Computer Design/December*, 1997, 57-68.
- [31] Shum, H.Y., Ikeuchi, K., and Reddy, R., "Principal Component Analysis with Missing Data and Its Applications to Polyhedral Object Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995, 17,9:854-865.
- [32] Berry, B.J.L., "A Note Concerning Methods of Classification," *Annals of Eugenics*, 1958, 300-303.
- [33] Coffman, G.A., Mezzich, J.E., and Strauss, J.S., "Clustering and Visualization of Patterns in a Psychiatric Database," *IEEE*, 1980.
- [34] Herrin, George, Jr., "Probability of Matching RFLP Patterns from Unrelated Individuals," *Am. J. Hum. Genet.*, 1993, 52:491-497.
- [35] Chakraborty, R., and Li, Zhaojue, "Correlation of DNA Fragment Sizes within Loci in the Presence of Non-detectable Alleles," *Genetica*, 1995, 96:27-36.
- [36] Clark, A.G., Hamilton, F, and Chambers, G.K., "Inference of Population Subdivision from VNTR Distributions of New Zealanders," *Genetics*, 1995, 96:37-49.

- [37] Maiste, P.J., and Weir, B.S., "A Comparison of Tests for Independence in the FBI RFLP Databases," *Genetica*, 1995, 96:125-138.
- [38] Birdwell J.D., Wang T.W., Horn R.D., Yadav Puneet, and Icové D.J., "Method of Storage and Retrieval of Multidimensional Information," U.S. Provisional Patent Application, 04842.83534, September 1999.
- [39] Richard Willing, USA TODAY, Arlington, Va., Oct. 12, 1998.
- [40] Tou, J.T., and Gonzalez, R.C., *Pattern Recognition Principles* Addison Wesley, 1974.
- [41] CODIS 5.1 GDIS Searching Specification (Draft), U.S. Department of Justice Federal Bureau of Investigation, July 23, 1998.
- [42] Isenberg, A. R., and Moore, J. M., "Mitochondrial DNA Analysis at the FBI Laboratory", *Forensic Science Communications*, July 1999, 1:2.
- [43] Niezgoda S.J., and Brown Barry, "The FBI Laboratory's COMbined DNA Index System Program" <http://www.promega.com/geneticidproc/ussymp6proc/neizgod.htm>
- [44] Crow, J. F., and Denniston C., *Genetic Distance*, Plenum Press, New York and London, 1974.
- [45] Strang, G., *Linear Algebra and its Applications*, 2nd ed., Academic Press, New York, 1980.
- [46] Treybal, R. E., *Mass Transfer Operations*, 3rd ed., New York, McGraw-Hill, c1980.
- [47] McCabe, W.L., and Smith, J. C., *Unit operations of chemical engineering*, New York, McGraw-Hill, 1956.

- [48] Massart, D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y., and Kaufman, L., *Chemometrics: a textbook*, 2nd ed., Elsevier Science Publishing Company, New York, 1989.
- [49] Budowle, B., and Moretti, T. R., “Genotype Profiles for Six Population Groups at the 13 CODIS Short Tandem Repeat Core Loci and other PCR Based Loci,” *Forensic Science Communications*, FBI Laboratory Division Publication 99-06, U.S. Department of Justice, Federal Bureau of Investigation, July 1999, V.1, n.2.
- [50] Berry, M. W., Drmac, Z., and Jessup, E., “Matrices, Vector Spaces, and Information Retrieval,” *SIAM Review* 41:335-362, 1999.
- [51] Green, P. E., Carroll, J. D., *Mathematical Tools for Applied Multivariate Analysis*, student ed., Academic Press, San Diego, 1978.
- [52] Balding, D. J., and Nichols, R. A., “A Method for Quantifying Differentiation Between Populations at Multi-Allelic Loci and its Implications for Investigating Identity and Paternity,” *Genetica*, V. 96, n. 1-2, June 1995.
- [53] Lynch, M., and Ritland, K., “Estimation of Pairwise Relatedness with Molecular Markers,” *Genetics*, V. 152, n. 4:1753-66, August 1999.
- [54] Martin, E. B., and Morris, A. J., “An Overview of Multivariate Statistical Process Control in Continuous and Batch Process Performance Monitoring,” *Transactions of the Institute of Measurement and Control*, V.18 n.1: 51-60, 1996.
- [55] Kourti, T., and MacGregor, J. F., “Process Analysis, Monitoring and Diagnosis, using Multivariate Projection Methods,” *Chemometrics and Intelligent Laboratory Systems*, V. 28, n. 3-21, 1995.

- [56] Wise, B. M., and Gallagher N. B., “ The Process Chemometrics Approach to Process Monitoring and Fault Detection,” *Journal of Process Control*, V. 6, n. 6:329-348, 1996.
- [57] Griffin, H. G., and Griffin, A. M., *PCR Technology: Current Innovations*, CRC Press, 1994.
- [58] Morrison, R.D., “E-Gels Allow DNA Results in 35 Minutes,” *Law Enforcement Technology*, August 1999, pp: 88.
- [59] Shannon, E., “DNA: Putting Bad Guys Away Too,” *Time Magazine*, September 13, 1999, pp: 28.
- [60] Casey, D.K., “Genes, Dreams, and Reality: The Promises and Risks of the New Genetics,” *Judicature Genes and Justice*, November-December 1999, V. 83(3).
- [61] Asplen, C.H., “From Crime Scene to Courtroom Integrating DNA Technology into the Criminal Justice System,” *Judicature Genes and Justice*, November-December 1999, V. 83(3).
- [62] Strange, C., “DNA’s Search for Truth,” *Judicature Genes and Justice*, November-December 1999, V. 83(3).

VITA

Puneet Yadav graduated First Division with Honors in May 1998, with a B.E. (Bachelor of Engineering) in Chemical Engineering, from University of Roorkee, India.

Puneet was awarded a graduate research assistantship to pursue a Master of Science in Chemical Engineering, at the University of Tennessee, Knoxville. After graduating in May 2001, he has taken employment with LAM Research, Fremont, CA.