Combined evidential value of forensic DNA profiles

A thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfillment of the requirements
for the degree

MASTER OF SCIENCE
in
APPLIED MATHEMATICS

by

Jacob de Zoete

Delft, the Netherlands
March 2012

Copyright © 2012 by Jacob de Zoete. All rights reserved.
Abstract

In forensic science it is common practice to work on problems where the likelihood ratio is based on observing single piece of evidence given two hypotheses. However, in a lot of cases, there is more than one piece of evidence. In this thesis three problems regarding combining DNA profiles are discussed. First, we derived a method to combine the evidential value of different partial Y-chromosomal DNA profiles of different stains. The method consists of finding a lower bound for the likelihood ratio when more than two propositions are compared and where we don’t need the prior probabilities of the different propositions. Second, we made a simulation model to investigate the dependence of autosomal and Y-chromosomal DNA profiles by assuming that everybody with the same Y-chromosomal DNA profile has a common ancestor and simulating the assignment of autosomal DNA profiles over different family tree structures. The results can be used in practice as scientific support for the assumption of independence between the Y and autosomal profile. Third, we developed a model to interpret (low-template) DNA profiles which is able to give the likelihood of observing the DNA profile given any allele-combination of the donor. This model assumes that a DNA profile is a result of a stochastic process where the input are the alleles of a possible donor. This model uses the information in peak heights without using any threshold. The model shows promising results in determining the combined evidential value of several low template DNA profiles that were obtained from the same stain.
Preface

This master thesis is not only the result of a final project, it is also the last part of six exiting years as a student. Six years ago, I started with my study in Mathematics. For six years I lived with the nicest people, first with the family Janssen in Nootdorp and after that, with my fellow students. I would like to thank Chris, Elwin, Teun, Jarno and Paolo for the great time we had at the Markt. Now, at the end of my study, I am proud and happy to be living with my girlfriend Charlotte.

I want to thank my parents for being enthusiastic and supportive in everything I did or wanted to do. Towards my brothers, Thomas and Karel, your proud brother would like to thank you for the interest and support you showed in the things I did.

I did this master thesis at the Statistics department of the Netherlands Forensic Institute (NFI). I would like to thank them for the educational environment and the interesting conversations during the last nine months. In this light, I especially want to mention Laurens Grol and Hinda Haned (NFI) who provided me with data to use in my models, and Sharda and Yvonne for the nice time we had as roommates. My thanks go out to Peter de Knijff from Leiden University who provided allele frequencies I used in my simulation model. I want to thank my study friends Linda, Gemma and Frank for the nice time we had working on our thesis projects. I also want to thank Teun, Joost and Christiaan for commenting me on my writing. The most important word of thanks goes out to my girlfriend Charlotte for the nice times and the support you gave me during this thesis project.

Lastly, I want to thank Eric and Marjan for supervising my during this project. I enjoyed working with Eric, who surprised me each week with other ways of solving or looking at problems. His enthusiasm on the problems we worked on inspired me during the past nine months. Marjan showed me the importance of mathematics as a tool to find relevant answers to interesting questions. She learned me that the outside world is more interested in what to do than in how to come to the decision on what to do. I felt privileged working with them and feel privileged that I will be working with Marjan for the coming four years.

Jacob Coenraad de Zoete
Delft, the Netherlands
May 25, 2012
## Contents

1 Introduction 10

2 Forensic Statistics 12
   2.1 Bayes’ Theorem 12
      2.1.1 The Likelihood Ratio 13

3 The combination of partial Y-chromosomal DNA profiles 14
   3.1 A method to determine a likelihood ratio when multiple partial Y-chromosomal DNA profiles are found 14
      3.1.1 Hypotheses 14
      3.1.2 $P(E|H_i)$ 16
      3.1.3 Conclusion 17
   3.2 Methods based on the previous method 18
      3.2.1 Selecting the priors 18
      3.2.2 Finding a lower bound for the likelihood ratio 20
      3.2.3 Conclusion 21

4 The combination of autosomal and Y-chromosomal DNA profiles 22
   4.1 Introduction 22
   4.2 DNA profiles 22
      4.2.1 Introduction 22
      4.2.2 DNA profiles 23
   4.3 Model 24
      4.3.1 Introduction 24
      4.3.2 Assumptions 25
      4.3.3 Population substructures 26
      4.3.4 The model 29
   4.4 Data 35
      4.4.1 Probability of dying 35
      4.4.2 Son-probability 35
      4.4.3 Allele frequencies 35
   4.5 Results 36
      4.5.1 General results 36
      4.5.2 A profile on three loci 38
      4.5.3 The profile probability 41
   4.6 Conclusion and Discussion 44

5 The combination of low template DNA profiles 46
   5.1 Introduction 46
      5.1.1 What is Low Template DNA? 46
   5.2 Procedure 46
      5.2.1 Obtaining DNA from an object 47
      5.2.2 Amplifying the DNA using PCR 47
      5.2.3 Producing the DNA profile 47
List of Figures

4.3.1 Representation of $\lambda$ .................................................. 30
4.3.2 A family tree .............................................................. 32
4.5.1 The size of the close family against the number of times we have seen zero matches 37
4.5.2 Estimated probability for the number of matching profiles for 10 trees, 100 men ...... 39
4.5.3 Estimated probability for the number of matching profiles for 10 trees, 1000 men .... 40
4.5.4 A histogram showing the $\log_{10}$ of our random match probability of all the trees . 42

5.2.1 Proposed mechanism to explain the formation of stutters ................................ 48
5.2.2 An example of a DNA profile .......................................... 49
5.5.1 A DNA profile on TH01 .................................................. 59
5.6.1 A plot of recorded stutter on different loci .................................. 67
5.6.2 A plot of stutter percentages on the different alleles of locus D10 ....................... 67
5.6.3 The probability density function of the contamination ....................................... 69
5.9.1 A histogram of the log of the simulated likelihoods for the allele combination 16-16 on vWA 79
5.9.2 A histogram of the log of the simulated likelihoods for the allele combination 7-7 on TH0 80
5.9.3 A histogram of the log of the simulated likelihoods for the allele combination 9-12 on D16 80
5.9.4 A histogram of the log of the simulated likelihoods for the allele combination 9-12 on D2S1338 ............................................................. 80

B.2.1 The estimated probability for the number of matching profiles for 20 trees without close family .......................................................... 95
B.2.2 The estimated probability for the number of matching profiles for 10 trees without close relatives .......................................................... 96
B.2.3 The estimated probability for the number of matching profiles for 20 trees without close relatives .......................................................... 97
List of Tables

3.1.1 Hypotheses that explain the findings ................................................................. 14
3.1.2 The notation of the evidence ............................................................ 15
3.1.3 Values for $P(E|H_i)$ ................................................................. 17
3.2.1 All possible hypotheses for a case with three crime stains and one suspect ........ 19
3.2.2 The hypotheses when we only consider situations where the suspect left either all or none of the stains ................................................................. 19
4.3.1 Allele-pairs probabilities after using our equations to account for subpopulations ... 29
4.3.2 Allele-pairs probabilities without sub-population assumption ..................... 29
4.3.3 The suspects profile  ........................................................................... 31
4.3.4 The profiles of the father and mother of the suspects ................................... 31
4.3.5 The profile of the suspect ....................................................................... 32
4.3.6 The profiles of everybody in the paternal line from the suspect to the root of the tree ..... 33
4.3.7 All the profiles of the people in our tree. Persons with the same profile as the suspect (person 12) are highlighted ................................................................. 33
4.3.8 Results after simulating the profile simulation 1000 times ....................... 34
4.4.1 Allele frequencies ........................................................................... 36
4.5.1 The number of simulations with a certain number of matching profiles with and without the close family, extreme case ................................................................. 37
4.5.2 Parameters for the different simulations ................................................. 38
4.5.3 The profile on 3 loci assigned to our suspect ............................................. 38
4.5.4 The mean, maximum and minimum relative frequencies computed over 50 trees with the probabilities for a random unrelated population ................................................................. 40
4.5.5 The mean, maximum and minimum computed over 50 trees .................... 40
4.5.6 The mean and standard deviation of the profile probabilities on 50 trees linking 100 people and profile information on 3 loci when we neglect the close family. ................................................................. 41
4.5.7 The mean and standard deviation of the profile probabilities on 50 trees linking 100 people and profile information on 3 loci when we neglect the close family. ................................................................. 42
4.5.8 The mean and standard deviation of the profile probabilities on 20 trees linking 1000 people and profile information on 4 loci when we neglect the close family. ................................................................. 43
4.5.9 The mean and standard deviation of the profile probabilities on 20 trees linking 10 000 people and profile information on 4 loci when we neglect the close family. ................................................................. 43
4.5.10 The mean and standard deviation of the profile probabilities on 20 trees linking 10 000 people and profile information on 5 loci when we neglect the close family. ................................................................. 43
4.5.11 The peak heights on the different alleles .................................................. 59
4.5.12 The peak heights on the different alleles without the alleles that are not relevant due to our prior distribution ................................................................. 59
4.5.13 Likelihoods for observing the DNA profile from our example ................... 64
4.5.4 Prior probabilities for all allele combinations on locus $THO1$ assuming Hardy-Weinberg equilibrium ................................................................. 65
5.5.1 Allele frequencies for the alleles on locus $D22S1045$ ................................. 66
5.5.2 Prior probabilities for all allele combinations on locus $D22S1045$ ................. 66
5.7.1 Alleles and peak heights of the different loci from our DNA profile ............... 70
5.7.2 The alleles we assume to be belonging to the donor.

5.7.3 The settings we will differ during the sensitivity analysis.

5.7.4 The expected peak heights per locus.

5.7.5 Likelihoods of DNA profile under standard assumptions.

5.7.6 Likelihoods of observing the DNA profiles, normal distribution stutter.

5.7.7 Likelihoods of observing the DNA profiles, standard deviation noise.0.

5.7.8 Likelihoods of observing the DNA profiles, standard deviation noise.2.

5.7.9 Likelihoods of observing the DNA profiles, standard deviation amplification factor.0.1.

5.7.10 Likelihoods of observing the DNA profiles, standard deviation amplification.0.3.

5.7.11 Likelihoods of observing the DNA profiles, uniform distribution amplification.

5.7.12 Likelihoods of observing the DNA profiles, profile wide expected peak height.

5.7.13 Likelihoods of observing the DNA profiles, original contamination.

5.7.14 Likelihoods of observing the DNA profiles, lognormal distribution contamination.

5.7.15 Likelihoods of observing the DNA profiles, original contamination.99% < 250.

5.8.1 Likelihoods of observing the DNA profiles, use direction to compute \( \tilde{c} \).

5.8.2 Likelihoods of observing the DNA profiles, use compute \( \tilde{c} \) each time.

5.10.1 The alleles on the different loci of the donor of the crime stain of which we obtained four different DNA profiles.

5.10.2 The expected peak heights on the different loci for the four profiles.

5.10.3 The log\(_{10}\) of the likelihoods of observing the DNA profile of the first sample.

5.10.4 The log\(_{10}\) of the likelihoods of observing the DNA profile of the second sample.

5.10.5 The log\(_{10}\) of the likelihoods of observing the DNA profile of the third sample.

5.10.6 The log\(_{10}\) of the likelihoods of observing the DNA profile of the fourth sample.

5.10.7 The sum of the peak heights of the 4 DNA profiles on the locus \( D10 \).

5.10.8 The log\(_{10}\) of the likelihoods of observing the sum of the four DNA profile.

B.1.1 The possible alleles on the different loci.

B.2.1 The profile on 4 loci assigned to our suspect.

B.2.2 The mean, maximum and minimum computed over 50 trees.

B.2.3 The mean, maximum and minimum computed over 20 trees.

B.2.4 The profile on 5 loci assigned to our suspect.

B.2.5 The mean, maximum and minimum computed over 50 trees.

B.3.1 DNA profile on 5 loci for our sensitivity analysis.

B.3.2 Settings.

B.3.3 The log\(_{10}\) of the likelihoods of observing the DNA profile given different allele pairs under our standard settings.

B.3.4 The log\(_{10}\) of the likelihoods of observing the DNA profile given different allele pairs where the expected peak height method was universal.

B.3.5 The log\(_{10}\) of the likelihoods of observing the DNA profile given different allele pairs where the standard deviation of the noise was 5.

B.3.6 The log\(_{10}\) of the likelihoods of observing the DNA profile given different allele pairs where the standard deviation of the noise was 5.

C.1.1 Reference DNA profile for the case example.

C.1.2 DNA profile 1 and 2 for the case example.

C.1.3 DNA profile 3 and 4 for the case example.
Chapter 1

Introduction

“When you have eliminated the impossible, whatever remains, however improbable, must be the truth.”

These words were spoken by Sherlock Holmes, in The Sign of Four by Arthur Conan Doyle [1]. His method to solve a mystery or a crime was merely based on this principle. For Sherlock Holmes, a case usually followed some standard steps. First of all, there was a mystery, a mystery that normal people could not solve. For example, the death of a person, say Sir Charles Baskerville. When Sherlock was confronted with the mystery dr. Watson and he would try to gather as much information on the mystery they could. First by just asking the person that consulted them and later by examining a possible crime scene or by examining objects related to the mystery. The biggest part of their investigation always consisted of thinking; thinking and studying. When they gathered enough information they started solving the mystery and Sherlock always followed the same steps to arrive at a conclusion. He would list all possible explanations for the mystery. Then, one by one, he tried to find a proof that showed that the particular explanation could not be true, since there was evidence that did not coincide with this explanation. This led him to his famous words, “When you have eliminated the impossible, whatever remains, however improbable, must be the truth.”

Forensic research today doesn’t differ a lot from the methods of Sherlock and his friend Watson. When there is a mystery (e.g. ‘does the highway shooter exist?’) occur or a crime (e.g. the Putteuse murder case) has been committed the Netherlands Forensic Institute is consulted to help solving it. Just as the method of Holmes and Watson, the first step consists of gathering evidence that is related to the mystery. The second step is to evaluate this evidence given possible explanations (hypotheses) of the occurred event. This is usually the moment where it becomes clear that there is a big difference between the fictional world, with certainty, Holmes and Watson solved their mysteries and the reality we live in. In most cases, it is impossible to rule out, with certainty, all explanations but one. Therefore, it is not possible to be absolutely sure that a suspect is guilty. Therefore, forensic scientists are asked to determine a likelihood for observing the evidence given a certain explanation. Using these likelihoods it is possible to report a likelihood ratio that shows whether it is more or less likely to observe some evidence (Gigantic footprints that might be left by an even more gigantic hound near the body of Sir Charles Baskerville) under one hypotheses (Sir Charles Baskerville died due to the attack of a gigantic hound) than under another (Sir Charles Baskerville committed suicide). In an ideal case, the likelihood ratio is a number and the forensic scientist experts can report: “It is 100 times more likely to find the gigantic footprints near the body of Sir Charles Baskerville when he died due to the attack of a gigantic hound than when he committed suicide.”

In forensic science, forensic science experts are competent to deal with one piece of evidence. In DNA the statement “It is $M$ times more likely to observe this DNA profile of the blood stain on the jacket of Henry Baskerville when our suspect Jack Stapleton is the donor of the stain than when an unknown person is
the donor.” is one of common practice. However, it becomes more difficult as well as challenging when there are multiple pieces of evidence. A forensic researcher should be capable of giving likelihoods for the distinct pieces of evidence, but in a lot of cases, there is much left to learn about the combination of these pieces. We hope to achieve synergy between different pieces of evidence: showing that the combined evidential value of two pieces of evidence is stronger than when we would regard them separately. The most common problem when combining evidence is that a lot of events are dependent.

In this master thesis we will concentrate on three problems that are based on combining DNA profiles. First we will work on a method to combine Y-chromosomal DNA profiles of different stains. After that we will work on a model that investigates the dependency between autosomal and Y-chromosomal DNA profiles of a single stain. This dependency is important when we would want to combine the evidential value of these DNA profiles. Finally we will develop a model that is able to analyze (low-template) DNA profiles. This model will help us to combine several DNA profiles from the same stain. The first chapter in this thesis is written to help the reader understand the basics of forensic statistics and in the appendix we will give the reader a better understanding of forensic DNA. Those people that are unfamiliar with forensic DNA profiles should read appendix A where we also give a glossary with the most important DNA related terms.

Most importantly, I hope that this thesis helps to close a part of the gap between the reality we live in and the fictional world of Mr. Holmes and Dr. Watson.
Chapter 2

Forensic Statistics

The main issue in forensic science is whether a fact has occurred or not. Since in almost any case, one can never be absolutely sure whether this is the case, the important question becomes how likely it is that the fact has occurred, given the evidence. The main question for forensic statisticians is, therefore, “What is the degree of belief that hypothesis $H_p$ is true given evidence $E$?”. To answer this question Bayes’ theorem is used. In this chapter we give some probability definitions and derive Bayes’ theorem as well as the equation for the Likelihood ratio.

2.1 Bayes’ Theorem

When a crime has been committed and there is a suspect, the evidence in the case is denoted with $E$. This evidence can be an eyewitness, a blood stain on the jacket of the suspect, a message on the phone of the suspect, . . . , etc. An explanation is needed to explain the evidence. Usually two explanations are of interest, one from the prosecutor (for instance the police) $H_p$ and one from the defense (the suspect) $H_d$ (this explanation is often unknown). For instance, when $E$ is a blood stain on the jacket of the suspect.

$H_p$: the suspect is the offender.
$H_d$: the suspect is not the offender.

Judges are interested in the answer to the question “Which hypothesis ($H_p$ or $H_d$) is more likely, given the evidence?” To answer this we derive Bayes’ Theorem.

Definition 2.1.1. Conditional Probability

Let $P(H_p)$ and $P(E)$ be the probabilities that propositions (or events) $H_p$ and $E$ are true. Then $P(H_p|E)$ is the probability that $H_p$ is true, given that the event $E$ is true. If $P(H_p) > 0$ and $P(E) > 0$, the following holds,

$$P(H_p|E) = \frac{P(H_p \cap E)}{P(E)}, \quad (2.1.1)$$

$$P(E|H_p) = \frac{P(H_p \cap E)}{P(H_p)}. \quad (2.1.2)$$

Now we can combine (2.1.1) and (2.1.2) and find

$$P(H_p|E)P(E) = P(E|H_p)P(H_p).$$

And with this result Bayes’ theorem is easily derived:

Theorem 2.1.1. Bayes’ theorem

Let $P(H_p)$ and $P(E)$ be the probabilities that propositions $H_p$ and $E$ are true. Then $P(H_p|E)$ is the probability that $H_p$ is true, given that proposition $E$ is true. If $P(H_p) > 0$ and $P(E) > 0$, then Bayes’ theorem states that

$$P(H_p|E) = \frac{P(E|H_p)P(H_p)}{P(E)}. \quad (2.1.3)$$
We have a formula for the conditional probability of \( H_p \) given \( E \), the posterior probability. The probabilities \( \mathbb{P}(H_p) \) and \( \mathbb{P}(E) \) are the prior probabilities, the probabilities that \( H_p \) and \( E \) will occur without considering any other information. \( \mathbb{P}(E|H_p) \) is the likelihood. Usually, a judge wants to know the ratio of posterior probabilities,

\[
\frac{\mathbb{P}(H_p|E)}{\mathbb{P}(H_d|E)}
\]

We use equation 2.1.3 to find,

\[
\frac{\mathbb{P}(H_p|E)}{\mathbb{P}(H_d|E)} = \frac{\mathbb{P}(E|H_p)\mathbb{P}(H_p)}{\mathbb{P}(E|H_d)\mathbb{P}(H_d)} = \frac{\mathbb{P}(E|H_p)}{\mathbb{P}(E|H_d)} \cdot \frac{\mathbb{P}(H_p)}{\mathbb{P}(H_d)}.
\]

(2.1.4)

The term \( \frac{\mathbb{P}(H_p)}{\mathbb{P}(H_d)} \), the ratio of prior probabilities, needs to be determined by the judge. The term \( \frac{\mathbb{P}(E|H_p)}{\mathbb{P}(E|H_d)} \), the likelihood ratio\(^{1}\) is answered by forensic scientists.

### 2.1.1 The Likelihood Ratio

The likelihood ratio (LR) is a measure of the value of evidence regarding two hypotheses. It indicates the extent to which the evidence is in favor of one hypothesis over the other. So,

\[
LR = \frac{\mathbb{P}(E|H_p)}{\mathbb{P}(E|H_d)}.
\]

(2.1.5)

If \( H_p \) stands for “the suspect is the offender” and \( H_d \) for “the suspect is not the offender” we could summarize the results as follows.

- If the likelihood ratio \( M \) is bigger than 1, the evidence is in favor of \( H_p \), we report it is \( M \) times more likely to find the evidence \( E \) when \( H_p \) is true, than when \( H_d \) is true.

- If the likelihood ratio \( M \) is smaller than 1, the evidence is in favor of \( H_d \), we report it is \( M \) times more likely to find the evidence \( E \) when \( H_d \) is true, than when \( H_p \) is true.

- If the likelihood ratio is 1, the evidence favors neither \( H_p \) nor \( H_d \). We report it is equally likely.

\(^{1}\)A better word may be Bayes Factor. This is currently being discussed. However in this text we will use the commonly used words likelihood ratio.
Chapter 3

The combination of partial Y-chromosomal DNA profiles

Introduction
In this chapter we will discuss an approach on how to combine partial Y-chromosomal DNA profiles found on the same crime scene. The toy-problem which is the basis of our approach is the following:

**Problem**  On a t-shirt of a victim of a sex offense two DNA-traces from separate stains are secured. The traces provide two partial Y-profiles, which partly overlap and the overlapping loci match. Both profiles match the DNA-profile of the suspect Bert. What is the value of the evidence of these observations?

In this chapter we will discuss a general method that can be used to solve this problem. First we will give the hypotheses we use to evaluate the evidence we found, the notations we will use and the assumptions we made. After that we will give an approach on how to compare more than two propositions/hypotheses and what we need to compute the likelihood ratio. Lastly, we will derive a lower bound for the likelihood ratio when comparing multiple hypotheses that can be used without determining the prior probabilities.

3.1 A method to determine a likelihood ratio when multiple partial Y-chromosomal DNA profiles are found

3.1.1 Hypotheses
In the toy-problem we recognize several possible explanations (hypotheses) for the findings, see table 3.1.1.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_a$</td>
<td>the suspect is the source of both crime stains.</td>
</tr>
<tr>
<td>$H_b$</td>
<td>the suspect is the source of the first crime stain, some unknown unrelated person is the source of the second crime stain.</td>
</tr>
<tr>
<td>$H_c$</td>
<td>the suspect is the source of the second crime stain, some unknown unrelated person is the source of the first crime stain.</td>
</tr>
<tr>
<td>$H_{d1}$</td>
<td>the suspect is not the source of the crime stains, one unknown unrelated person is the source of the crime stains.</td>
</tr>
<tr>
<td>$H_{d2}$</td>
<td>the suspect is not the source of the crime stains, two unknown unrelated (to the suspect and to each other) persons are the sources of the crime stains.</td>
</tr>
</tbody>
</table>

Table 3.1.1: Hypotheses that explain the findings

The evidence we have will be denoted in the following way, see table 3.1.2.
$E_{c1}$: the correspondence between the DNA of the suspect and the DNA obtained from the first crime stain (seeing the same alleles on the DNA profile from crime stain 1 as on the profile obtained from the suspect).

$E_{c2}$: the correspondence between the DNA of the suspect and the DNA obtained from the second crime stain.

$E_s$: the sample from the suspect.

$E$: the correspondence between the suspect sample ($E_s$) and the crime samples 1 and 2.

Table 3.1.2: The notation of the evidence

We will use $E$ in our derivations.

**Assumptions**

To make statements on the likelihood of the hypotheses we will make some assumptions on the evidence. The assumptions are as follows.

- One Y-chromosomal DNA profile is not entirely overlapped by the other.
- The DNA stains each contain Y-chromosomal DNA of a single person (not necessarily the same person), i.e. they are not a DNA-mixture.

**Combining Hypotheses**

Since we are not in a classical case where we only have two hypotheses (he did it versus someone else did it) we need to combine hypotheses, in order to compare one with a group of the others. There are two possible combinations of subsets that are interesting in this case.

1. The suspect left both crime stains versus the suspect did not leave both crime stains. We combine $H_b, H_c, H_{d1}$ and $H_{d2}$ and compare them with $H_a$.

2. The suspect left at least one of the crime stains versus the suspect did not leave any crime stain. In this case we combine $H_a, H_b$ and $H_c$ and compare them with $H_{d1}$ and $H_{d2}$.

Taroni and Aitken give an approach for comparing more than two propositions \[16\]. Consider a number $n$ of competing exclusive propositions, $H_1, \ldots, H_n$. Let $E$ denote the evidence to be evaluated under each of the $n$ propositions, and consider $P(E|H_i)$, with $i = 1, \ldots, n$. If the prior probabilities of each of the propositions $H_i$ are available, the ratio of the probability of $E$ given each of the pair of competing propositions $H_1$ and $H_2$ can be evaluated as follows:

$$
\frac{P(E|H_1)}{P(E|\overline{H}_1)} = \frac{P(E|H_1) \{ \sum_{i=2}^{n} P(H_i) \}}{\sum_{i=2}^{n} P(E|H_i)P(H_i)} \tag{3.1.1}
$$

**The suspect left both crime stains**  When we want to compare the two collections of propositions

$H_1$: the suspect left both crime stains, $H_a$.

$\overline{H}_1$: the suspect did not leave both crime stains, $H_b \cup H_c \cup H_{d1} \cup H_{d2}$.

We can use the equation of Aitken and Taroni \[3.1.1\], from which it follows that

$$
\frac{P(E|H_1)}{P(E|\overline{H}_1)} = \frac{P(E|H_1) \{ \sum_{i=2}^{n} P(H_i) \}}{\sum_{i=2}^{n} P(E|H_i)P(H_i)} = \frac{P(E|H_a) \{ P(H_b) + P(H_c) + P(H_{d1}) + P(H_{d2}) \}}{P(E|H_b) \cdot P(H_b) + P(E|H_c) \cdot P(H_c) + P(E|H_{d1}) \cdot P(H_{d1}) + P(E|H_{d2}) \cdot P(H_{d2})} \tag{3.1.2}
$$
The suspect left at least one of the crime stains  Here we want to compare two separate subsets of hypotheses.

\[ H_p: \text{The suspect left at least one of the crime stains,} \quad H_p = H_a \cup H_b \cup H_c. \]

\[ H_d: \text{the suspect did not leave a crime stain,} \quad H_d = H_{d1} \cup H_{d2}. \]

Since we are comparing two subsets of hypotheses, we can’t use the equation given by Aitken and Taroni and have to derive one ourself.

\[
\mathbb{P}(E|H_d) = \frac{\mathbb{P}(E \cap H_d)}{\mathbb{P}(H_d)} = \frac{\mathbb{P}(E|H_{d1}) \cdot \mathbb{P}(H_{d1}) + \mathbb{P}(E|H_{d2}) \cdot \mathbb{P}(H_{d2})}{\mathbb{P}(H_{d1}) + \mathbb{P}(H_{d2})} \tag{3.1.3}
\]

and,

\[
\mathbb{P}(E|H_p) = \frac{\mathbb{P}(E \cap H_p)}{\mathbb{P}(H_p)} = \frac{\mathbb{P}(E|H_a) \cdot \mathbb{P}(H_a) + \mathbb{P}(E|H_b) \cdot \mathbb{P}(H_b) + \mathbb{P}(E|H_c) \cdot \mathbb{P}(H_c)}{\mathbb{P}(H_a) + \mathbb{P}(H_b) + \mathbb{P}(H_c)} \tag{3.1.4}
\]

so, by combining (3.1.3) and (3.1.4),

\[
\mathbb{P}(E|H_a) = \frac{\mathbb{P}(E|H_a) \cdot \mathbb{P}(H_a) + \mathbb{P}(E|H_b) \cdot \mathbb{P}(H_b) + \mathbb{P}(E|H_c) \cdot \mathbb{P}(H_c)}{\mathbb{P}(H_a) + \mathbb{P}(H_b) + \mathbb{P}(H_c)} \tag{3.1.5}
\]

General formula for the evaluation of evidence with more than two propositions

We can also derive a general formula which can be used to evaluate the evidence when there are more than two propositions. Suppose we want to compare the following pair of propositions,

\[ H_p := \cup_{i \in I} H_i, \]
\[ H_d := \cup_{j \in J} H_j, \]

where \( I, J \in [1, 2, \ldots, n] \) and \( I \cap J = \emptyset \). When we know the prior probabilities of each of the propositions \( H_i \) we can evaluate the ratio of the probability of \( E \), the evidence, given each of the pair of competing propositions, as follows

\[
\mathbb{P}(E|H_p) = \frac{\sum_{h_i \in H_p} \mathbb{P}(E|H_i) \cdot \mathbb{P}(H_i)}{\sum_{h_j \in H_d} \mathbb{P}(E|H_j) \cdot \mathbb{P}(H_j)} \tag{3.1.6}
\]

3.1.2  \( \mathbb{P}(E|H_i) \)

We return to our toy-problem and see that when we want to compare different subsets of propositions, we need two probabilities: \( \mathbb{P}(E|H_i) \) and \( \mathbb{P}(H_i) \). In this section we will discuss the first one. We will derive a method that won’t need the second in 3.2.2.

We need to know the probabilities of finding the evidence given the different hypotheses. We will do this in the following paragraphs.

\( \mathbb{P}(E|H_a) \)

The probability of finding the evidence given that our suspect Bert is the source of both crime stains is easily done. Since the DNA-profiles of the stains match the DNA profile of Bert we get (assuming no errors were made)

\[
\mathbb{P}(E|H_a) = 1. \tag{3.1.7}
\]
Now we wonder what the probability is of finding the evidence given that our suspect Bert is the source of the first crime stain and that some unknown person is the source of the second crime stain. These two events are independent given $H_b$. If $H_b$ is true, then the probability of $E_{c1}$, the correspondence between the suspect and the first crime sample, remains unaffected (assuming that the unknown person is unrelated to the suspect) if in addition knowledge about $E_{c2}$, the correspondence between the suspect and the second crime sample, becomes available. Analogously, $E_{c2}$ remains unaffected by knowledge about $E_{c1}$, if $H_b$ is known. Therefore we can multiply the different probabilities to find

$$
P(E|H_b) = P(E_{c1}|H_b) \cdot P(E_{c2}|E_{c1}, H_b) = 1 \cdot \gamma_2 = \gamma_2,
$$

where $\gamma_2$ is the random match probability of the DNA profile obtained from the second crime stain.

**P(E|H_e)**

If we argue analogous to the previous part we find

$$
P(E|H_e) = \gamma_1,
$$

where $\gamma_1$ is the random match probability of the DNA profile obtained from the first crime stain.

**P(E|H_{d1}) and P(E|H_{d2})**

The probability of finding the evidence given that none of the stains were left by our suspect Bert is interesting. In that case there are two possible scenarios. Either one unknown person is the source of both stains or two unknown persons are the sources of the separate stains.

1. When there is one unknown person that left both stains we can combine the two partial Y-profiles into one bigger (partial) Y-profile and the probability of the evidence could be estimated by using the profile probability

$$
P(E|H_{d1}) = \gamma,
$$

where $\gamma$ is the random match probability of the DNA profile obtained from combining the profiles from the first and second crime stain.

2. When there are two unknown persons that left the separate stains we can estimate the probability of the evidence for the separated partial Y-profiles by using the profile probabilities, $P(E_{c1}|H_{d2}) = \gamma_1$ and $P(E_{c2}|H_{d2}) = \gamma_2$. The probability that two unknown persons (unrelated to each other) left the separate stains is given by the product of these two probabilities times two (there are two combinations)

$$
P(E|H_{d2}) = 2\gamma_1\gamma_2
$$

We can combine the results in table 3.1.3.

| $P(E|H_a)$ | $P(E|H_b)$ | $P(E|H_c)$ | $P(E|H_{d1})$ | $P(E|H_{d2})$ |
|------------|------------|------------|---------------|---------------|
| 1          | $\gamma_2$ | $\gamma_1$ | $\gamma$      | $2\gamma_1\gamma_2$ |

Table 3.1.3: Values for $P(E|H_i)$

### 3.1.3 Conclusion

A problem that arises when considering a case with multiple pieces of evidence is that there is a possibility that there were multiple offenders. Therefore, it is (almost always) impossible to make a pair of hypotheses that are both mutually exclusive and collectively exhaustive (this is not necessary). A way to overcome this problem is to define more than two hypotheses. To evaluate the evidence in these cases we should combine some of these hypotheses in a way that we have a pair of subsets of hypotheses. Interesting pairs
are ones that distinguish between the suspect is the source of at least one of the pieces of evidence or the suspect is the source of all the pieces of evidence against their complementary hypotheses (the suspect did not leave any of the stains and the suspect is not the source of all the pieces of evidence). This can be done by using the general formula for the evaluation of evidence with more than two propositions, given in section 3.1.1.

To use this formula we need to know the prior probabilities of all the hypotheses as well as the posterior ones. In the case that we are considering DNA-evidence, the posterior probabilities for all the hypotheses can be assigned by using the profile probability of the DNA-profiles obtained from the different pieces of evidence. The prior probabilities need to be assigned by a judge. We will derive a method that won’t use these prior probabilities in section 3.2.2.

This report was written on the basis of a toy problem. This problem is based on two crime stains that both provided a partial Y-chromosomal DNA profile. The same approach can be used when considering other pieces of evidence. When the evidence gathered in a case consists of a footprint and a fingerprint we are still sure that either one or two people left these. In general the following rule holds, the number of pieces of evidence is smaller or equal to the number of people that left them (where we assume that there is no evidence that can be left two persons like DNA-mixture profiles).

When writing a report on a crime and making a likelihood ratio on the evidence of the DNA match of a suspect profile with multiple profiles obtained from various crime stains it is important to note that the likelihood ratio score is based on the assumption that the unknown other people with which we compare are unrelated to the suspect. When doing this one can come up with the same kind of conclusions as in a case where there was only one piece of evidence: The probability of finding the evidence $E$ (which consists of two parts) given that the suspect left at least one of the crime stains is $M_1$ times more likely than when he did not or The probability of finding the evidence $E$ given that the suspect left both of the crime stains is $M_2$ more likely than when he did not.

3.2 Methods based on the previous method

3.2.1 Selecting the priors

A problem that arises with the first method is that the number of hypotheses grows rapidly when the number of crime stains increases. When we only have one crime stain we get two hypotheses, one that states that the suspect left the crime stain and one that states that an unknown person left the crime stain. As we saw, when there are two crime stains we will use five hypotheses. Three stains will lead to the list of possible hypotheses given in table 3.2.1.
Table 3.2.1: All possible hypotheses for a case with three crime stains and one suspect

It becomes very difficult to assign prior probabilities to all these hypotheses, and slightly different prior probabilities will lead to (slightly) different likelihood ratio’s. Another approach might be desirable here. It is unlikely that a suspect would testify that either of the stains originated from him. In most cases the relevant comparison is between the hypotheses that either states that the suspect left all the stains versus that the suspect did not leave any of the stains.

**Only considering certain hypotheses**

Consider a situation where the evidence consists of three crime stains of which the DNA-profiles are matching with the DNA profile of a suspect. A method to deal with this situation is to consider only the hypotheses which imply that the suspect left all the crime stains or the ones that imply that the suspect did not leave any of the stains, see table 3.2.2.

<table>
<thead>
<tr>
<th>$H_p$</th>
<th>$H_1$: the suspect is the source of all three crime stains.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_2$:</td>
<td>the suspect is not the source of the crime stains, some unknown person is the source of all three crime stains.</td>
</tr>
<tr>
<td>$H_3$:</td>
<td>the suspect is not the source of the crime stains, some unknown person is the source of the first crime stain, another person is the source of the last two crime stains.</td>
</tr>
<tr>
<td>$H_4$:</td>
<td>the suspect is not the source of the crime stains, some unknown person is the source of the second crime stain, another person is the source of the first and last crime stains.</td>
</tr>
<tr>
<td>$H_5$:</td>
<td>the suspect is not the source of the crime stains, some unknown person is the source of the third crime stain, another person is the source of the first two crime stains.</td>
</tr>
<tr>
<td>$H_6$:</td>
<td>the suspect is not the source of the crime stains, three unknown persons left the three separate crime stains.</td>
</tr>
</tbody>
</table>

Table 3.2.2: The hypotheses when we only consider situations where the suspect left either all or none of the stains

In this case we can use our general formula for the evaluation of evidence with more than two propositions (3.1.6). We get as a results

$$
\frac{P(E|H_p)}{P(E|H_d)} = \frac{P(E|H_1) \cdot (P(H_2) + P(H_3) + P(H_4) + P(H_5) + P(H_6))}{P(E|H_2)P(H_2) + P(E|H_3)P(H_3) + P(E|H_4)P(H_4) + P(E|H_5)P(H_5) + P(E|H_6)P(H_6)}
$$

To evaluate this we need to know the conditional probabilities of finding the evidence given the different hypotheses, which can be computed using the profile frequencies of the DNA profiles obtained from the
crime stains. The prior probabilities need to be assigned, just like before, but now we only have to assign probability to 6 propositions instead of 15.

We recognize the same downside to this method as in the method where we used all hypotheses. The prior probabilities to all the hypotheses need to be selected.

3.2.2 Finding a lower bound for the likelihood ratio

Here we discuss another method to combine the evidence. One where we try to find a lower bound for the likelihood ratio and where we don’t need the prior probabilities of all the hypotheses. Consider the case where we had two crime stains (the original problem). We want to combine the evidence and work on the following couple of hypotheses.

\[ H_p : \text{the suspect left both crime stains.} \]
\[ H_d : \text{the suspect did not leave both crime stains.} \]

In section 3.1 (see equation 3.1.2) we derived a form for the likelihood ratio.

\[
\frac{\mathbb{P}(E|H_p)}{\mathbb{P}(E|H_d)} = \frac{\mathbb{P}(E|H_b) \cdot (\mathbb{P}(H_b) + \mathbb{P}(H_c) + \mathbb{P}(H_d) + \mathbb{P}(H_{d2}))}{\mathbb{P}(E|H_b) \cdot \mathbb{P}(H_b) + \mathbb{P}(E|H_c) \cdot \mathbb{P}(H_c) + \mathbb{P}(E|H_{d1}) \cdot \mathbb{P}(H_{d1}) + \mathbb{P}(E|H_{d2}) \cdot \mathbb{P}(H_{d2})}
\]

We can do the following derivation:

\[
\frac{\mathbb{P}(E|H_p)}{\mathbb{P}(E|H_d)} = \frac{\mathbb{P}(E|H_b) \cdot (\mathbb{P}(H_b) + \mathbb{P}(H_c) + \mathbb{P}(H_d) + \mathbb{P}(H_{d2}))}{\mathbb{P}(E|H_b) \cdot \mathbb{P}(H_b) + \mathbb{P}(E|H_c) \cdot \mathbb{P}(H_c) + \mathbb{P}(E|H_{d1}) \cdot \mathbb{P}(H_{d1}) + \mathbb{P}(E|H_{d2}) \cdot \mathbb{P}(H_{d2})}
\]

\[
\geq \frac{\mathbb{P}(E|H_b)}{\max_{i \in \{b,c,d1,d2\}} \mathbb{P}(E|H_i)}
\]

(3.2.1)

Our result is a lower bound for the likelihood ratio where we do not need the prior probabilities of the hypotheses. However, there is something unsatisfying when using this method. The likelihood ratio is of the same order as the likelihood ratio would be when we would compare the hypotheses that states that the suspect left both crime stains and the hypotheses with the largest conditional probability. This comes from the fact that the alternative hypotheses in our example consists of hypotheses that state that the suspect left one of the crime stains and that an unknown unrelated person left the other. Therefore another approach might be desirable here. In practice it is more desirable to compare the following pair of hypotheses.

\[ H_p : \text{the suspect left all crime stains.} \]
\[ H_d : \text{the suspect did not leave any of the crime stains.} \]

We will assume that there are \(k\) distinct pieces of evidence. The alternative hypothesis, \(H_d\) consists of \(n - m\) propositions that state that the stains were not left by the suspect and that either 1, 2, \ldots, \(k\) other unknown unrelated men left them. Equation [3.1.2] becomes

\[
\frac{\mathbb{P}(E|H_p)}{\mathbb{P}(E|H_d)} = \frac{\mathbb{P}(E|H_p) \cdot (\mathbb{P}(H_m) + \mathbb{P}(H_{m+1}) + \ldots + \mathbb{P}(H_n))}{\mathbb{P}(E|H_m) \cdot \mathbb{P}(H_m) + \mathbb{P}(E|H_{m+1}) \cdot \mathbb{P}(H_{m+1}) + \ldots + \mathbb{P}(E|H_n) \cdot \mathbb{P}(H_n)}
\]

Which we can derive to a lower bound

\[
\frac{\mathbb{P}(E|H_p)}{\mathbb{P}(E|H_d)} = \frac{\mathbb{P}(E|H_p) \cdot (\mathbb{P}(H_m) + \mathbb{P}(H_{m+1}) + \ldots + \mathbb{P}(H_n))}{\mathbb{P}(E|H_m) \cdot \mathbb{P}(H_m) + \mathbb{P}(E|H_{m+1}) \cdot \mathbb{P}(H_{m+1}) + \ldots + \mathbb{P}(E|H_n) \cdot \mathbb{P}(H_n)}
\]

\[
\geq \frac{\mathbb{P}(E|H_p)}{\max_{i \in \{m,\ldots,n\}} \mathbb{P}(E|H_i)}
\]

(3.2.2)
When the evidence consists of Y-chromosomal DNA profiles, the alternative hypothesis with the highest likelihood will usually be the one where one unrelated unknown person left all the stains.

### 3.2.3 Conclusion

A downside to the method described in section 3.1 is that the number of propositions that explain the findings grows rapidly when the number of pieces of evidence increases. Another downside is that to evaluate the weight of the evidence given the different propositions we need to know the prior probabilities of all the propositions. In this section we have shown different ways to overcome the big number of possible propositions.

One way that we described is by only considering propositions that state that the suspect either did leave all the crime stains or where he did not leave any of the crime stains. This greatly reduces the total number of propositions. This method doesn’t take away the problem of multiple prior probabilities that need to be estimated.

The last method that we suggested was deriving a lower bound for the likelihood ratio. We derived a lower bound for the original, and true, likelihood ratio which doesn’t contain prior probabilities. This method can be used immediately and there is no need to estimate certain parameters. We suggest that the best way to use it is by only comparing the hypotheses the suspect is the source of all the evidence versus the suspect is not the source of any piece of evidence. Where the latter consists of all the possible propositions (i.e. one unknown unrelated person is the source of all the evidence, one unknown unrelated person is the source of the first piece of evidence another unknown unrelated person is the source of the other pieces, ...). In a forensic report one could write,

> “The probability of finding the evidence E given that the suspect left all the crime stains is at least M times more likely than when he did not leave any crime stain”.

Where $M$ is the lower bound we derived. We expect that the method will only need small adaptations to make it applicable to other sorts of evidence, such as fingerprints.
Chapter 4

The combination of autosomal and Y-chromosomal DNA profiles

4.1 Introduction

Suppose that a crime stain is found at a crime scene. The stain is examined by specialists and two DNA profiles are obtained from it: a Y-chromosomal DNA profile and an autosomal DNA profile. Since the two profiles are obtained from the same stain we can be almost sure that they were left by the same person. Unfortunately, the profiles are not discriminating enough separately. Suppose that the Y-chromosomal DNA profile has a profile frequency of 1 in 10,000, meaning that approximately one in ten thousand people has the same profile, and the (partial) autosomal DNA profile has a profile frequency of 1 in 1000. Combining these two frequencies would lead to a combined profile frequency of 1 over 10,000,000 (ten million), which is extremely discriminating. Combining these two frequencies can only be done when the two profiles are independent, the Y-chromosomal DNA profile has to be independent of the autosomal DNA profile. Since the Y-chromosomal of a man is directly inherited from his father, and the autosomal DNA profile is a combination of the autosomal DNA profiles of his parents, assuming this independence seems to be wrong. When we assume that two men are completely unrelated, then we are certain that the Y-chromosomal DNA profile is not giving any information about the autosomal DNA profile. And this is exactly where our problem arises: if two people share the same Y-chromosomal DNA it is more likely that they are somehow related to each other, and relatedness implies a dependency on their autosomal DNA profiles. We want to examine to which extent this dependency should be taken into account, i.e. how much the ‘true’ match probability deviates from multiplying. At first, it seems clear that there is a direct dependency between these two profiles for a father and his son, but how about a grandfather and his grandson, or dependency between two second cousins? In this chapter we will answer these questions.

4.2 DNA profiles

In this section we will provide some basic knowledge about DNA profiles that is required to understand some of the assumptions we made in the model. For a more complete explanation of DNA and DNA profiles we suggest to read the appendix on DNA, (appendix A).

4.2.1 Introduction

A DNA profile is a representation of some parts of somebody’s DNA. There are different sorts of DNA profiles, most notably autosomal and Y-chromosomal DNA profiles. Within cells DNA is organized into long structures called chromosomes. Each person has 46 chromosomes that occur in pairs, 22 pairs of autosomal and one pair of sex chromosomes.

DNA consists of two very long strands. A strand consists of a sequence of nucleobases. Only 2% of the DNA provides for the genetic characteristics. The other 98% are ‘non-coding’ repetitive sequences and can be used to make a DNA profile. On different locations of the DNA, DNA specialists count the number
of times a small part of nucleobases is repeated. For example the part ‘ATCGATCGATCGATCGATCGATCG’ consists of the sequence of nucleobases ATCG that is repeated six times. These locations are called the loci (sing. locus). The number of times a sequence is repeated is called the allele. Since DNA consists of two paired strands, on each locus their are two alleles. The alleles on 15 different loci combined is the autosomal DNA profile.

4.2.2 DNA profiles

The true power of the analysis of a DNA profile is in its statistical power of discrimination. Because the 15 loci that are currently used for discrimination are assumed to be approximately independently assorted (having a certain number of repeats on one locus doesn’t change the likelihood of having any number of repeats at any other locus), we can apply the product rule of probabilities (with a small correction for dependency due to population substructures, see section 4.3.3). At each locus we can find different alleles. 2085 men in the Netherlands approved use of their DNA profiles to estimate the frequency of each allele in the Dutch population. For example, on the locus with code ‘D2S1388’ the allele combination 19/25 has an estimated frequency of 3.7%, i.e. of the Dutch population approximately 3.7% have the same alleles at that locus. This can lead to very strong evidence against a suspect. Suppose an entire DNA-profile (alleles at 15 loci) is obtained from a crime scene. The police also obtain a sample from the suspect and find that its profile matches with the profile found on the crime scene. Taking the product of the different frequencies can lead to a profile frequency estimate way less than one in a billion (assuming a population that is unrelated to the suspect and with small correction for population substructure and sample size)[14].

Autosomal profiles

Autosomal, or standard, DNA-research is the analysis of the DNA of at least ten different loci. The gender of the corresponding person is also determined. The combination is an autosomal DNA-profile. It is possible that a DNA analysis of a stain does not yield a ‘complete’ DNA profile. For instance, if a body is found somewhere floating at sea and a blood stain is found on the shirt of the victim which DNA profile doesn’t match with the victim, it is possible that the DNA is degraded due to the fact that the body has been laying in the water for quite some time. The profile that is obtained from the stain is called a partial DNA profile, only the alleles on some loci can be determined. This means that the profile frequency of the partial profile is larger than it would have been when a full profile was obtained. In these cases it becomes interesting to combine the profile frequency of the autosomal DNA profile with the profile frequency of a Y-chromosomal DNA profile. A full autosomal DNA profile is almost always discriminating enough by itself.

Y-chromosomal profiles

Of the 23 chromosomal pairs, one decides the sex of the person. For women this is represented with two X’s, XX, for men it is represented with one X and one Y, XY. Each child inherits one chromosome from both parents. If the father passes the Y-chromosome the child will be a boy, if the father passes the X-chromosome the child will be a girl. During the passing of the Y-chromosome it is not or barely affected, which means that it stays the same over many generations. Mutations of the Y-chromosome are possible but rare.

In some cases DNA of different people is mixed within one stain or trace (think of a sex-offense where the DNA of the victim is ‘mixed’ with the DNA of the culprit). In these cases the autosomal research method can hardly distinguish the alleles of the victim and the ones of the culprit. Y-chromosomal DNA-research makes it possible to determine the specific DNA-characteristics of the male DNA in a mixed DNA trace. Since the Y-chromosome barely changes from father to son it is important to remember that when a suspect matches the Y-chromosomal DNA profile of a stain left at the crime you can be almost sure that his father, brothers, sons, etc. will also match the same DNA profile.
Heritage

On each locus of your autosomal DNA profile, one of the alleles is inherited from your father, and one is inherited from your mother. For two unrelated persons, it is very unlikely that they share the same DNA profile. For two siblings, it is still not very likely that they have the exact same DNA profile, but since their DNA profile is a random combination of the DNA profiles of their parents, it is much more likely than for two unrelated persons. Identical twin brothers or sisters share the exact same DNA profile (except for mutations).

Since all men have a Y-chromosome, and this Y-chromosome is directly inherited from the person’s father; all fathers have the exact same Y-chromosomal DNA profile as their sons (except for mutations). Therefore, an autosomal DNA profile is much more discriminating than a Y-chromosomal DNA profile. A lot of people share the same Y-chromosomal DNA profile whereas an autosomal DNA profile is usually unique.

4.3 Model

In this section we will explain why a simulation is a good tool for examining the dependency of the Y and autosomal profiles. After that we will explain how the model is constructed and which assumptions are made. We especially focus on the assumption of the existence of sub-populations within a population, and how we added this in our model.

4.3.1 Introduction

We are interested in the probability that the culprit of some crime has a combination of a Y-chromosomal and an autosomal DNA profile, given that the suspect has the same combination of DNA profiles and that he did not commit the crime. If we set \( H_d \) to be “The suspect is not the culprit” and let \( G_c \) denote the DNA profiles (Y-chromosomal and autosomal) of the culprit, where \( G_s \) represents the same for the suspect, we could write

\[
P(G_c = (g_y, g_a) | G_s = (g_y, g_a), H_d) = P(G_{cy} = g_y | G_s = (g_y, g_a), H_d) \cdot P(G_{ca} = g_a | G_s = (g_y, g_a), G_{cy} = g_y, H_d)
\]

The probability that we will examine in our model is,

\[
P(G_{ca} = g_a | G_s = (g_y, g_a), G_{cy} = g_y, H_d)
\]

In other words, we will examine the probability that the culprit has the same autosomal DNA profile, given that they share the same Y-chromosomal DNA profile and that the suspect is not the culprit. In our model we will work with a ‘best case’ scenario for the suspect, i.e. all the model assumptions that are made will be either in favor of the suspect or not in favor of anybody.

In order to examine the dependency of Y-chromosomal and autosomal DNA profiles we make a model that represents a population. For each Y-chromosomal DNA profile we can estimate the profile frequency. This profile frequency gives us an idea of the total number of people that bear it in a certain population (for example the Dutch population). Our model assumes that all these people are related in some way, which basically means that we can find a common ancestor for everybody sharing the same Y-chromosomal DNA profile. If the estimated number of people that bear the same Y-chromosomal DNA profile is big, this might mean that some people in our simulation are only considered to be related because they share the same great-grandfather. The model can be divided into two parts. First we make a family tree that links all the people with the same Y-chromosomal DNA profile, secondly we will simulate the process of assigning autosomal DNA profiles to all the people in the tree. We will do this by taking into account the population substructures and heritage of autosomal DNA profiles between people.

Why simulation

Simulating this whole process in order to investigate the dependency between the autosomal and the Y-chromosomal DNA profiles gives us some advantages with regard to an analytical approach. First of
all, we can calculate analytically the probability for pairs of people, e.g., that a cousin of a man shares the same autosomal DNA profile. We can also calculate analytically the probability of a match among a group of relatives [15]. However, it is very hard to derive analytically the match probability for various family tree structures, number of living people sharing some autosomal DNA profile, etc., when doing a simulation this can be examined very fast and easy. Therefore we use simulation as a tool to calculate this probability.

4.3.2 Assumptions
In order to make the model we have to make some assumptions. These are listed here:

1. For any Y-chromosomal DNA profile we know the number of people that bear it in a certain population.
2. All people that share the same Y-chromosomal DNA profile are somehow related to each other, it is possible to find a common ancestor that links all of them.
3. When a Y-chromosome is inherited from father to son there are no copying mistakes (mutations). All father-son pairs share the exact same Y-chromosomal DNA profile.
4. The population of which we know the allele-frequencies consists of several sub-populations. In these sub-populations people mate with each other. This means that the allele-frequencies known for the whole population are slightly different for the several sub-populations. We assume that the family of our suspect is in the same sub-population as our suspect.
5. For a man it is impossible to become older than 120. If a person reaches the age of 120 he will die in the year that follows.
6. The probability of becoming father of a son for a person in the tree is 0.0625 each year when he is between the 20 and 40 years old. This number is chosen on the basis of trial and error. We started with probability 0.5 (which is an overestimation of the actual number, see section 4.4.2), checked the results, divided it by 2 when the number of sons seemed to be an overestimation and repeated this procedure until the results were acceptable in the sense that each man got a number of sons that came close to the current actual number of sons a man gets.
7. The last person that is added to the family tree represents our suspect.

We have to be careful that our assumptions are not disadvantageous for a suspect. One thing can be said directly about what would be advantageous for the suspect. First of all, when the number of generations that is needed to link all the people in the family tree becomes smaller, then the number of times that the profile of the suspect gets modified becomes smaller as well (every time we simulate the profile of somebody’s father or son the original profile gets modified once more). Simply stated, for the suspect it would be best when all of the Y-chromosomal bearers would be brothers of the suspect since they are the people that have the highest probability of sharing the same autosomal DNA profile. We will now discuss the consequences for the suspect for each assumption above in turn.

1. The first one is not disadvantageous to the suspect. The profile frequency of a certain Y-chromosomal DNA profile can be estimated based on a DNA-database. This profile frequency tells us how frequent the profile is and can be turned into an approximate number that bear it, given a total number of people of a population.
2. This assumption is to investigate to what extent a related person has a higher probability to share the same autosomal DNA profile than an unrelated one. We assume that everybody with the same autosomal DNA profile is related to each other and compare the results with an unrelated population.
3. This is clearly in favor of the suspect. It basically says that the number of generations that are needed to link everybody with the same Y-chromosomal DNA profile is lower than when mutations are possible.
4. This assumption is further explained in section 4.3.3.

5. This assumption might be explained as being disadvantageous for the suspect, since the older people become the better for the suspect. On the other hand, in all history the oldest acknowledged male ever lived became only 115 years. Also, it is very unlikely to find a man of 115 years old to be the culprit.

6. As shown in section 4.4.2, this number leads to a higher mean number of sons than the current mean number of sons. This implies a smaller number of generations that is needed to complete the family tree, which is advantageous to a suspect.

7. This assumption might be causing trouble. A consequence of this assumption is that the number of brothers of the suspect is lower than it might be, since it is possible that he gets another brother (or even more) in the following years. Another problem arises since the suspect doesn’t have any children (or even grandchildren), and that the same holds for his brothers. All these problems will disappear when we neglect some close family when comparing our results. This means that even if he had more brothers, children or grandchildren, these people would not be taken into account anyway. Since we link a fixed number of people in the tree, and we neglect the close family, there are more people we will take into account when the suspect doesn’t have any brothers or children. We take the total number of people minus the people that are in the close family, i.e. a smaller close family will give a bigger population that is taken into account. This makes that our assumption is in favor of the suspect.

4.3.3 Population substructures

In our model we assume that a population exists of a finite number of disjoint sub-populations. People from a sub-population mate with other people from the sub-population more often. We know the allele frequencies of all the loci on all the alleles for the entire population. We assume that the actual allele frequencies in the sub-population can differ from the allele frequencies for the entire population. Suppose we have a suspect of a crime and his DNA profile is the same as the DNA-profile obtained from the crime scene. We assume that the actual culprit is from the same sub-population as our suspect. When we want to calculate the likelihood ratio concerning the DNA match for the hypothesis that the suspect leaving the crime stain versus an unknown person (from the same sub-population) leaving the crime stain, we need to know the allele probabilities for the sub-population where the suspect is from.

Our knowledge of the allele frequencies is based on the population as a whole. It is assumed that the alleles of the suspect occur more often in his sub-population than in the whole population. We account for this by using the sampling formula given by Balding and Nichols [12] [13]. The sampling formula gives us the ‘new’ probability of finding allele $a$, when $n$ alleles were sampled of which $x$ alleles were of type $a$:

$$
x\theta + (1 - \theta)p_a, \quad \frac{x\theta + (1 - \theta)p_a}{1 + (n - 1)\theta}, \quad (4.3.1)
$$

where $\theta$ is the co-ancestry coefficient and $p_a$ is the probability of allele $a$ in the population. Or in other words; suppose that there is a long lane of people standing before your desk and you check for each person whether they bear allele $a$, then for the first allele of the first person, where you did not see any alleles $a$ yet (so $x = 0$) and you did not check for any alleles yet (so $n = 0$) the probability of observing allele $a$ is just the probability that you know from a database $p_a$. Suppose that the first allele of this person is $a$, then the sampling formula states that this increases the probability of seeing another allele $a$. The probability of finding allele $a$ when we check his second allele becomes ($x = 1$, $n = 1$), $\theta + (1 - \theta)p_a$. This sampling formula has two nice features, first of all, the individual probabilities for observing the alleles remain to add up to 1; if we let $x_i$ denote the number of times we saw allele $i$ in the first $n$ samples, and $i \in \{1, \ldots, m\}$, then the sum of the new probabilities due to the sampling formula is,

$$
\sum_{i=1}^{m} \frac{x_i \theta + (1 - \theta)p_i}{1 + (n - 1)\theta} = \frac{(\sum_i x_i) \theta + (1 - \theta) (\sum_i p_i)}{1 + (n - 1)\theta} = \frac{n\theta + (1 - \theta)}{1 + (n - 1)\theta} = 1
$$
The second nice feature is, that when \( n \gg 0 \), the probability of observing allele \( a \) is approximately \( x/n \), the fraction of times you have seen allele \( a \).

In order to simulate our population we need to change our allele probabilities given the alleles of the suspect. Our suspect can either be homozygote or heterozygote. In both cases we have to compute the ‘new’ probabilities for a number of cases.

**Homozygote alleles**

\( \mathbb{P}(aa|aa) \) When our suspect has \( aa \) and we want to compute the probability of finding another \( aa \),

\[
\mathbb{P}(aa|aa) = \frac{\mathbb{P}(aa \cap aa)}{\mathbb{P}(aa)}.
\]

We compute this probability following \([12] \ [13] \ [14]\). We start by computing the numerator. We consider each allele in turn. The first allele \( a \) gives probability (using the sampling formula (4.3.1) with \( n = 0 \) and \( x = 0 \)).

\[
\frac{(1 - \theta)p_a}{1 - \theta} = p_a. \quad (4.3.2)
\]

The second allele \( a \) gives \((n = 1 \text{ and } x = 1)\)

\[
\frac{\theta + (1 - \theta)p_a}{1} = \theta + (1 - \theta)p_a. \quad (4.3.3)
\]

The third one \((n = 2, x = 2)\)

\[
\frac{2\theta + (1 - \theta)p_a}{1 + \theta}. \quad (4.3.4)
\]

And the last one \((n = 3, x = 4)\)

\[
\frac{3\theta + (1 - \theta)p_a}{1 + 2\theta}. \quad (4.3.5)
\]

Now, we do the same for the denominator. We will obtain the same expressions as \(4.3.2\) and \(4.3.3\).

Combining all these we can get an expression for our conditional probability.

\[
\mathbb{P}(aa|aa) = \frac{(2\theta + (1 - \theta)p_a) \cdot (3\theta + (1 - \theta)p_a)}{(1 + \theta) \cdot (1 + 2\theta)} \quad (4.3.6)
\]

\( \mathbb{P}(ab|aa) \) Now we do the same for the probability of finding \( ab \) when the suspect has \( aa \). We use the sampling formula to get

\[
\mathbb{P}(ab|aa) = \mathbb{P}(ba|aa) = 2\frac{(1 - \theta)p_a \cdot (2\theta + (1 - \theta)p_a)}{(1 + \theta) \cdot (1 + 2\theta)} \quad (4.3.7)
\]

This probability is the same when we want to know \( \mathbb{P}(ba|aa) \).

\( \mathbb{P}(bb|aa) \) Finding \( bb \) when the suspect has \( aa \).

\[
\mathbb{P}(bb|aa) = \frac{(1 - \theta)p_b \cdot (\theta + (1 - \theta)p_b)}{(1 + \theta) \cdot (1 + 2\theta)} \quad (4.3.8)
\]

\( \mathbb{P}(bc|aa) \) Finding \( bc \) (or \( cb \)) when the suspect has \( aa \).

\[
\mathbb{P}(bc|aa) = 2\frac{p_b(1 - \theta)^2p_c}{(1 + \theta) \cdot (1 + 2\theta)} \quad (4.3.9)
\]

**Heterozygote alleles** Now we look at the probabilities when the suspect has an heterozygote combination (i.e. two different numbered alleles) of alleles on a locus.
\( \mathbb{P}(ab|ab) \) Finding \( ab \) (or \( ba \)) when the suspect has \( ab \).

\[
\mathbb{P}(ab|ab) = \mathbb{P}(ba|ab) = 2 \frac{(\theta + (1 - \theta)p_a) \cdot (\theta + (1 - \theta)p_b)}{(1 + \theta) \cdot (1 + 2\theta)}
\] (4.3.10)

\( \mathbb{P}(aa|ab) \) Finding \( aa \) when the suspect has \( ab \).

\[
\mathbb{P}(aa|ab) = \frac{(\theta + (1 - \theta)p_a) \cdot (2\theta + (1 - \theta)p_a)}{(1 + \theta) \cdot (1 + 2\theta)}
\] (4.3.11)

\( \mathbb{P}(ac|ab) \) Finding \( ac \) (or \( ca \)) when the suspect has \( ab \).

\[
\mathbb{P}(ac|ab) = \mathbb{P}(ca|ab) = 2 \frac{(1 - \theta)p_c \cdot (\theta + (1 - \theta)p_a)}{(1 + \theta) \cdot (1 + 2\theta)}
\] (4.3.12)

\( \mathbb{P}(cd|ab) \) Finding \( cd \) (or \( dc \)) when the suspect has \( ab \).

\[
\mathbb{P}(cd|ab) = \frac{p_c(1 - \theta)^2 p_d}{(1 + \theta) \cdot (1 + 2\theta)}
\] (4.3.13)

\( \mathbb{P}(cc|ab) \) Finding \( cc \) when the suspect has \( ab \).

\[
\mathbb{P}(cc|ab) = \frac{(1 - \theta)p_c \cdot (\theta + (1 - \theta)p_c)}{(1 + \theta) \cdot (1 + 2\theta)}
\] (4.3.14)

In our model we assume that the entire family of our suspect is in the same sub-population as our suspect. We don’t account for sub-populations in the sub-population of the suspect (sub-sub-populations). Therefore, the allele-frequencies in the sub-population are only changed once, when we assign our suspect profile. A simulated profile for, e.g., the father or other relatives from the suspect are drawn from this population and cannot change the allele-frequencies again.

To explain what this sub-population does with our allele-frequencies, we give a short example.

**Example of the change in probabilities when considering sub-populations**

Suppose that at some locus, which we will refer to as Locus 1, there are four possible alleles. The allele frequencies for these alleles are known for the entire population, and are given by,

<table>
<thead>
<tr>
<th>Allele</th>
<th>( p(\text{allele}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
</tr>
</tbody>
</table>

For a certain crime we have a suspect who’s autosomal DNA profile matches with the DNA profile obtained from a stain at the crime scene. The suspect (and the crime stain) has at our locus the allele pair,

\[
\begin{array}{c|c}
\text{Locus 1} & 1 \\
\hline
1 & 2
\end{array}
\]

We assume that the actual culprit is from the same subpopulation as the suspect, remember that the suspect and the culprit are not necessarily the same person. We wonder what the probability is of finding the same allele pair on Locus 1 in this sub-population. Therefore we have to change our allele probabilities since we know the autosomal DNA profile of our suspect. We do this by using the formulas we derived in section 4.3.3 and setting \( \theta \), the co-ancestry coefficient to 0.01 (this value was suggested by the National
Research Council Committee on DNA technology in Forensic Science, [20]). By using formula 4.3.10 we compute the following probability.

$$P\left(\frac{1}{2} \bigg| \text{suspect} = \frac{1}{2}\right) = 2 \cdot \frac{(0.01 + 0.99 \cdot p_1) \cdot (0.01 + 0.99 \cdot p_2)}{1.01 \cdot 1.02}$$

$$= 2 \cdot \frac{(0.109) \cdot (0.208)}{1.0302}$$

$$= 0.044$$

We can use these equations for all the possible combinations of alleles given the suspect profile. The probabilities for all the allele pairs after using our equations are given in table 4.3.1.

<table>
<thead>
<tr>
<th>First allele</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0126</td>
<td>0.0440</td>
<td>0.0628</td>
<td>0.0838</td>
</tr>
<tr>
<td>2</td>
<td>0.0440</td>
<td>0.0440</td>
<td>0.1199</td>
<td>0.1599</td>
</tr>
<tr>
<td>3</td>
<td>0.0628</td>
<td>0.1199</td>
<td>0.0885</td>
<td>0.2283</td>
</tr>
<tr>
<td>4</td>
<td>0.0838</td>
<td>0.1599</td>
<td>0.2283</td>
<td>0.1561</td>
</tr>
</tbody>
</table>

Table 4.3.1: Allele-pairs probabilities after using our equations to account for subpopulations

This table is symmetric. The probability of seeing 1/2 is the same as the probability of seeing 2/1. The probabilities for all the different combinations add up to 1.

If we assume that the population is homogeneous and subpopulations do not exist, then we obtain other probabilities using the product rule. These are given in table 4.3.2.

<table>
<thead>
<tr>
<th>First allele</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.01</td>
<td>0.04</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>0.04</td>
<td>0.04</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>0.06</td>
<td>0.12</td>
<td>0.09</td>
<td>0.24</td>
</tr>
<tr>
<td>4</td>
<td>0.08</td>
<td>0.16</td>
<td>0.24</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Table 4.3.2: Allele-pairs probabilities without sub-population assumption

We see that the difference between the numbers are quite small for allele frequencies in this range. If we look at the profile probability of our suspect profile, 1/2, we see that we have 0.044 when we do take sub-population into account and 0.04 when we don’t.

4.3.4 The model

We can divide our model into two parts, one where a family tree is made and one where autosomal DNA profiles are distributed among the people in the tree.

Making the family tree

To make the family tree we start with one person, the root of the tree. This person is assigned an age (e.g. 20), and the corresponding probabilities for becoming father and dying in that particular year. When there is only one living person in our simulation there is a very limited amount of possible events that can occur each year. First of all, the person can die during that year. The other thing that can happen is that the person becomes father of a son that year. It is also possible that both events occurred. The last thing that can happen is that the year passes and that none of the above events happened. The number of events that occur in a year are drawn from a Poisson distribution with parameter $\lambda$ equal to the sum of all probabilities for the different events. When we only have one person, $\lambda$ is given by

$$\lambda = P(\text{the person dies}) + P(\text{the person becomes father of a son}).$$
When the first person gets a son we have two people in our simulation. They both have separate probabilities for dying and becoming father. The parameter $\lambda$ in these years is computed as the sum of all four of these probabilities.

Now we have drawn the number of events each year, we need to decide which events occur. Suppose that $\lambda$ consists of four probabilities, 0.0625, 0, 0.3, 0.0625 and 0.25. Then, $\lambda$ is 0.675 and we could represent it by the following line (figure 4.3.1).

![Figure 4.3.1: Representation of $\lambda$](image)

If it turned out that there were 2 events that year, then, for the first event we draw a random number from a uniform distribution on (0, 0.675). Suppose we get 0.5. In figure 4.3.1 we divided $\lambda$ into little blocks. 0.5 falls in the block belonging to the event with probability 0.25. We conclude that the event with probability 0.25 occurred. Suppose that this belongs to the event person 2 dies. Now we need to decide what the second event was. If we assume that one of the two events with probability 0.0625 was the event person 2 gets a son then we put the probability that this happens to 0. We are left with two possible events, one where person 1 gets a son and one where person 1 dies. We draw a random number from a uniform distribution on (0,0.3625). If the outcome is less or equal to 0.0625 the event that occurred is person 1 gets a son and when the outcome is greater than 0.0625 then the event that occurred is person 1 dies.

We let time pass and count the number of people that are alive. For all persons that are simulated we save how they are related to each other. At some point in the future the number of people that are alive in the simulation is equal to or greater than the number $N$, the total number of people estimated to bear a certain Y-chromosomal DNA profile. At that point we stop our simulation. We have a family tree that links $N$ living people. Usually, the number of people that are linked exceeds this number $N$ since there is also a number of people that already died in the tree.

**NOTE** After these simulations were done, it appeared to us that it might be a better idea to simulate the family tree generating as a Poisson process, where the time between events was exponentially distributed with parameter $\lambda$. Each moment that an event occurred we could decide which event that was just like we did now, and than determine the new parameter $\lambda$ and so on. We ran one simulation using this new way of simulating, and the results were similar to the results using the above method. Since it would take a long time to do all simulation for results using this slightly different method we decided to use the results that were generated using the above technique (note that this method only changes the way we make our family tree, not the assignment of the autosomal profiles!).

**Assigning autosomal DNA-profiles**

The family tree is build on the assumption that all people that bear a certain Y-chromosomal DNA profile are related to each other. Now we assign an autosomal DNA profile to one of the people in the tree and simulate how it is distributed over the tree. Remember that the profile that is assigned to the suspect changes the allele-frequencies that we will use during the rest of the simulation due to the assumption on the existence of subpopulations. This is done in two parts. First we simulated autosomal DNA profiles to the father of the suspect, then to his father’s father, . . . , etc, until we reach the root of the tree. After that we simulate autosomal DNA profiles for all the other people in the tree.

**1st part** For the father of the suspect we know that he contributed half of the DNA profile of the suspect, i.e. half of the DNA profile of the suspect comes directly from his father. The other half from
his mother. We assign these halves randomly, based on the assumption of Mendelian inheritance. For example, when the suspect is assigned the following profile (on four loci).

\[
\begin{array}{cccc}
1 & 2 & 3 & 4 \\
5 & 6 & 7 & 8 \\
\end{array}
\]

Table 4.3.3: The suspect’s profile

His father might get

\[
\begin{array}{cccc}
1 & 3 \\
6 & 8 \\
\end{array}
\]

which implies that his mother will get

\[
\begin{array}{cccc}
2 & 4 \\
5 & 7 \\
\end{array}
\]

Now we draw the empty parts of the DNA profile of the suspect’s parents from the distribution of allele combinations on that locus, given the allele they already have, based on the formulas in section 4.3.3. The parents might get the profiles in table 4.3.4.

\[
\begin{array}{cccc}
1 & 5 & 3 & 4 \\
1 & 6 & 4 & 8 \\
1 & 2 & 7 & 4 \\
5 & 3 & 7 & 8 \\
\end{array}
\]

Table 4.3.4: The profiles of the father and mother of the suspect

Now we continue with the profile of the father, for this profile we simulate a father’s and a mother’s profile the same way as we did here. This procedure is repeated until we reach the root of the tree. This way we have simulated the autosomal DNA profiles of all the people in the direct paternal (and their wives) line of the suspect to the root of the tree.

2nd part Now we have simulated the autosomal DNA profiles of the suspect and of the people between the suspect and the root of the tree, we can simulate the DNA profiles of all the other people in the tree. We start with the man at the root of the tree. We already simulated the autosomal DNA profile of one of his sons. Suppose he has more than one son. In that case we simulate the profile of these remaining sons by randomly taking allele-scores from the known autosomal profiles from the root of the tree and his wife. This way we simulate all the profiles in the family tree top-down. For everybody who did not already have a autosomal profile there are two possible scenarios. Either the profile of his mother is already known since we already simulated the DNA profile of one of the person’s brothers, or it is still unknown in which case we have to simulate it. The profile of this person’s mother is drawn randomly using the probabilities on all the allele combinations that were made by assigning a suspect profile. Hence, we use the allele frequencies in the database the assigned suspect’s profile, and derive the probabilities as in table 4.3.1. This means that on each locus, we will draw a combination of alleles (1-2, 1-1, etc.) using the probabilities that belong to these combinations. If we do this at each locus, we have simulated a DNA profile for this mother.

For example, suppose that the suspect from the first part (with profile as in table 4.3.3) has a brother. The profile of his brother will be assigned by randomly selecting from his father’s and mother’s profiles (see table 4.3.4). A possible outcome is

---

1Mendelian inheritance is a scientific theory of how hereditary characteristics are passed from parent organisms to their offspring. This theoretical framework was derived from the work of Gregor Johann Mendel.

31
Counting the number of matching profiles  After the family tree simulation and the assigning of autosomal profiles we can count the number of people in the tree who have the exact same autosomal profile as the one that was given to our suspect. We only count the number of matching profiles of the people that are alive in our tree.

Since we want to examine the dependency between the autosomal and the Y-chromosomal DNA profiles we count the number of matching profiles of all the living people in our tree and the number of matching profiles of all the people in the tree except for the people who are ‘closely’ related to the suspect. For example, we could neglect all the matching profiles from people who share the grandfather of the suspect as an ancestor.

Example of the model

In order to visualize all of the above we give a short example. Suppose that the number of people who bear a certain Y-chromosomal is known to be 8 in a certain population. To link all of them we begin with one root. We then simulate a tree according to section 4.3.4 ‘Making the family tree’. The simulated family tree turned out to be as in figure 4.3.2.

Our suspect is the last person added, in this case person 12. In this case, his autosomal profile is simulated using the database frequencies and turned out to be as in table 4.3.5 (normally the suspect profile is known and we would not need to simulate this).

This profile is used to change all the probabilities for pairs of alleles. We use the formulas we derived in section 4.3.3. Now, all the profiles of the people (and the corresponding mothers) who are in between of the root of the tree and the suspect are simulated. For each person we take one of the two alleles of his son randomly. The other allele of the son is assigned to his mother.
Now we have to draw the other allele on the loci of the parents. These are drawn conditioned on the first allele they got. When both persons have a complete profile we go to the next person in the tree (the father of the father we just assigned a profile). We repeat this procedure until we reach the root of the tree. The profiles turned out to be as in table 4.3.6.

<table>
<thead>
<tr>
<th>person</th>
<th>profile</th>
<th>profile of the person’s wife</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>D22 D10</td>
<td>D22 D10</td>
</tr>
<tr>
<td>1</td>
<td>16 13</td>
<td>16 13</td>
</tr>
<tr>
<td>2</td>
<td>16 15</td>
<td>16 15</td>
</tr>
<tr>
<td>3</td>
<td>16 14</td>
<td>16 15</td>
</tr>
<tr>
<td>4</td>
<td>16 13</td>
<td>16 13</td>
</tr>
<tr>
<td>5</td>
<td>16 15</td>
<td>16 15</td>
</tr>
<tr>
<td>6</td>
<td>16 15</td>
<td>16 15</td>
</tr>
<tr>
<td>7</td>
<td>16 15</td>
<td>16 15</td>
</tr>
<tr>
<td>8</td>
<td>16 15</td>
<td>16 15</td>
</tr>
<tr>
<td>9</td>
<td>16 15</td>
<td>16 15</td>
</tr>
<tr>
<td>10</td>
<td>15 15</td>
<td>15 15</td>
</tr>
<tr>
<td>11</td>
<td>15 15</td>
<td>15 15</td>
</tr>
<tr>
<td>12</td>
<td>16 13</td>
<td>16 13</td>
</tr>
</tbody>
</table>

Table 4.3.6: The profiles of everybody in the paternal line from the suspect to the root of the tree.

Since we have reached the top of the tree now we can do the second part of the profile assignment, that of all the other men in the tree, starting in the top of the tree. The complete profile assignment is given in table 4.3.7.

<table>
<thead>
<tr>
<th>person</th>
<th>profile</th>
<th>profile of the person’s wife</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>D22 D10</td>
<td>D22 D10</td>
</tr>
<tr>
<td>1</td>
<td>16 13</td>
<td>16 13</td>
</tr>
<tr>
<td>2</td>
<td>16 14</td>
<td>16 14</td>
</tr>
<tr>
<td>3</td>
<td>16 15</td>
<td>16 15</td>
</tr>
<tr>
<td>4</td>
<td>16 13</td>
<td>16 13</td>
</tr>
<tr>
<td>5</td>
<td>16 15</td>
<td>16 15</td>
</tr>
<tr>
<td>6</td>
<td>16 15</td>
<td>16 15</td>
</tr>
<tr>
<td>7</td>
<td>16 15</td>
<td>16 15</td>
</tr>
<tr>
<td>8</td>
<td>16 15</td>
<td>16 15</td>
</tr>
<tr>
<td>9</td>
<td>16 15</td>
<td>16 15</td>
</tr>
<tr>
<td>10</td>
<td>15 15</td>
<td>15 15</td>
</tr>
<tr>
<td>11</td>
<td>15 15</td>
<td>15 15</td>
</tr>
<tr>
<td>12</td>
<td>16 13</td>
<td>16 13</td>
</tr>
</tbody>
</table>

Table 4.3.7: All the profiles of the people in our tree. Persons with the same profile as the suspect (person 12) are highlighted.

Remember that only the profiles of the men are compared with the suspect’s profile. The profiles of the corresponding wives are only needed when assigning the profiles of their children. For some men we did not simulate a wife for the simple reason that the person doesn’t have any children.

The persons with numbers 1 and 2 died before the end of the simulation. This means that we won’t use them when we count the number of corresponding profiles. So, although the profile of person 1 is the
same as our suspect profile, we won’t use it when counting the number of matching profiles since the person died before the end of the simulation. The total number of people that is alive in our tree is ten. We see that there is only one living person with the same autosomal profile as our suspect. It is the suspect’s brother and their profiles are highlighted in green. Now suppose we neglect everybody who comes from the same grandfather as our suspect. In that case we don’t look at the profiles of everybody who comes from person number 4. We neglect the profiles of 4, 7, 11 and 12 (which is neglected anyway, since we compare it with all the other profiles and not with itself). In that case there are no corresponding profiles. We can do this profile simulation $M$ times, and look at the results on the number of matching profiles before and after we neglected the close relatives. Our results might turn out to be like in table 4.3.8 when $M = 1000$.

<table>
<thead>
<tr>
<th>number of matching profiles</th>
<th># simulations in related population</th>
<th># simulations in related population without close relatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>400</td>
<td>490</td>
</tr>
<tr>
<td>1</td>
<td>400</td>
<td>350</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>75</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.3.8: Results after simulating the profile simulation 1000 times
4.4 Data

In order to examine the dependency between an autosomal and a Y-chromosomal DNA profile we made a simulation model. The data that are used are provided in this section.

4.4.1 Probability of dying

For the probability of dying on each age we used data from the Statistics Netherlands (CBS) [11]. These data consider all ages up to one-hundred and give the probability that a man from the Netherlands dies in the year that he has a certain age. Since the CBS only gives these probabilities for people from 0 to 100 years, and give a probability of dying at an older age than 120, we have to add some part. We assume that it is impossible that people become older than 120 years and added this last part (from 100 till 120 years) by linear continuation.

\[ E[\text{number of sons}] = \sum_{t=1}^{120} P(\text{being alive at time } t) \cdot P(\text{getting a son at } t) \]

\[ = \sum_{t=20}^{40} P(\text{being alive at time } t) \cdot 0.0625 \]

\[ = \sum_{t=20}^{40} \prod_{i=0}^{t-20} (1 - P(\text{dying at age } i | \text{being alive at time } i - 1)) \cdot 0.0625 \approx 1.29. \]

This is clearly in favor of a suspect, since large families will lead to more matching profiles. A smaller number of sons often makes it hard to form a family tree, almost every time all people in the specific family die before the tree is big enough. A more realistic choice would be much smaller than our estimate. This assumption doesn’t match with actual real life situations. In practice, there are men that get a son before the age of 20 and after the age of 40 years. The probability of getting a son is not uniformly distributed over the years. It is more likely that a man gets a son around the age of 30 than at the age of 20. Another thing that we see in practice but which is not in our model is that there is a certain dependence on whether you already have a son and that you will get another one; People often decide whether they want any children. It is important to note that these assumptions which do not match with a practical setting are not that important. These assumptions are only made to link all the people with the same Y-chromosomal DNA profile. At which age they get children is not that important. One might argue that getting children at a young age makes it more likely that generations follow up each other faster which would make the number of generations that is needed to complete the family tree is lower. On the other hand, the uniform distribution between the ages of 20-40 can be regarded as quite conservative. The number of children one gets is more important, since this really affects how the family tree grows.

4.4.2 Son-probability

We assume that every man between 20 and 40 years has a probability of 0.0625 to get a son each year. We chose this number by trial and error. We started with a probability of 0.5 and divided it by 2 until we got acceptable results. The mean number of children a father has in the Netherlands is 1,719 [10]. This corresponds with a mean number of (almost) 0.86 sons. When using the probability of becoming father of a son we chose during the ages at which this probability holds, we can compute the mean number of sons fathers will get. We therefore need the probabilities of dying. The mean number of sons is computed the following way,

4.4.3 Allele frequencies

For the allele frequencies we used data samples that were gathered by prof dr Peter de Knijff from Leiden University. Analysis was performed both in Leiden and at the Netherlands Forensic Institute (NFI). We only used data on six different loci, which we chose randomly. Adding more alleles would be possible,
but we will show that it won’t change any conclusions. The numbers we chose here are representative in the sense that they are used in real cases but can be substituted for any other numbers. If you want to simulate with a profile with information on more than eight loci it won’t be necessary to combine the profile frequencies. We won’t give this data (the one we used) in this report due to the fact that they will be published at another moment. To give the reader an idea of the real data, we will give data that were used in practice before these data became available and which are similar to the ones used in the simulation; this data can be found in table 4.4.1.

<table>
<thead>
<tr>
<th></th>
<th>D3S1358</th>
<th>VWA</th>
<th>D16S539</th>
<th>D18S51</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.015</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.123</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.067</td>
<td>0.011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.340</td>
<td>0.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.279</td>
<td>0.134</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.002</td>
<td>0.162</td>
<td>0.108</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.091</td>
<td>0.067</td>
<td>0.013</td>
<td>0.182</td>
</tr>
<tr>
<td>15</td>
<td>0.281</td>
<td>0.076</td>
<td></td>
<td>0.117</td>
</tr>
<tr>
<td>16</td>
<td>0.253</td>
<td>0.203</td>
<td></td>
<td>0.152</td>
</tr>
<tr>
<td>17</td>
<td>0.193</td>
<td>0.303</td>
<td></td>
<td>0.141</td>
</tr>
<tr>
<td>18</td>
<td>0.167</td>
<td>0.223</td>
<td></td>
<td>0.071</td>
</tr>
<tr>
<td>19</td>
<td>0.011</td>
<td>0.110</td>
<td></td>
<td>0.039</td>
</tr>
<tr>
<td>20</td>
<td>0.002</td>
<td>0.013</td>
<td></td>
<td>0.026</td>
</tr>
<tr>
<td>21</td>
<td>0.004</td>
<td></td>
<td></td>
<td>0.006</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td>0.004</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td>0.002</td>
</tr>
</tbody>
</table>

Table 4.4.1: Allele frequencies

In table B.1.1 which can be found in appendix B.1 we give the possible alleles on each locus.

4.5 Results

In this section we will give the results of our simulations. We will compare these with the random match probabilities of the used suspect profile. The factor $\theta$ that we use for the sub-populations is set to 0.01 in all simulations. This value is used in standard casework of the NFI, and is based on recommendation of the report of the National Research Council of the US [20]. We will start with some general results that show the effect of some of our assumptions on the results.

4.5.1 General results

In this section we will discuss what the effect of excluding all the people that came from the same grandfather when counting the number of matching profiles is.

Excluding the close family

When comparing the results on the number of matching profiles between a random population and a related population we exclude the close family from the related population. The close family is defined as all the men that have the same grandfather of our suspect as an ancestor. It is known that it is far more likely for the brothers of our suspect to have a matching autosomal DNA profile than some distant cousin. However, a person usually has far more distant relatives than brothers. Hence, it is not straightforward to reason whether most matching profiles will be found among close or among distant relatives. When simulating 1000 profile simulations over 50 family trees containing 1000 living men (so we have 50 family trees containing 1000 living men, and on each of these trees we do the profile simulation
1000 times) where the suspect profiles is as in table 4.5.3 And we look at the tree with the most extreme results (where the most extreme result is the result where there was the biggest difference in the results on the number of matching profiles between the related population and the related population without close relatives), we got the following result on the number of matching profiles (table 4.5.1).

<table>
<thead>
<tr>
<th>Number of matching profiles</th>
<th>Related population</th>
<th>Without close family</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>422</td>
<td>903</td>
</tr>
<tr>
<td>1</td>
<td>360</td>
<td>84</td>
</tr>
<tr>
<td>2</td>
<td>168</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.5.1: The number of simulations with a certain number of matching profiles with and without the close family, extreme case

From table 4.5.1 we see that the majority of the matching profiles can be found in the close family of our suspect. The mean number of matching profiles lowers from 0.848 to 0.011. Since this is an extreme case, we should also look at the differences over all the simulated trees. In Figure 4.5.1 we put the number of men in the close family on the x-axis and plotted them against the number of times we found zero matches in our profile simulation. We give the results for a related and a related-without-close-family population and added a least squares line through our points.

![Figure 4.5.1: The size of the close family against the number of times we have seen zero matches](image)

Figure 4.5.1: The size of the close family against the number of times we have seen zero matches

From the figure it is clear that, as expected, the number of people in the close family has a big influence on the number of times we see no matching profiles in our family tree. When the number of people that are in the close family grows, the number of times we see no matching profiles lowers (or the number of times we see matching profiles grows). We see that when we exclude the close family we don’t get this result. Since we want to find a way to combine the autosomal and a Y-chromosomal DNA profile frequencies, which means that we need to show how dependent the two are of each other (and in the best case that they are independent), this is a good result. Remember that we only give this figure to help the reader understand the dependence between the number of people in the close family and the number of times we see no matching profiles in the tree. It is not necessarily a representation on how the number...
of matching profiles in- or decreases as the number of living people in the close family in- or decreases. In the following sections we will give the results for profile simulations with different profiles on different number of living people in the family tree structure. In table 4.5.2 we give the parameters we used in the simulations.

<table>
<thead>
<tr>
<th>simulation</th>
<th>number of loci of the profile</th>
<th>number of living men in the tree</th>
<th>number of times we did the profile simulation for each tree</th>
<th>number of trees we did the profile simulations on</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>100</td>
<td>1000</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1000</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1000</td>
<td>1000</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>10000</td>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>10000</td>
<td>100</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 4.5.2: Parameters for the different simulations

The number of family trees we will use for each simulation is not very high (between 20 and 50) but since it takes a lot of time to run our simulation over each family tree and the fact that most family trees will approximately have the same shape we expect the number is sufficient enough to work with.

4.5.2 A profile on three loci

In this section we do simulations using a profile of 3 loci. The profile we used, which was randomly drawn based on the database frequencies, is as follows,

<table>
<thead>
<tr>
<th>D22S1045</th>
<th>D10S1248</th>
<th>D3S1358</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>16</td>
<td>13</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 4.5.3: The profile on 3 loci assigned to our suspect

The random match probability of this profile is computed by taking into account the changed probabilities for all the allele combinations. We want to compute the following probability.

\[
P \left( \begin{array}{c|c|c|c|c}
\text{unknown stain donor} = & D22 & D10 & D3S \\
15 & 16 & 13 & 14 \\
\text{suspect} = & D22 & D10 & D3S \\
15 & 16 & 13 & 15 \\
\end{array} \right), I, H_d
\]

Where \(H_d\) is the hypothesis that the suspect is not the donor to the stain, and \(I\) represents that the suspect and the unknown stain donor have the same Y-profile.

Therefore we change our probabilities for finding the combinations of certain alleles as explained in section 4.3.3. We assume that the alleles on the different loci are independent, and compute our profile probability by multiplying these different probabilities.

\[
\text{Profile probability} = P \left( \begin{array}{c|c}
15 & 16 \\
16 & 16 \\
\end{array} \right) \cdot P \left( \begin{array}{c|c}
13 & 13 \\
13 & 13 \\
\end{array} \right) \cdot P \left( \begin{array}{c|c}
14 & 14 \\
15 & 15 \\
\end{array} \right)
\]

\[
= 0.2432 \cdot 0.1108 \cdot 0.0649
\]

\[
= 0.0017
\]

We simulated this profile over family trees containing 100 and 1000 men.

Family trees relating 100 living men

We run our profile simulation 1000 times over 50 family trees that related 100 living men. We counted the number of matching profiles for all these men. To give an idea of the results, we give the following figure.
(figure 4.5.2) representing the frequency distribution of the number of matching profiles but only for ten trees. We see the number of matching profiles against the relative frequency of finding this number of matching profiles for the ten different trees. The relative frequency is computed as the number of times we have seen the number of matching profiles divided by the total number of times we simulated the profiles over the tree (1000). Each color represents another family tree structure. The question that we need to answer here is: do these number differ from what we would expect when we would regard an unrelated population? What is the probability that we find 0,1,... matching profiles in a random population of 100 people? We can find these numbers from the probability density function of the binomial distribution with \( N \) the number of people and \( p \) the profile probability. Since the results in the figure are without the close family, \( N \) is less than 100. The number of people in the close family is different for each tree so we take the mean number of people in the close family over all the trees and subtract it from 100 for our parameter \( N \). It turns out that for these trees, \( N = 94 \). In the figure are the results that we expect to find in an unrelated random population, i.e. the numbers from the binomial distribution are given with a dotted line.

![Figure 4.5.2: The estimated probability for the number of matching profiles for 10 trees with 100 men without close family](image)

We see that the mean relative frequencies for these trees are 85% to find no matches in the tree, 12% to find one match and 3% to find more than one match. In the next table (4.5.4) we give the mean relative frequency over all the 50 trees as well as the minimum and the maximum over all the separate trees and compare it with the probability we expect to find in a random unrelated population.

**Family trees relating 1000 living men**

We did the same computations as in section 4.5.2, but now with 50 trees of 1000 living people and only 100 profile simulations on each tree. We used the same profile, that had a random match probability of 0.0017. We give a similar figure as we did with 100 living men, showing the results for 10 randomly selected different family tree structures and the expected numbers for a random unrelated population. Figure 4.5.3 shows the approximate probabilities of 10 different trees. The results seem to be more differentiated than before, notice that the y-axis is between 0 and 0.4 instead of from 0 to 1. The table with the approximate probabilities computed over all 50 family trees are given in table 4.5.5. We see little difference between the mean frequency over 50 trees and the probability for finding that number of matching profiles in a random unrelated population. The results of profiles on four and five loci are similar and can be found in appendix B.2.
### Table 4.5.4: The mean, maximum and minimum relative frequencies computed over 50 trees with the probabilities for a random unrelated population.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Max</th>
<th>Min</th>
<th>random unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(no matching profiles)</td>
<td>0.8553</td>
<td>0.893</td>
<td>0.830</td>
<td>0.8483</td>
</tr>
<tr>
<td>P(1 matching profile)</td>
<td>0.1183</td>
<td>0.145</td>
<td>0.086</td>
<td>0.1397</td>
</tr>
<tr>
<td>P(2 matching profiles)</td>
<td>0.0224</td>
<td>0.034</td>
<td>0.011</td>
<td>0.0114</td>
</tr>
<tr>
<td>P(3 matching profiles)</td>
<td>0.0033</td>
<td>0.012</td>
<td>0</td>
<td>0.0006</td>
</tr>
<tr>
<td>P(4 matching profiles)</td>
<td>0.0005</td>
<td>0.002</td>
<td>0</td>
<td>0.0000</td>
</tr>
<tr>
<td>P(5 matching profiles)</td>
<td>0.0001</td>
<td>0.001</td>
<td>0</td>
<td>0.0000</td>
</tr>
<tr>
<td>P(6 matching profiles)</td>
<td>0.0000</td>
<td>0.001</td>
<td>0</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

### Figure 4.5.3: The estimated probability for the number of matching profiles for 10 trees with 1000 men without close family

### Table 4.5.5: The mean, maximum and minimum computed over 50 trees

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Max</th>
<th>Min</th>
<th>random unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(no matching profiles)</td>
<td>0.2076</td>
<td>0.31</td>
<td>0.12</td>
<td>0.1753</td>
</tr>
<tr>
<td>P(1 matching profile)</td>
<td>0.2932</td>
<td>0.39</td>
<td>0.23</td>
<td>0.3055</td>
</tr>
<tr>
<td>P(2 matching profiles)</td>
<td>0.2286</td>
<td>0.32</td>
<td>0.15</td>
<td>0.2659</td>
</tr>
<tr>
<td>P(3 matching profiles)</td>
<td>0.1452</td>
<td>0.24</td>
<td>0.07</td>
<td>0.1542</td>
</tr>
<tr>
<td>P(4 matching profiles)</td>
<td>0.0762</td>
<td>0.15</td>
<td>0.03</td>
<td>0.0670</td>
</tr>
<tr>
<td>P(5 matching profiles)</td>
<td>0.0306</td>
<td>0.07</td>
<td>0</td>
<td>0.0232</td>
</tr>
<tr>
<td>P(6 matching profiles)</td>
<td>0.0118</td>
<td>0.04</td>
<td>0</td>
<td>0.0067</td>
</tr>
<tr>
<td>P(7 matching profiles)</td>
<td>0.0052</td>
<td>0.02</td>
<td>0</td>
<td>0.0017</td>
</tr>
<tr>
<td>P(8 matching profiles)</td>
<td>0.0012</td>
<td>0.01</td>
<td>0</td>
<td>0.0004</td>
</tr>
<tr>
<td>P(9 matching profiles)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0001</td>
</tr>
<tr>
<td>P(10 matching profiles)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0000</td>
</tr>
<tr>
<td>P(11 matching profiles)</td>
<td>0.0004</td>
<td>0.01</td>
<td>0</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Table 4.5.5: The mean, maximum and minimum computed over 50 trees
Conclusion

Our results show that the difference in the number of matching profiles, when we exclude the close family, are similar over the different trees. We also see that the number of times we see \( i \) matching profiles in the trees are similar to what we would expect in a random unrelated population. We conclude that our simulations give us no reason to expect a different number of people with a similar autosomal DNA profile as our suspect in the population which has the same Y-chromosomal DNA profile as the suspect (where the close family is excluded), than in a random unrelated population.

4.5.3 The profile probability

In this section we will look at the differences between the profile probability as given for a random population, which is just the product of the probabilities of the allele-combinations on the loci, and the profile probability as given by our family trees. Is there a (big) difference between the (approximate) profile probabilities for different family tree structures? What happens to the approximate profile probabilities when we take family trees that link 10 times as many living men?

A profile on three loci

Here we will discuss the results of our simulations where the suspect profile was known on three loci.

Family trees linking 100 living men

We assigned the following profile to our suspect;

\[
\begin{array}{ccc}
D22S1045 & D10S1248 & D3S1358 \\
15 & 13 & 14 \\
16 & 13 & 15 \\
\end{array}
\]

We saw that the profile probability of this profile (given that our suspect has the profile) is 0.0017. Now we want to compare this profile probability with the profile probabilities we have in our family trees. We made 50 different family tree structures linking 100 living men. Over these family trees we did 1000 profile simulations. We can compute the approximate profile probabilities based on the results of our profile simulation for each tree.

For each tree we did multiple profile simulations. For each profile simulation we got the number of times we found 0, 1, \ldots, matching profiles in his family tree. If we count the total number of times we have seen a matching profile over all the simulations and divide this number by the number of simulations we did and the number of people that are alive in the family tree, we get an estimation for the random match probability in the family tree. If we do this for all the family trees we did simulations on, than we can look at the differences between the trees and the estimated mean random match probability over all the trees. The mean and the standard deviation of the profile probabilities over all these trees are in table 4.5.6.

| profile probability | 0.0017482 |
| mean profile probability over the trees | 0.0017490 |
| standard deviation | 0.0001760 |

Table 4.5.6: The mean and standard deviation of the profile probabilities on 50 trees linking 100 people and profile information on 3 loci when we neglect the close family.

We also give the results in a histogram, since we deal with very small numbers, we give the \( \log_{10} \) of our results, see figure 4.5.4.
Figure 4.5.4: A histogram showing the $\log_{10}$ of our random match probability of all the trees. The red bar represents the profile probability of the suspect’s profile.

The red bar represents the profile probability as given by our formula. The figure shows the same result as our table. The mean of the blue bars is close to the red bar, and the standard deviation is quite small.

Family trees linking 1000 men

In our simulation we used the same profile as our suspect profile as in the trees linking 100 men.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>D22S1045</td>
<td>15</td>
</tr>
<tr>
<td>D10S1248</td>
<td>13</td>
</tr>
<tr>
<td>D3S1358</td>
<td>14</td>
</tr>
<tr>
<td>VWA</td>
<td>16</td>
</tr>
</tbody>
</table>

The results of this simulation are presented in table 4.5.7.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>profile probability</td>
<td>0.0017482</td>
</tr>
<tr>
<td>mean profile probability over the trees</td>
<td>0.0017650</td>
</tr>
<tr>
<td>standard deviation</td>
<td>0.0001403</td>
</tr>
</tbody>
</table>

Table 4.5.7: The mean and standard deviation of the profile probabilities on 50 trees linking 1000 people and profile information on 3 loci when we neglect the close family.

The result is satisfying. The profile probability over the trees is approximately the same as the profile probability of the suspect profile. The standard deviation is very small compared with the mean.

A profile on 4 loci

Now we do the same computations but on a profile on 4 loci. The profile we used was

<table>
<thead>
<tr>
<th>Locus</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>D22S1045</td>
<td>15</td>
</tr>
<tr>
<td>D10S1248</td>
<td>16</td>
</tr>
<tr>
<td>D3S1358</td>
<td>19</td>
</tr>
<tr>
<td>VWA</td>
<td>16</td>
</tr>
<tr>
<td>VWA</td>
<td>15</td>
</tr>
<tr>
<td>VWA</td>
<td>14</td>
</tr>
<tr>
<td>VWA</td>
<td>14</td>
</tr>
<tr>
<td>VWA</td>
<td>17</td>
</tr>
</tbody>
</table>

Which has a profile probability $5.1814 \cdot 10^{-6}$.
Table 4.5.8: The mean and standard deviation of the profile probabilities on 20 trees linking 1000 people and profile information on 4 loci when we neglect the close family

| profile probability | $5.1814 \times 10^{-6}$ |
| mean profile probability over the trees | $5.8000 \times 10^{-6}$ |
| standard deviation | $7.5835 \times 10^{-6}$ |

Family trees linking 1000 men

We see that the standard deviation is of the same order as the mean profile probability over the trees. The mean profile probability over the trees is approximately the same as the profile probability. The relatively high standard deviation can be explained by the fact that the profile probability is very small. Therefore, in a lot profile simulations over a tree, we will find no result at all. In some cases we find one or two matching profiles. This could explain the high value of the standard deviation. The standard deviation is still of the same order as the mean. If it would be of order $10^{-5}$, the difference would be really disturbing.

Family trees linking 10 000 men

| Profile probability | $5.1814 \times 10^{-6}$ |
| mean profile probability over the trees | $4.2727 \times 10^{-6}$ |
| standard deviation | $1.9022 \times 10^{-6}$ |

Now we use trees that links 10 times as many people, we see that the standard deviation becomes smaller again. The results are surprising in the sense that the profile probability for a random person is higher than the mean profile probability over the trees. However, they are of the same order so the difference is not really significant.

A profile on 5 loci

Lastly, we will give some results for a profile on 5 loci with a tree that links 10 000 people.

<table>
<thead>
<tr>
<th>D22S1045</th>
<th>D10S1248</th>
<th>D3S1358</th>
<th>VWA</th>
<th>D16S39</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>13</td>
<td>17</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>16</td>
<td>13</td>
<td>14</td>
<td>14</td>
<td>12</td>
</tr>
</tbody>
</table>

Which has profile probability $1.5722 \times 10^{-5}$.

| profile probability | $1.5722 \times 10^{-5}$ |
| mean profile probability over the trees | $1.6182 \times 10^{-5}$ |
| standard deviation | $4.2853 \times 10^{-6}$ |

Table 4.5.10: The mean and standard deviation of the profile probabilities on 20 trees linking 10 000 people and profile information on 5 loci when we neglect the close family

Once again we see a satisfying result. The profile probability over the trees is of the same order as the profile probability of the profile. The standard deviation is small when we compare it with the mean. The profile probability is also not that small for the number of people that we linked (10 000).
Conclusion

We see that the profile probability that we can compute over the trees is of the same order as the profile probability of the input profile in all our simulations. In one case, the standard deviation was notably high. However, we could explain this high value due to the fact that the profile probability was quite low when we compared it with the number of people we linked in the family tree. Therefore we conclude that there is no evidence to suggest that the approximated autosomal DNA profile probability for people that share the same Y-chromosomal DNA profile differs from the profile probability for a random individual.

4.6 Conclusion and Discussion

Our model shows that the profile probabilities that come from our trees are of the same order as the profile probability of the profile given by our equations. The standard deviation is either small or explainable high with regard to our approximate probability. We conclude that the dependence between autosomal DNA profiles and Y-chromosomal DNA profiles may be neglected for persons other than the close relatives of a suspect, where the close relatives are the people that come from the same grandfather. In a forensic report one could write.

“Given that we exclude all the family of the suspect that comes from the same grandfather, we can multiply the profile frequency of the autosomal DNA profile with the profile frequency of the Y-chromosomal DNA profile to find a profile frequency of the two profiles combined.”

We did not do a sensitivity analysis on the model since all parameters were chosen to benefit the defending party. We think a sensitivity analysis would make sense when the parameters would be more realistic. For further research regarding the investigation of the dependence between autosomal and Y-chromosomal DNA profiles using this model we would advise to test it on other sizes of the population. In a very small population, the factor $\theta$ to account for subpopulations becomes bigger and for very big populations it will be possible to test autosomal DNA profiles on more loci.
Chapter 5

The combination of low template DNA profiles

5.1 Introduction

When the amount of DNA found on a crime scene is very small, or when DNA has been degraded due to circumstances, low template DNA analysis is used. The toy problem which is the basis of our approach is the following:

Problem  A body is found in the forest, and it appears to have been laying there for quite some time. A blood stain is found of which it is believed that it belongs to the culprit. Since the stain could have been affected by bad conditions, four different extracts are obtained from the same stain. And from this extracts, four DNA profiles are made. What is the combined evidential value of the DNA profiles? And how to we interpret these low template DNA profiles?

5.1.1 What is Low Template DNA?

Low template DNA or low copy number DNA (LCN) refers to samples that contain less than the 250 pg (picogram \(10^{-12}\) gram) required to produce a complete profile. Those small amounts are usually found when we have ‘touch’ DNA, meaning DNA that has been transferred from a person simply by touching an item. It is also of use in degraded samples, for example when a stain has been laying somewhere for a long time; or under bad circumstances (e.g. moist).

A low template DNA analysis consists of multiple analysis on the same profile. This will lead to multiple (slightly) different DNA profiles belonging to the same original, and true, DNA profile. The question that arises is, “How to combine the evidential value of the different profiles?”.

5.2 Procedure

In this section we will discuss the standard procedure for obtaining a DNA profile from a crime stain. Basically we can divide this process in three steps.

1. DNA is obtained from the object (for example the glove). (sampling)
2. The DNA is amplified. (amplification)
3. The DNA profile of the amplified DNA is made. (interpretation)

We will discuss the different steps in the following subsections.
5.2.1 Obtaining DNA from an object

We cannot put the entire glove in a device to obtain a DNA profile. Therefore, we need an extract which hopefully contains DNA. Once extracted, samples are quantified to see how much DNA has been extracted from a swab/sample. This is particularly important to highlight samples that require extra cycles of PCR (polymerase chain reaction).

5.2.2 Amplifying the DNA using PCR

The amount of obtained DNA is usually too small to make a DNA profile. Typically, 500 ng (nanogram $10^{-9}$ gram) is required for a successful test, but when we amplify the DNA, less than 1 ng could be analyzed. This amplifying is done by using a polymerase chain reaction (PCR). The PCR reaction is like a molecular photocopier. PCR involves a number of replication ‘cycles’. Each cycle has the potential to double the amount of DNA, although actual amplification is slightly less than a doubling. The DNA strings are detached from each other and put in to a mixture of enzymes. These enzymes attach to the DNA strings, leading to two couples of DNA strings. In many cases, standard casework is undertaken at 28 cycles. At perfect amplification this should amplify the starting template by a factor of $2^{28} = 268,435,456$. However, perfect amplification is never achieved.

With each amplification there is a probability of changing the actual DNA profile. This is best compared with a zipper. Suppose you unzip a zipper after which we have two loose parts. If we attach two other parts to the unzipped zipper parts it is possible that there is loop in one of the two new zippers. The same can happen when we attach enzymes to our DNA strings. In other words, amplification can have a very significant effect on the profile morphology.

When the amount of DNA before amplification is very small (for example 15 to 20 cells), we speak of Low Copy Number (LCN) DNA analysis. In this case the number of amplification cycles is usually increased from 28 to 34 (which would lead to a theoretical factor of $2^{34} = 17\,179\,869\,184$). As said, increasing the number of amplification cycles will usually also have a significant effect on the profile morphology for very few amplified templates. One can imagine that when the first cycle 5 of the 15 cells give a different DNA, the final batch of DNA will include a lot of DNA that is different from the original.

Possible errors due to amplification of the DNA

We will discuss the possible errors that come from amplifying the DNA.

Stutters  Stutters occur due to error in the amplification of the DNA. A stutter is a smaller peak just before the peak from the allele that is in the DNA. These stutters usually have a peak height between 6% and 15% of the allelic peak. In figure 5.2.1 is an explanation on what happens in the amplification for a stutter to occur.

Allelic drop-out  Allele drop-out occurs when a sample is typed and one or more alleles are not present. This can be due to an initial input quantity of DNA that is too low, resulting in the failure to amplify one or more alleles in the sample. Another cause for allelic drop-out is that a crime stain is ‘degraded’. It has been laying somewhere for a long time or under bad circumstances (e.g. moist) and some of the DNA is degraded. This affects mostly the longer repeat sequences.

Allelic drop-in  A contamination may result in what is called an allelic drop-in. In our profile we see an allele that is not in the actual DNA obtained from the stain.

5.2.3 Producing the DNA profile

After the amplification we should have enough DNA material to make a DNA profile. A DNA profile consists of a continuous line containing peaks at certain points. The profile can be subdivided into different loci. A ‘perfect’ DNA profile has peaks at each locus. An example of a DNA profile can be found in figure 5.2.2
Figure 5.2.1: Diagrammed in this figure is a proposed mechanism to explain the formation of stutters one repeat unit shorter than the main allele band. The template strand (main allele) in this figure contains seven repeat units, as represented by the dark gray shaded boxes. In Step 1, the DNA polymerase has extended through four repeat units, represented by the light gray boxes. When the polymerase falls of the extending strand, the template and extending strand can breath apart, as shown in Step 2. When the two strands re-anneal in Step 3 the template strand has looped out and the extending strand aligns out of register by one repeat unit. Thus the polymerase can only add an additional two repeat units, instead of the correct three repeat units. Step 4 demonstrates that at the completion of synthesis the newly extended strand contain only six repeat units, while the template strand has seven repeat units.

On this profile we have information on eleven different loci, which are marked in purple above the profile. At all these loci we see one or two peaks. If we want to summarize the results from this profile in a table we get,

<table>
<thead>
<tr>
<th></th>
<th>D3S1358</th>
<th>vWA</th>
<th>D16S539</th>
<th>D2S1338</th>
<th>. .</th>
<th>D8S1179</th>
<th>D21S11</th>
<th>D18S51</th>
<th>D19S433</th>
<th>THO</th>
<th>FGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3S1358</td>
<td>14</td>
<td>14</td>
<td>11</td>
<td>17</td>
<td>X</td>
<td>11</td>
<td>28</td>
<td>14</td>
<td>13</td>
<td>6</td>
<td>22</td>
</tr>
<tr>
<td>D16S539</td>
<td>16</td>
<td>16</td>
<td>11</td>
<td>17</td>
<td>X</td>
<td>11</td>
<td>28</td>
<td>14</td>
<td>13</td>
<td>6</td>
<td>22</td>
</tr>
</tbody>
</table>

The dots stand for the pair of sex-chromosomes. In this case we find an X and a Y, which means that the profile belongs to a man. If this profile would have been obtained from a crime stain we could compare it with the DNA profile of a suspect or try to find a match in a DNA database.

Unfortunately, the profiles obtained from crime stains usually are not as nice as in figure 5.2.2. In that case, multiple DNA profiles are made from the same stain and compared with each other. Especially when the original amount of DNA is very small, this is done multiple times (3, 4 or even more). Since copy mistakes can lead to stutters, drop-outs or drop-ins; there might be some differences between the
Figure 5.2.2: An example of a DNA profile

different profiles, although they come from the same original DNA. The question here is; what can we say about the profile belonging to the original DNA given these profiles?

There are different 'kits' that can be used to analyze DNA. Each kit comes with a certain level. When the peak height is bigger than this level the allele is assumed to be in the profile. At each locus you can be either a hetero- or a homozygote.

**Heterozygotes and Homozygotes** A homozygote has two identical copies of a certain gene at a locus. In figure 5.2.2 we see this at the locus D16S593. The peak height at 11 is approximately twice as high as the peak heights in vWA, which is a heterozygote since it has two different alleles (14 and 16).

**Different lengths for the repetitive parts** The repetitive parts that determine the allele on a locus (see section 4.3) can have different lengths. Most common is to use a repetitive part of length 4 but on some loci the repetitive parts have lengths of 1, 2 or 3 units. These alleles are represented with 10.1, 10.2 or 10.3 (when the number of times the repetitive part is repeated is 10). The alleles from repetitive parts of length 4 are only represented by the number of times the part is repeated, i.e. 10. An allele with a repetitive part of length 2 can only stutter to an allele with a repetitive part of length 2 (the same holds for lengths 1, 3 and 4).

### 5.3 Analysis of existing methods

For analyzing (low template) DNA profiles a lot of methods have been developed. In this section we will discuss very briefly some of these methods. We will divide the methods into two groups; one group with
methods that use a threshold for the peak heights and one group that uses a drop-out probability.

### 5.3.1 Methods with a threshold

Two widely used methods for evaluating multiple DNA profiles from the same trace are the composite- and the consensus method (Benschop, (2012)[2]).

**Composite method**

The composite method works with a threshold for the peak heights. For each profile, the alleles that have a peak height that is higher than this threshold are collected. For each locus, all the alleles that were observed at least one time in any of the profiles is put into the final profile. This profile can be compared with a reference profile from a suspect.

**Consensus method**

The consensus method also works also with a threshold for the peak heights. Again, for each profile the alleles that have a peak height higher than this threshold are collected. The final profile consists of all the alleles that occurred at least half of the times in the individual profile. This profile can be compared with a reference profile from a suspect. There are several variations on this method.

An important downside of both these methods is that they use a threshold to determine whether a peak should be marked as an allele or not. Suppose that the threshold is set to 50 rfu. A peak height that is just below this threshold (i.e. 49 rfu) is discarded from the profile, where a peak that is just above threshold (i.e. 51 rfu) is considered to be in the profile. In other words, a lot of information is discarded using these methods. A peak height very far above the threshold seems to be containing more information than a peak that is just above the threshold, but they are treated equally.

**Roberto Puch Solis et al. (Submitted in 2011)**

In *Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters* [3] by Roberto Puch Solis et al. a statistical model is introduced to calculate likelihood ratios for evaluating DNA evidence arising from multiple known and unknown contributors that allows for dropouts and stutters by incorporating peak heights. The model incorporates peak heights to interpret DNA profiles from any number of known and unknown sources, some of whom may have contributed low levels and/or degraded DNA. The sum of observed peak heights at a locus is used as a proxy for DNA quantity, which allows coping with differential effects over loci of degradation and inhibition.

In the model, a detection threshold is used. All peaks lower than 30 rfu, are ignored. The model does not incorporate noise. Each peak height higher than the detection threshold is assumed to be either a stutter of another allele, or is allelic, meaning that it originates from a donor with this allele at that locus.

**Mark W. Perlin et al. (2001)**

In *Linear Mixture Analysis: A Mathematical Approach to Resolving Mixed DNA Samples* [5] Mark W. Perlin and Beate Szabady developed a quantita

\[ d = G \cdot w \]

Here, column \(d\) describes the mixture profile peak quantitation data, matrix \(G\) represents the genotypes (column \(j\) gives the alleles for individual \(j\)) and \(w\) is the weight column vector that reflects the relative proportions of template DNA or PCR product. Since, in general, the weights of the different contributors are unknown, a workable approach is searching for the best weight \(w\) in the \([0,1]\) interval that satisfies additional constraints on the problem. Before the method (LMA) can be applied, it is important that DNA analysts already have made proper allele calls to exclude artifact peaks, such as stutter and pull-up peaks. We put this method in the section with threshold methods because DNA analysts need to make these proper calls.
Wang et al. (2006)

In *Least-Square Deconvolution: A Framework for Interpreting Short Tandem Repeat Mixtures* [4] by Tsewei Wang, Ning Xue and J. Douglas Birdwell a least-square-based interpretation framework is presented for interpreting two-contributor STR mixture samples using the quantitative allele peak data information. The model gives a framework on how to obtain a ranking of the most likely DNA profile pairs that contribute to the mixture profiles. Before the method (LSD) can be applied, it is important that DNA analysts already have made proper allele calls to exclude artifact peaks, such as stutter and pull-up peaks. We put this method in the section with threshold methods because DNA analysts need to make these proper calls.

The model acts on allele peak data fed to it for each locus and returns the best-fit mass proportions for each possible genotype combination. The method can be summarized by the following steps:

1. Take the peak heights for the alleles from one locus (only the ones that were selected by DNA analysts) and normalize them.
2. Take each possible genotype combination for this locus, and compute the best-fit mass proportion, i.e. take each possible genotype combination on the locus and use least-squares-estimation to find the proportion between these genotypes that comes the closest to the actual peak heights.
3. Calculate the fitting error for each possible combination, this is the difference between the actual peak heights, and the peak heights given the genotype combination and the mass proportion.
4. Calculate the error residual (the sum of the squares of the entries of the vector from step 3) for each possible combination.
5. Rank the possible genotype combinations according to their residuals, from the smallest to the largest.
6. Calculate the ratio of the error residual of each possible combination case to the residual of the top-ranked combination case.
7. Calculate the mass proportion ratio of each genotype combination case by dividing the larger mass proportion element by the smaller mass proportion element of the mass proportion vector of step 2.
8. Take the next locus and apply steps 1-7. Continue until all loci have been processed.

After this procedure the results should be given to a DNA analyst. He gets the LSD rankings of the possible genotype combinations. This way one can find the allele combinations on all loci that yield the highest likelihood of the DNA profile, by taking into account that the mass proportions on all the loci should be similar.

David Balding et al. 2009

In *Interpreting low template DNA profiles* [8] David J. Balding and John Buckleton develop a statistical model for evaluating DNA profile evidence, including complex mixtures and when profiles are subject to drop-out. The model is able to estimate the relative DNA template levels originating from different contributors. The model has two important limitations/downsides. They do not use peak height information, only the presence or absence of peaks. They assume a dropout rate for a given allele that is the same over all loci.

Hinda Haned and Peter Gill (2011)

In *Analysis of complex DNA mixtures using the Forensim package* [9] Hinda Haned and Peter Gill present a module within Forensim, LRmix. It enables the calculation of likelihood ratios for complex STR profiles with drop-in and drop-out, multiple contributors and multiple replicates. For this method a software package was written for R. A downside of the method is that it does not use the peak heights and that it needs an expert opinion on the drop-out probability.
5.3.2 Methods using a drop-out probability

Some methods assume a drop-out probability to be able to interpret a DNA profile. We will give a short overview of some models that assume such a drop-out probability.

**Torben Tvedebrink et al. (2012)**

In *Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out* Torben Tvedebrink, Poul Svante Eriksen, Helle Smidt Mogensen and Niels Morling [7] present a method for measuring the degree of degradation of a sample and demonstrate how to incorporate this in estimating the probability of allelic drop-out. They argue that when a DNA profile is made from a degraded sample, there is a significant probability that an allele has ‘dropped-out’. This makes it harder to deduce the evidential value of the profile when comparing the profile with the DNA profile of a suspect. They assume that longer sequences, i.e. higher allele-numbers, have a higher probability of degradation, and take as degradation probability,

\[ p^{bp} \]

where \( p \) is some number, \( 0 \leq p \leq 1 \), and \( bp \) is the number of base pairs, i.e. the allele number. They model the amount of DNA as the average peak height, and use this number to scale the degradation of the DNA.

When comparing a DNA profile obtained from a crime stain with a DNA profile from a suspect, and in the DNA profile from the crime stain there is a peak missing on an allele that is in the DNA profile of the suspect, we need to correct the likelihood with the probability that that particular allele has dropped-out. Also, when considering the possibility that the DNA profile originated from somebody else, we need to take the drop-out probabilities into account.

**R.G. Cowell et al. (2011)**

In *Probabilistic expert systems for handling artifacts in complex DNA mixtures* by R.G. Cowell et al. [6] present a coherent probabilistic framework for taking account of allelic drop-out, stutter bands and silent alleles when interpreting STR DNA profiles from a mixture sample using peak size information. In an older paper, Cowell et al. presented a Bayesian network for modeling peak area values in the absence of such artifacts. In this paper, they show how to extend this model to handle silent alleles, drop-out and stutter bands. The gamma model of Cowell et al. considers \( I \) potential contributors to a DNA mixture. Each contributor added some proportion of the DNA, denoted with \( \theta = (\theta_1, \theta_2, \ldots, \theta_I) \), where \( \sum \theta_i = 1 \). On each locus, they consider the peak-weight, which is the product of the peak height at an allele times the allele number. They compute the likelihood using an assumption that states that each contribution to the peak weight at an allele from an individual is gamma distributed where the shape parameter is an indication of the amount that the individual contributed. They derive (using properties of the gamma distribution) an expression for the relative weights, which is the peak weight at an allele divided by the total peak weights over all the alleles on a locus. These relative weights follow a Dirichlet distribution and the Dirichlet distribution determines the likelihood of the observed profile. Their model assumes a drop-out probability to account for possibly missing allelic peaks.

5.3.3 Conclusion

We briefly discussed some methods that were developed to interpret DNA profiles. A downside to some of these methods is that they use a threshold on the peak heights to determine which alleles are present and which alleles are not present in the profile. The other methods that we discussed here assume some drop-out probability in order to explain “missing peaks” in the DNA profiles. This drop-out probability is also based on a threshold. An allele is said to be dropped-out when that allele fails to produce a peak height that is higher than some threshold.

Our model does not use a threshold. The peak heights on all loci are used to determine the likelihoods of observing the DNA profile given all possible pairs of alleles on all loci. Our model also makes some assumptions, but most of them can be easily changed without changing our method. Especially the assumptions made on the parameters and distributions of the different stochastic variables can be easily substituted by more realistic ones.
5.4 Method to compute the likelihood of a combination of alleles to be contributed to a stain, given the DNA profile

In this section we will develop a method which can determine the likelihoods of different combinations of alleles to be the contributor to a non-mixture DNA profile. To do so, we will assume the following:

- We have a DNA profile that comes from a stain that was left by one person, i.e. it is not a mixture.
- The DNA profile is a result of a stochastic process where the only input are the alleles of the donor of the stain.
- Stutters can occur in the amplification of the DNA that was obtained from the stain stutters can occur, but only stutters to one allelic position less than the alleles that the donor bears.
- Knowing the amount of input DNA is information for the peak heights one will see in the DNA profile.
- When the DNA profile is made from the amplified DNA, some normal distributed noise is added to the signal.
- Although the stain was left by one person, it is possible that a contamination or artifact occurred which led to additional peaks in the DNA profile.
- The stutter amounts on the alleles are independent of each other and independent of the noise and amplification of the process.
- Stutters from alleles can only stutter to an allele where the length of the repetitive part is the same as the allele where the stutter originated from.

In the remaining of this section, we will use the following notation.

\[ x := \text{the DNA profile that we obtained from the crime stain, the original profile, we will see it as m vector(s) with the peak heights on locus m at the corresponding allele numbers.} \]

\[ P := \text{the input profile, i.e. the combination of alleles of which we want to know what the likelihood is when this combination of alleles was the contributor of the original profile. It will be m vector(s) with at each position the recorded peak height at allele i, locus m.} \]

\[ \pi_i := \text{a measure for the proportion of stutters (the amount of stutters on each peak). } \pi_i \text{ is a stochastic variable. It tells us how much of the original DNA material stays in } i, \text{ and how much goes to } i - 1. \text{ For instance, } \pi_i = (0.1) \text{ says that 10% of the peak height is 'stuttered' to the allele } i - 1, \text{ and 90% stays at the original allele } i. \text{ The value of } \pi_i \text{ depends on } m, \text{ the total amount of DNA. If this value is high we expect lower stutters, and when this value is low we expect higher stutters. For now, we will work with } \pi_i \text{ being lognormally distributed over all loci. The mean and standard deviation of the distribution of } \pi_i \text{ is different for all alleles and on all loci. We will tell more about this in section 5.6.1.} \]

\[ c_i := \text{the contamination-amount on each allele, it tells us how much irrelevant DNA was added to our relevant DNA (for instance, DNA from an imprudent researcher), it is a measure of which we choose the cumulative distribution function to be } P(c \leq a) = 1 - \frac{1}{\sqrt{1 + 99.9a}}. \text{ This was chosen in such a way that 90\% would be less than 1 and that the distribution was 'heavy tailed' to the right. In figure 5.6.3 we show the probability density function of } c. \text{ More on why we choose this distribution can be found in section 5.6.3.} \]

\[ \Lambda_i := \Lambda \text{ is a factor that accounts for differences in the actual peak height compared with the peak height one would expect, given a certain amount of input. We assume it is normally distributed with mean 1 and standard deviation 0.2. More on why we choose this values can be found in section 5.6.3.} \]

\[ \epsilon_i := \epsilon \text{ is the normally distributed noise, } \epsilon_i \text{ is the noise on allele } i. \text{ We assume it has mean 0 and standard deviation 10, see also section 5.6.3.} \]
Our model assumes a relation between the DNA profile $x$ and the variables $P$, $\pi$, $c$, $\Lambda$ and $\epsilon$. We assume that the peak height at an allele is a combination of five things. First, a contribution from $P$, the donor of the profile, that bears that allele. Second, a possible contamination on that allele. Third, a stutter from a possible contribution $P$ on a higher allele combined with a possible contamination on that higher allele. The last thing that is added to the peak height is the noise on that allele $\epsilon$. We add amplification factors to get the system of equations given in (5.4.1).

$$
x_1 = \Lambda_1(P_1 + c_1) \cdot (1 - \pi_1) + \Lambda_2(P_2 + c_2) \pi_2 + \epsilon_1
$$
$$
x_2 = \Lambda_2(P_2 + c_2) \cdot (1 - \pi_2) + \Lambda_3(P_3 + c_3) \pi_3 + \epsilon_2
$$
$$
\vdots
$$
$$
x_i = \Lambda_i(P_i + c_i) \cdot (1 - \pi_i) + \Lambda_{i+1}(P_{i+1} + c_{i+1}) \pi_{i+1} + \epsilon_i
$$
$$
\vdots
$$
$$
x_n = \Lambda_n(P_n + c_n) \cdot (1 - \pi_n) + \epsilon_n
$$

This system of equations is for the peak heights on one locus with $n$ alleles. The system is similar for all other loci. We can describe $x$ as a function $\Psi$ of $P, \pi, c, \Lambda$ and $\epsilon$.

$$
x = \Psi(P, \pi, c, \Lambda, \epsilon)
$$

We can write this system of equations as a matrix-vector multiplication.

$$
x = A(\pi, \Lambda)c + b(P, \pi, \Lambda) + \epsilon
$$

where,

$$
A = \begin{pmatrix}
\Lambda_1 \cdot (1 - \pi_1) & \Lambda_2 \cdot \pi_2 & 0 & \cdots & \cdots & 0 \\
0 & \Lambda_2 \cdot (1 - \pi_2) & \Lambda_3 \cdot \pi_3 & 0 & \cdots & \cdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\vdots & \ddots & \ddots & \ddots & 0 & \Lambda_i \cdot (1 - \pi_i) & \Lambda_{i+1} \cdot \pi_{i+1} & 0 \\
\vdots & \ddots & \ddots & \ddots & \cdots & \ddots & \ddots & \cdots & 0 & \Lambda_n \cdot (1 - \pi_n)
\end{pmatrix}
$$

and,

$$
b = \begin{pmatrix}
\Lambda_1 P_1 \cdot (1 - \pi_1) + \Lambda_2 P_2 \pi_2 \\
\Lambda_2 P_2 \cdot (1 - \pi_2) + \Lambda_3 P_3 \pi_3 \\
\vdots \\
\Lambda_i P_i \cdot (1 - \pi_i) + \Lambda_{i+1} P_{i+1} \pi_{i+1} \\
\vdots \\
\Lambda_n P_n \cdot (1 - \pi_n)
\end{pmatrix}
$$

As said, we are interested in the likelihood of observing the DNA profile $x$ given that the contributor had alleles $P$, $f(x|P)$, but since we do not know the values of $\pi, \Lambda, c$ and $\epsilon$ that led to this profile, we will estimate this likelihood by simulating values for these parameters (although we ‘know’ the probability distribution of $\pi, \Lambda, c$ and $\epsilon$, it is still very hard to determine the likelihood analytically, therefore we will use a simulation to estimate it). In this method we will simulate values for $\pi, \Lambda$ and $c$ and determine the corresponding value of $\epsilon$ that will lead to the profile $x$. The likelihood of $x$ will be determined by the following procedure.

### 5.4.1 Theoretical Procedure

The likelihood of interest is

$$f(x|P).$$

We will estimate this likelihood by simulation. First, we summarize our method in a number of steps:
1. For each combination of alleles we simulate values for $\pi$ and $\Lambda$.

2. Using these simulated values, we will determine the maximum likelihood estimator, $\tilde{c}$, for our contamination amounts $c$, given $x$ and $P$ and the distribution of $\epsilon$.

3. We will simulate a vector $c$ for the contamination from a distribution that is centered at the maximum likelihood estimator, $\tilde{c}$.

4. Using the simulated values of $\pi$, $\Lambda$ and $c$, we will determine the vector $\epsilon$ that will result in the desired vector $x$.

5. The values for $\epsilon$ and the simulated values for $c$ determine the likelihood for the combination of alleles.

We will explain how we determine the likelihood of a DNA profile $x$ using the previous steps.

**Simulating $\pi$ and $\Lambda$**

By definition, (5.4.5) is the same as,

$$f(x|P) = \int f(x|P, \Lambda, \pi)d\pi d\Lambda \quad (5.4.6)$$

If we simulate $M$ values for $\pi$ and $\Lambda$, we know that (5.4.6) can be approximated with

$$\int f(x|P, \Lambda, \pi)d\pi d\Lambda \approx \frac{1}{M} \sum_{i=1}^{M} f(x|P, \Lambda^{(i)}, \pi^{(i)}). \quad (5.4.7)$$

Where $\pi$ and $\Lambda$ are assumed to be independent of each other.

**Determine the MLE $\tilde{c}$**

It is not a good idea to simulate values $c$ for our contamination. Since it is a distribution with a ‘heavy tail’ (one would expect to have no contamination but there is a non negligible probability that the contamination is high), it will take a very large amount of simulations to cover the whole distribution (remember that $c$ is a vector and that we would need to simulate the contamination amounts on each of the alleles in this vector). Even more important is that a big part of the possible outcomes of $c$ is irrelevant. We are interested in the values of $c$ that will result in a high likelihood. For instance, if we have a DNA profile on a locus with one homozygote peak at allele 10, we are not interested in the case where there was a very high contamination on allele 14, since we would only be able to explain our profile with a very negative noise amount $\epsilon$. In other words, when you have a DNA profile and want to know the likelihood of finding it given some combination of alleles, one already has a clue of what the contamination needs to be on that locus. Therefore, we will determine the maximum likelihood estimator of $c$, $\tilde{c}$ and simulate values of $c$ around this maximum likelihood estimator $\tilde{c}$. This way, the variation in the simulated likelihoods will be much smaller. To determine the maximum likelihood estimator, we first need to do some derivations. We will continue to work with the term $f(x|P, \Lambda, \pi)$ from (5.4.7). We know that it is equal to

$$f(x|P, \Lambda, \pi) = \int f(x|P, \Lambda, \pi, c)f_c(c)dc. \quad (5.4.8)$$

Where $f_c$ is the distribution of the contamination. The term $f(x|P, \Lambda, \pi, c)$ represents the likelihood of $x$, given the values for $P, \Lambda, \pi, c$. We assumed the following relation between $x$ and $P, \Lambda, \pi, c$ and $\epsilon$,

$$x_i = \Lambda_i \cdot (P_i + c_i) \cdot (1 - \pi_i) + \Lambda_{i+1} \cdot (P_{i+1} + C_{i+1}) \pi_{i+1} + \epsilon_i.$$ 

Since $x$ is known, the only unknown in our system, given $P, \Lambda, \pi$ and $c$, is the vector $\epsilon$. Therefore, we can find $A$ and $b$ such that

$$x = Ac + b + \epsilon$$
The likelihood of \( \epsilon \) happens with the likelihood of our profile when the contamination is \( \tilde{h} \). Hence, we can choose some probability density function for \( g \). From (5.4.8) we have

\[
\prod_{i=1}^{n} f(\epsilon_i; \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{\epsilon_i}{\sigma} \right)^2 \right) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 \right) = \gamma \cdot \exp \left( -\frac{\|\epsilon\|^2}{2\sigma^2} \right). \tag{5.4.9} \]

Therefore, we see that when \( \|\epsilon\|^2 \) (the square of the euclidean norm of \( \epsilon \)) is at its minimum, then the likelihood of \( \epsilon \) is maximal. So, the maximum likelihood estimator \( \tilde{c} \) for \( c \) is given by,

\[
\tilde{c} = \min_{c \geq 0} \|c\|_2^2 = \min_{c \geq 0} \|x - b - Ac\|_2^2. \tag{5.4.10} \]

**Simulate the vector \( c \)**

From (5.4.8) we have

\[
f(x|P, \Lambda, \pi) = \int_c f(x|P, \Lambda, \pi, c) f_c(c) \frac{g_c(c)}{g_c(c)} dc. \tag{5.4.11} \]

Hence, we can choose some probability density function for \( g \) that tells us something about how \( c \) is distributed around our maximum likelihood estimator \( \tilde{c} \). This makes (5.4.11) into an expectation given the probability density function \( g \),

\[
f(x|P, \Lambda, \pi) = \mathbb{E}_g \left[ f(x|P, \Lambda, \pi, c) \frac{f_c(c)}{g_c(c)} \right]. \tag{5.4.12} \]

We know that \( \tilde{c} \geq 0 \). We assume independence between the different values of \( c \) under \( g \), so we assume the different components of \( c \) to be independent of each other under our distribution \( g \) (around \( \tilde{c} \)). We want to simulate these values of \( c \) that are around the maximum likelihood estimator \( \tilde{c} \). Therefore we assume that the contamination on each allele \( i \) is normally distributed around \( \tilde{c}_i \) with standard deviation \( \tilde{\omega} \). Now we are interested in what a good choice is for the standard deviation \( \tilde{\sigma} \). Therefore, we check what happens with our likelihood when we add a little bit to our maximum likelihood estimator \( \tilde{c} \). So, what happens with the likelihood of our profile when the contamination is \( \tilde{c} + \delta \) instead of \( \tilde{c} \). A contamination of \( \tilde{c} + \delta \) will give an \( \epsilon \) of

\[
\epsilon = x - b - A(\tilde{c} + \delta) - A\epsilon - A\delta. \]

The likelihood of \( \epsilon \) becomes

\[
\gamma \exp \left( -\frac{1}{2\sigma^2} \|x - b - A(\tilde{c} + \delta)\|_2^2 \right) \]

The difference we encounter is in the exponent. We have,

\[
\|x - b - A(\tilde{c} + \delta)\|_2^2 = \|x - b - A\tilde{c} - A\delta\|_2^2 = \|x - b - A\tilde{c}\|_2^2 + 2|A\delta, x - b - A\tilde{c}| + \|A\delta\|_2^2. \tag{5.4.13} \]

We see that by adding \( \delta \) to our contamination, our likelihood of \( \epsilon \) gets multiplied by an extra term of

\[
\exp \left( -\frac{1}{2\sigma^2} \{2 \cdot |A\delta, x - b - A\tilde{c}| + \|A\delta\|_2^2\} \right) \]

56
If we take $\delta$ as the solution of, (remember that $\sigma$ is the standard deviation of the noise $\epsilon$),

$$2\delta \langle A\epsilon_j, x - b - A\tilde{c} \rangle + \delta^2 \| A\epsilon_j \|^2 = 2\sigma^2 \sqrt{\log(100)},$$

we get that are likelihood of $\epsilon$ gets multiplied by the extra term

$$\exp \left( -\frac{1}{2\sigma^2} \{2\sigma^2 \sqrt{\log(100)}\} \right).$$

So, we get that the likelihood of $\epsilon$ becomes $\frac{1}{100}$ smaller per allele when we add $\delta$. If we take $\tilde{\sigma}$ to be $\delta/3$ on each allele (so $\tilde{\sigma}$ is a vector with the standard deviations for our distribution $g$ around the maximum likelihood estimator of $c$, $\tilde{c}$), we suspect the majority of our simulations to be in the range,

$$(\tilde{c} - \delta, \tilde{c} + \delta)$$

This means that we will choose $c$ to be (remember that $\tilde{c}$ is the maximum likelihood estimator of $c$),

$$c \sim \mathcal{N}(\tilde{c}, \tilde{\sigma})$$

Now we encounter a small problem. We know that our maximum likelihood estimate $\tilde{c} \geq 0$, but it might be very close or even equal to 0 on some alleles. We also know that $c$ can only be greater or equal to zero, but since it will be simulated from a normal distribution around $\tilde{c}$, it can occur that it becomes less than 0 when $\tilde{c}$ is close to 0. Therefore, we will take $c$ as the absolute value,

$$|\tilde{c} + \eta|$$

where

$$\eta \sim \mathcal{N}(0, \tilde{\sigma})$$

Therefore, the likelihood of $c$ on each allele is a sum of two probabilities: Suppose that $c_1 = 40$, there are two possibilities that could give us this result. Either $\tilde{c} + \eta = -40$ or $\tilde{c} + \eta = 40$. Therefore,

$$g_c(c_1) = h(-40; \tilde{c}_1, \sigma_1) + h(40; \tilde{c}_1, \sigma_1),$$

where $g_c$ is the distribution of $c$ around our maximum likelihood estimator $\tilde{c}$ and where $h$ is the probability density function of a normal distribution with mean $\tilde{c}_1$ and standard deviation $\sigma_1$. When $\tilde{c}_1$ is close to 0, this will result in similar terms that will be added, in other cases (where $\tilde{c}_1 \gg 0$), one term will be approximately zero.

**Determine $\epsilon$**

Up till now, we simulated values for $\pi, \Lambda$ and $c$, we know $x$ and we assumed some allele combination $P_1, P_2$ to be the allele combination of the donor of the stain of which the DNA profile was obtained. With all these values we can easily obtain the vector representing the noise $\epsilon$,

$$\epsilon = x - b - Ac$$

We need to be careful here. We are interested in $f(x|P)$. We know that there is a function $\Psi$, such that,

$$x = \Psi(P, \pi, c, \Lambda, \epsilon)$$

When we want to know the likelihood $f(x|\pi, \Lambda, c, P)$, we can use the fact that we know the function $\Psi$:

$$P(x \in B|\pi, \Lambda, P, c) = \mathbb{P}(\epsilon \in \Pi_{-1,\Lambda,P,c}(B)|\pi, \Lambda, P, c)$$

$$= \int_{\Pi_{-1,\Lambda,P,c}(B)} f_\epsilon(y)dy$$

$$= \int_B f_\epsilon(\Psi^{-1}_{\pi,\Lambda,P,c}(x)) \cdot |D\Psi^{(-1)}_{\pi,\Lambda,P,c}(x)|dx$$

Note that $|D\Psi^{(-1)}_{\pi,\Lambda,P,c}(x)|$ is 1, so,

$$f_\epsilon(x|\pi, \Lambda, P, c) = f_\epsilon(x - b - Ac) \cdot 1 \quad (5.4.14)$$

So we need to multiply our likelihood with the Jacobian $|D\Psi^{(-1)}_{\pi,\Lambda,P,c}(x)|$, but since this is 1 we will not notice it.
Determine the likelihood

For each simulation, we get a vector $c$ and a vector $c$ that together determine the likelihood. We derived that the likelihood that we are interested in is (see (5.4.5)-(5.4.12))

$$f(x|P) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{\pi} \left[ f(x|P, \Lambda^{(i)}, \pi^{(i)}, c) \frac{f_{c}(c)}{g_{c}(c)} \right]$$

(5.4.15)

In this equation, $f(x|P, \Lambda^{(i)}, \pi^{(i)}, c)$ is as in (5.4.14), note that $f_{c}$ is the product of the individual $f_{c}(c_{i})$. The function $f_{c}$ represents the probability density function of the contamination $c$. We set it to be

$$f_{c}(a) = \frac{99}{2 \cdot (1 + 99a)^{3/2}},$$

a distribution which is heavy tailed, and we choose it to be in such a way that 90% of the values is below 1. The function $g_{c}$ represents the distribution we choose of $c$ around the maximum likelihood estimator for $c$, $\hat{c}$. We showed that,

$$g_{c}(a) = h(-a; \hat{c}, \hat{\sigma}) + h(a; \hat{c}, \hat{\sigma})$$

where $h$ is a normal distribution with mean $\hat{\sigma}$ and standard deviation $\hat{\sigma}$. The product of the likelihood of $x$ given $P, \Lambda, \pi$ and $c$ and the fraction $\frac{f_{c}(c)}{g_{c}(c)}$ will result in the estimated likelihood for observing the profile $x$ given the allele combination $P$. Since we will compute the likelihood $M$ times for each combination of alleles, we will take the mean value of these simulated likelihoods as our approximation for the likelihood for observing the profile $x$ given that the donor of stain of which the profile was made had the allele combination $P$, which is a vector with on the two allele positions the expected peak height.

5.4.2 The method in practice

We developed a method to find the likelihood for observing a DNA profile given the allele combination of its donor and explained the theory behind it. In practice, we can/will alternate some elements to make our method faster. In equation (5.4.10) we set the vector $\tilde{c}$ to be the solution of a minimization problem. This is an expensive computation, in the sense that it takes quite some time. If we would determine this value for each new simulation, it will make our simulation very slow (remember that we do $M$ simulations on each locus for each possible pair of alleles). Therefore, we suggest to slightly change our method in order to be faster. We will examine this decision in section 5.8.

Instead of determining the vector $\tilde{c}$ each time by solving a minimization problem, we will determine the vector $\tilde{c}$ just once for each possible pair of alleles on each locus. We will do this by taking the vectors $\Lambda$ and $\pi$ equal to their expected values and making the matrix $\Lambda$ and $\pi$ using these values. The vector $\Lambda$ will be a vector only containing ones and the vector $\pi$ will be a vector containing the mean stutter percentages on each allele (based on data). With this we can obtain the maximum likelihood estimator $\tilde{c}$ for these values of $\pi$ and $\Lambda$. We will use this vector $\tilde{c}$ as a direction for approximating the contamination with the maximum likelihood during the simulations for $\Lambda$ and $\pi$.

In each of the $M$ simulations, we will determine $\Lambda$ and $\pi$ by the simulated values for $\Lambda$ and $\pi$ of that simulation $j$. We assume that the maximum likelihood estimator $\tilde{c}^{(j)}$ can be approximated by taking the maximum likelihood estimator over all possible vectors $c$ that have the same direction as $\tilde{c}$ that was minimized using the expected values for $\pi$ and $\Lambda$. In other words, we will take the direction of $\tilde{c}$ to look for a vector for our contamination. We will take the contamination vector that has the maximum likelihood from the set $\chi c$. This way, we only need to solve our minimization problem once instead of $M$ times. As said, we will examine this decision in section 5.8.

Another thing we added to our simulation to make it faster is that we include a lower bound on the sum of the peaks of the alleles of interest to be considered as relevant. In other words, suppose that we set the lower bound to 10, we will only compute the likelihoods for those pairs of alleles where the sum of the peak heights of these alleles in the profile $x$ is higher than 10.

58
5.5 Example to help the reader understand what happens in the simulation

To get a better understanding of the method we use to determine the likelihood for different contributors, we will give an example. Suppose we want to know the most likely contributor for the profile in figure 5.5.1, which is a profile on one locus, TH01. The profile suggests a heterozygote donor with a contamination.

![DNA profile on TH01](image)

The peak heights on the different alleles can be found in table 5.5.1.

<table>
<thead>
<tr>
<th>allele</th>
<th>peak height (rfu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>44</td>
</tr>
<tr>
<td>6</td>
<td>530</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>6.3</td>
<td>50</td>
</tr>
<tr>
<td>7.3</td>
<td>370</td>
</tr>
<tr>
<td>8.3</td>
<td>17</td>
</tr>
<tr>
<td>9.3</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 5.5.1: The peak heights on the different alleles

Since our prior probability for each combination of alleles is given by the allele distribution over our population, we can discard some of our information. These peaks can only be due to noise or contamination. Our prior distribution on the alleles of TH01 (based on a population sample) says that the possible contributors can have 5, 6, 7, 8, 9, 10, 8.3, 9.3. This means that the relevant peak heights for our method are the peaks on 4, 5, 6, 7, 8, 9, 10, 7.3, 8.3, 9.3. (we include 4 and 7.3 since the peak heights at these alleles can be stutters from alleles at 5 and 8.3, due to our prior, it is impossible that a person has an allele combination with 4 and/or 7.3). All other peaks can only come from noise or contamination. Therefore, the profile we will use in our model is slightly different from the profile in table 5.5.1. The profile we will use in the model is given in table 5.5.2.

<table>
<thead>
<tr>
<th>allele</th>
<th>peak height (rfu)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>44</td>
</tr>
<tr>
<td>6</td>
<td>530</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>7.3</td>
<td>50</td>
</tr>
<tr>
<td>8.3</td>
<td>370</td>
</tr>
<tr>
<td>9.3</td>
<td>17</td>
</tr>
</tbody>
</table>

Table 5.5.2: The peak heights on the different alleles without the alleles that are not relevant due to our prior distribution

Now we will compute the likelihood of observing the profile in 5.5.2, given all possible combinations of alleles on TH01.

A profile is made from a certain amount of input DNA. This amount corresponds with a certain peak height. We need to give this estimated peak height to our program. In this case, we set the estimated peak height to 500 rfu (this is an arbitrary but easy to work with value, but since this is just an example that does not matter). In this example, we will show how the likelihood of observing the DNA profile given that the donor was 6-7 is determined. The likelihoods of observing the profile given any other combination of donor alleles is computed likewise.
Suppose that the donor of the profile had the alleles 6-7 on TH01. We want to compute the likelihood for this pair.

\[
\text{Likelihood} = \mathbb{P}(\text{DNA profile|donor is 6-7}) = \mathbb{P}(x_4, x_5, \ldots, x_{9,3}|\text{donor is 6-7}) \tag{5.5.1}
\]

Where \(x_i\) is the peak height at allele \(i\).

In our method, the peak height at an allele is a combination of the amount from the donor, stutters, contamination and noise. To determine the likelihood we have the following set of equations:

\[
\begin{align*}
x_4 & = \Lambda_4 (P_4 + c_4)(1 - \pi_4) + \Lambda_5 (P_5 + c_5)\pi_5 + \epsilon_4 \\
x_5 & = \Lambda_5 (P_5 + c_5)(1 - \pi_5) + \Lambda_6 (P_6 + c_6)\pi_6 + \epsilon_5 \\
x_6 & = \Lambda_6 (P_6 + c_6)(1 - \pi_6) + \Lambda_7 (P_7 + c_7)\pi_7 + \epsilon_6 \\
x_7 & = \Lambda_7 (P_7 + c_7)(1 - \pi_7) + \Lambda_8 (P_8 + c_8)\pi_8 + \epsilon_7 \\
x_8 & = \Lambda_8 (P_8 + c_8)(1 - \pi_8) + \Lambda_9 (P_9 + c_9)\pi_9 + \epsilon_8 \\
x_9 & = \Lambda_9 (P_9 + c_9)(1 - \pi_9) + \Lambda_{10} (P_{10} + c_{10})\pi_{10} + \epsilon_9 \\
x_{10} & = \Lambda_{10} (P_{10} + c_{10})(1 - \pi_{10}) + \epsilon_{10} \\
x_{7,3} & = \Lambda_{7,3} (P_{7,3} + c_{7,3})(1 - \pi_{7,3}) + \Lambda_{8,3} (P_{8,3} + c_{8,3})\pi_{8,3} + \epsilon_{7,3} \\
x_{8,3} & = \Lambda_{8,3} (P_{8,3} + c_{8,3})(1 - \pi_{8,3}) + \Lambda_{9,3} (P_{9,3} + c_{9,3})\pi_{9,3} + \epsilon_{8,3} \\
x_{9,3} & = \Lambda_{9,3} (P_{9,3} + c_{9,3})(1 - \pi_{9,3}) + \epsilon_{9,3}
\end{align*}
\]

In this example, the donor has as DNA profile 6-7, therefore we know all the values of \(P\), we get,

\[
\begin{pmatrix}
x_4 \\
x_5 \\
x_6 \\
x_7 \\
x_8 \\
x_9 \\
x_{10} \\
x_{7,3} \\
x_{8,3} \\
x_{9,3}
\end{pmatrix} =
\begin{pmatrix}
\Lambda_4 (1 - \pi_4) & \Lambda_5 \pi_5 & 0 & \cdots & 0 \\
0 & \cdots & \cdots & \cdots & 0 \\
0 & 0 & \Lambda_6 (1 - \pi_6) & \Lambda_7 \pi_7 & \cdots \\
0 & 0 & 0 & \cdots & \cdots \\
0 & 0 & 0 & 0 & \Lambda_{9,3} (1 - \pi_{9,3})
\end{pmatrix} \begin{pmatrix}
c_4 \\
c_5 \\
c_6 \\
c_7 \\
c_8 \\
c_9 \\
c_{10}
\end{pmatrix} +
\begin{pmatrix}
0 \\
0 \\
0 \\
\Lambda_6 \cdot 500 \cdot \pi_6 + \Lambda_7 \cdot 500 \cdot \pi_7 + \epsilon, \\
\Lambda_8 \cdot 500 \cdot \pi_8 + \Lambda_9 \cdot 500 \cdot \pi_9 + \epsilon
\end{pmatrix}
\]

or,

\[
\begin{pmatrix}
x_4 \\
x_5 \\
x_6 \\
x_7 \\
x_8 \\
x_9 \\
x_{10} \\
x_{7,3} \\
x_{8,3} \\
x_{9,3}
\end{pmatrix} =
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\Lambda_6 \cdot 500 \cdot \pi_6 + \Lambda_7 \cdot 500 \cdot \pi_7 + \epsilon, \\
\Lambda_8 \cdot 500 \cdot \pi_8 + \Lambda_9 \cdot 500 \cdot \pi_9 + \epsilon
\end{pmatrix}
\]

which is the same as,

\[
x = \Lambda c + b + \epsilon \quad \Leftrightarrow \quad \epsilon = x - \Lambda c - b
\]

We want to determine the maximum likelihood estimate of \(c\), in order to draw the random vectors for \(c\) in the region which will result in the most relevant values for the likelihood (i.e. the largest values). We do this by making the matrix \(A\) and the vector \(b\) where we take the expected values for \(A\) and \(\pi\) as
entries (for more information on the expected value of $\pi$, see section 5.6.1).

$$E \begin{pmatrix} \pi_4 \\ \pi_5 \\ \pi_6 \\ \pi_7 \\ \pi_8 \\ \pi_9 \\ \pi_{10} \\ \pi_{7,3} \\ \pi_{8,3} \\ \pi_{9,3} \end{pmatrix} = \begin{pmatrix} 0.0064 \\ 0.0118 \\ 0.0172 \\ 0.0225 \\ 0.0279 \\ 0.0333 \\ 0.0386 \\ 0.0241 \\ 0.0295 \end{pmatrix}, \quad E \begin{pmatrix} \Lambda_4 \\ \Lambda_5 \\ \Lambda_6 \\ \Lambda_7 \\ \Lambda_8 \\ \Lambda_9 \\ \Lambda_{10} \\ \Lambda_{7,3} \\ \Lambda_{8,3} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

The matrix $A$ and the vector $b$ become,

$$A = \begin{pmatrix} 0.9936 & 0.0118 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.9882 & 0.0172 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.9828 & 0.0225 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.9775 & 0.0279 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.9721 & 0.0333 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.9667 & 0.0386 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.9614 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9759 & 0.0295 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9705 & 0.0349 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9651 \end{pmatrix}, \quad b = \begin{pmatrix} 8.5840 \\ 502.6815 \\ 488.7345 \end{pmatrix}$$

Now we can determine $\tilde{c}$ by finding the solution of,

$$\tilde{c} = \min_c \| x - b - Ac \|_2^2$$

We use Matlab, and in particular the command \texttt{lsqnonlin}, to do this. The obtained result is,

$$\tilde{c} = \begin{pmatrix} \tilde{c}_4 \\ \tilde{c}_5 \\ \tilde{c}_6 \\ \tilde{c}_7 \\ \tilde{c}_8 \\ \tilde{c}_9 \\ \tilde{c}_{10} \\ \tilde{c}_{7,3} \\ \tilde{c}_{8,3} \end{pmatrix} = \begin{pmatrix} 2.5936 \\ 35.8391 \\ 0.0000 \\ 41.4843 \\ 14.4110 \\ 0.0000 \\ 3.1155 \\ 39.7298 \\ 380.6151 \\ 17.6141 \end{pmatrix}$$

We will use this as the direction of future vectors $\tilde{c}$. Now we start to simulate values for $\Lambda$ and $\pi$ from their distributions. We obtain for our first simulation,

$$\Lambda = \begin{pmatrix} 1.3232 \\ 1.1490 \\ 1.1591 \\ 1.1951 \\ 1.2074 \\ 0.7842 \\ 1.0114 \\ 0.7193 \\ 1.1297 \end{pmatrix}, \quad \pi = \begin{pmatrix} 0.0000 \\ 0.0000 \\ 0.0096 \\ 0.0096 \\ 0.1142 \\ 0.0028 \\ 0.0200 \\ 0.1255 \end{pmatrix}$$
With these values we can make the matrix $A$ and the vector $b$,

$$
A = \begin{pmatrix}
1.3231 & 0.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1.1490 & 0.0111 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1.1480 & 0.0115 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1.1836 & 0.1379 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1.0696 & 0.0022 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.7820 & 0.0202 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.9912 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.6291 & 0.0283 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.1014 & 0.0020 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9454 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix},
$$

$$
b = \begin{pmatrix}
0.55676 \\
5.797625 \\
591.7816 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{pmatrix}.
$$

Now we want to find our approximation for the maximum likelihood estimator $\tilde{c}$ by checking which vector that has the same direction as the $\tilde{c}$ we found by using the expected values for $\Lambda$ and $\pi$ will give the highest likelihood.

$$
\tilde{c}^{(j)} = \chi \cdot \hat{c}
$$

where $\chi$ is a scalar and the solution of,

$$
\min_{\chi} \| x - b - A(\chi \cdot \hat{c}) \|
$$

It turns out that $\chi = 0.8566$, and, therefore, $\tilde{c}^{(j)}$ is,

$$
\tilde{c}^{(j)} = \begin{pmatrix}
2.2218 \\
30.7012 \\
0.0000 \\
35.5372 \\
12.3450 \\
0.0000 \\
2.6689 \\
34.0342 \\
326.0504 \\
15.0890
\end{pmatrix}
$$

We will simulate our random vectors $c$ around $\tilde{c}^{(j)}$ in such a way that the simulated vectors are ‘close’ to $\tilde{c}^{(j)}$. Therefore, we will need to find the solution of $\delta$, for each unit vector $e_i$,

$$
2\delta |\langle A e_j, x - b - A\hat{c} \rangle| + \delta^2 \| A e_j \|^2 = 2\sigma^2 \cdot \sqrt{\log(100)}
$$

Under the assumption that the noise is normally distributed with mean 0 and standard deviation $\sigma = 10$ and when we are interested in $\hat{\sigma}_4$, the standard deviation around the fourth component (allele 4) of our maximum likelihood estimator at that allele $\hat{c}_4$, we have,

$$
A e_j = \begin{pmatrix}
1.3231 & 0.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
$$

so,

$$
2 |\langle A e_j, x - b - A\hat{c} \rangle| = 0.1591
$$

$$
\Rightarrow \| A e_j \|^2 = 1.3231
$$

$$
2\sigma^2 \cdot \sqrt{\log(100)} = 429.1932
$$

We could compute this value for all $\delta_i$ and find the standard deviation we will use for the variable we
The likelihood of the vector \( c \) under the distribution of where the likelihood of the vector is the product of these values. The likelihood of these individual values add to \( \tilde{c} \). Now we have determined the vectors \( \tilde{c} \) and \( \hat{\sigma} \) we can start computing the likelihood of observing the DNA profile given that the donor had the allele combination 6/7.

First we simulate a value \( c \) where \( c = \tilde{c}^{(j)} + \eta \), where \( \eta \sim \mathcal{N}(0, \hat{\sigma}) \). We get,

\[
c = |\tilde{c}^{(j)} + \eta| = \begin{pmatrix}
2.2218 \\
30.7012 \\
0.0000 \\
35.5372 \\
12.3450 \\
0.0000 \\
2.6689 \\
34.0342 \\
326.0504 \\
15.0890
\end{pmatrix} + \begin{pmatrix}
-5.2962 \\
-1.7908 \\
-0.2866 \\
0.4756 \\
1.3222 \\
-0.3184 \\
5.0115 \\
1.6197 \\
-1.3166 \\
-1.3905
\end{pmatrix} = \begin{pmatrix}
3.0744 \\
28.9104 \\
0.2866 \\
36.0128 \\
13.6672 \\
0.3184 \\
7.6804 \\
35.6539 \\
324.7337 \\
13.6985
\end{pmatrix}
\]

The likelihood of these new individual values under our distribution around the maximum likelihood estimate \( \tilde{c} \), is,

\[
g_c(c) = g_c \begin{pmatrix}
3.0744 \\
28.9104 \\
0.2866 \\
36.0128 \\
13.6672 \\
0.3184 \\
7.6804 \\
35.6539 \\
324.7337 \\
13.6985
\end{pmatrix} = h \begin{pmatrix}
-3.0744 \\
-28.9104 \\
-0.2866 \\
-36.0128 \\
-13.6672 \\
-0.3184 \\
-7.6804 \\
-35.6539 \\
-324.7337 \\
-13.6985
\end{pmatrix} + h = \begin{pmatrix}
3.0744 \\
28.9104 \\
0.2866 \\
36.0128 \\
13.6672 \\
0.3184 \\
7.6804 \\
35.6539 \\
324.7337 \\
13.6985
\end{pmatrix}
\]

The likelihood of the vector is the product of these values. The likelihood of these individual values under the distribution of \( c \), \( f_c(c) \), \( f_c(a) = \frac{99}{2(1+99a)^{99/2}} \) is,

\[
f_c(c) = f_c \begin{pmatrix}
3.0744 \\
28.9104 \\
0.2866 \\
36.0128 \\
13.6672 \\
0.3184 \\
7.6804 \\
35.6539 \\
324.7337 \\
13.6985
\end{pmatrix} = \begin{pmatrix}
0.0093 \\
0.0003 \\
0.3110 \\
0.0002 \\
0.0010 \\
0.2670 \\
0.0024 \\
0.0002 \\
0.0000 \\
0.0010
\end{pmatrix}
\]

The likelihood of the vector \( c \) is the product of these values. With \( A, b \) and \( c \), we can compute what \( \epsilon \) should be, as well as the likelihood for this \( \epsilon \) (we will give the individual likelihoods of the coefficients in
$\epsilon$ in a vector, the likelihood of the vector $\epsilon$ is the product of these values),

$$
\epsilon = x - Ac - b = \begin{pmatrix}
-1.0679 \\
5.2123 \\
-90.5068 \\
-106.2895 \\
-0.6186 \\
-0.4040 \\
-4.6130 \\
18.3733 \\
12.3070 \\
4.0497 \\
\end{pmatrix},
so, \quad f_\epsilon = \begin{pmatrix}
-1.0679 \\
5.2123 \\
-90.5068 \\
-106.2895 \\
-0.6186 \\
-0.4040 \\
-4.6130 \\
18.3733 \\
12.3070 \\
4.0497 \\
\end{pmatrix}.
$$

Now we can compute the likelihood for this simulation,

$$
f(x|\text{donor is 6-7}) = \prod f_\epsilon(\epsilon) \cdot \prod f_c(c) \cdot \prod g_c(c) = 5.5496 \cdot 10^{-77}
$$

We compute this likelihood $M$ times and take the mean as our estimation for the likelihood. This is computed for each possible combination of alleles. The results of this simulation are given in the next subsection (5.5.1).

### 5.5.1 Results

The results after simulation were as follows (we took $M = 100,000$).

<table>
<thead>
<tr>
<th>allele-pair</th>
<th>$\log_{10}$ likelihood</th>
<th>allele-pair</th>
<th>$\log_{10}$ likelihood</th>
<th>allele-pair</th>
<th>$\log_{10}$ likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,8,3</td>
<td>-29.5544</td>
<td>6,9</td>
<td>-62.5486</td>
<td>8,8</td>
<td>-112.1494</td>
</tr>
<tr>
<td>6,7</td>
<td>-30.4262</td>
<td>7,10</td>
<td>-63.7268</td>
<td>9,3,9,3</td>
<td>-116.4896</td>
</tr>
<tr>
<td>7,8,3</td>
<td>-32.4951</td>
<td>7,9,3</td>
<td>-64.5385</td>
<td>5,8</td>
<td>-159.0740</td>
</tr>
<tr>
<td>7,7</td>
<td>-35.5102</td>
<td>7,8</td>
<td>-65.5933</td>
<td>8,9,3</td>
<td>-180.2220</td>
</tr>
<tr>
<td>6,6</td>
<td>-37.5522</td>
<td>5,7</td>
<td>-68.2620</td>
<td>5,10</td>
<td>-189.7327</td>
</tr>
<tr>
<td>8,3,8,3</td>
<td>-38.4137</td>
<td>8,3,9,3</td>
<td>-70.8842</td>
<td>5,9</td>
<td>-200.8951</td>
</tr>
<tr>
<td>5,6</td>
<td>-53.9335</td>
<td>7,9</td>
<td>-73.4591</td>
<td>5,9,3</td>
<td>-201.0075</td>
</tr>
<tr>
<td>5,8,3</td>
<td>-55.2076</td>
<td>6,10</td>
<td>-77.8182</td>
<td>8,10</td>
<td>-217.4320</td>
</tr>
<tr>
<td>6,9,3</td>
<td>-57.7477</td>
<td>5,5</td>
<td>-79.8455</td>
<td>10,9,3</td>
<td>-220.6531</td>
</tr>
<tr>
<td>6,8</td>
<td>-58.1526</td>
<td>10,8,3</td>
<td>-83.7774</td>
<td>8,9</td>
<td>-221.9224</td>
</tr>
<tr>
<td>8,8,3</td>
<td>-60.8104</td>
<td>9,8,3</td>
<td>-85.7567</td>
<td>9,9,3</td>
<td>-239.4455</td>
</tr>
</tbody>
</table>

Table 5.5.3: The $\log_{10}$ of the likelihoods of observing the DNA profile given different allele pairs. The settings were, $\Lambda \sim \mathcal{N}(1, 0.2), \pi_i \sim \ln\mathcal{N}(\mu_i, 1), \epsilon \sim \mathcal{N}(0, 10)$, Expected peak height = 500 rfu.

We see that the pairs 6-8,3, 6-7 and 7-8,3 have the highest likelihood.

### 5.5.2 Conclusion

Suppose that we found this profile of figure 5.5.1 on a crime scene, and we have a suspect that has the allele pair 6-7 on TH01, the hypotheses pair of which we want to know the likelihood ratio is,

- $H_p$: the suspect is the donor of the DNA profile on TH01
- $H_d$: some unknown man is the donor of the profile on TH01

To compute this, we need the prior probabilities of all the possible allele combinations (see table 5.5.4) as well as the corresponding likelihoods that are given in table 5.5.3. The likelihood ratio of the evidence is...
\( E \) (which is the DNA profile) given the two hypotheses is,

\[
\frac{\mathbb{P}(E|H_s)}{\mathbb{P}(E|H_d)} = \frac{\text{likelihood of } E \text{ given donor is } 6-7}{\sum_{\text{allele pairs}} \text{prior probability}(\text{pair } | H_d) \cdot \text{likelihood of } E \text{ given donor has pair}} 10^{-30.4262} \\
= \frac{p_{6,8.3} \cdot 10^{-29.5544} + p_{6,7} \cdot 10^{-30.4262} + \cdots + p_{8,9} \cdot 10^{-221.9224} + p_{9,9.3} \cdot 10^{-239.4455}}{13.0494}
\]

We see that it is 13 times more likely to find this DNA profile when the suspect is the donor of the profile than when an unknown man with an unknown pair of alleles is the donor of the profile.

In a case where there is no suspect, an idea is to use the list of likelihoods for the allelic pairs and their corresponding prior probabilities to make a list of posterior probabilities that give us an idea what the most likely DNA profile of the contributor of the crime stain is. This can be used to search for a suspect.

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>8.3</th>
<th>9.3</th>
<th>allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.0001</td>
<td>0.0030</td>
<td>0.0026</td>
<td>0.0016</td>
<td>0.0019</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0050</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>0.0440</td>
<td>0.0759</td>
<td>0.0466</td>
<td>0.0568</td>
<td>0.0038</td>
<td>0.0001</td>
<td>0.1454</td>
<td>0.0001</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>0.0327</td>
<td>0.0401</td>
<td>0.0489</td>
<td>0.0033</td>
<td>0.0001</td>
<td>0.1253</td>
<td>0.0000</td>
<td>0.0769</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>0.0123</td>
<td>0.0300</td>
<td>0.0020</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0769</td>
<td>0.0000</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>0.0183</td>
<td>0.0025</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>0.0000</td>
<td>0.1201</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 5.5.4: Prior probabilities for all allele combinations on locus TH01 assuming Hardy-Weinberg equilibrium

---

65
5.6 Data, assumptions and parameter estimation

In this section we will explain, discuss and examine the decisions we made for the different parameters in our model and show the data that was used in the model. In section 5.7 we will perform a sensitivity analysis on the different parameters.

5.6.1 Data

Loci and alleles

In our model we only included information of the following loci, D10, vWA, D16, D2S1338, D8, D21, D18, D22, D19, TH01, FGA, D2S441, D3, D1S1656 and D12. The prior distribution on the allele pairs is achieved by simply multiplying the allele frequencies of that loci (and for heterozygotes, with a factor 2). We assume that having one allele on a locus does not say anything about the other allele on that locus, which means we can simply multiply the two probabilities to find the priors. This assumption can be changed, but for now we assume the Hardy-Weinberg equilibrium\(^1\). For example, the probabilities for observing the different alleles of locus D22S1045 are given in table 5.6.1.

<table>
<thead>
<tr>
<th>allele</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0.0014</td>
<td>0.1405</td>
<td>0.0132</td>
<td>0.0060</td>
<td>0.0508</td>
<td>0.3257</td>
<td>0.3705</td>
<td>0.0842</td>
<td>0.0067</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

Table 5.6.1: Allele frequencies for the alleles on locus D22S1045

Using these values we can compute the prior probabilities for all the possible allele combinations on the locus. Their prior probabilities are given in table 5.6.2.

<table>
<thead>
<tr>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000</td>
<td>0.0004</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0001</td>
<td>0.0009</td>
<td>0.0011</td>
<td>0.0002</td>
<td>0.0000</td>
<td>0.0000</td>
<td>10</td>
</tr>
<tr>
<td>0.00197</td>
<td>0.0037</td>
<td>0.0017</td>
<td>0.0143</td>
<td>0.00915</td>
<td>0.001041</td>
<td>0.0015237</td>
<td>0.00019</td>
<td>0.0003</td>
<td>0.0003</td>
<td>11</td>
</tr>
<tr>
<td>0.0002</td>
<td>0.0002</td>
<td>0.0013</td>
<td>0.0086</td>
<td>0.0098</td>
<td>0.0022</td>
<td>0.0002</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>12</td>
</tr>
<tr>
<td>0.0000</td>
<td>0.0006</td>
<td>0.00039</td>
<td>0.0044</td>
<td>0.000010</td>
<td>0.000010</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>13</td>
</tr>
<tr>
<td>0.00026</td>
<td>0.033</td>
<td>0.0377</td>
<td>0.0086</td>
<td>0.00007</td>
<td>0.00001</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>14</td>
</tr>
<tr>
<td>0.01061</td>
<td>0.2413</td>
<td>0.0548</td>
<td>0.0044</td>
<td>0.00006</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>15</td>
</tr>
<tr>
<td>0.1373</td>
<td>0.0624</td>
<td>0.0050</td>
<td>0.00007</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>16</td>
</tr>
<tr>
<td>0.0071</td>
<td>0.011</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>17</td>
</tr>
<tr>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>18</td>
</tr>
<tr>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 5.6.2: Prior probabilities for all allele combinations on locus D22S1045

The alleles that we consider on each locus are chosen due to data on the allele frequencies of the alleles on each locus. It is possible that a person has an allele on a locus that is not in our prior distribution, since the prior distribution is made on the basis of a DNA profile database that cannot contain data of all living people. In our model, the alleles that we take into account are given in table B.1.1 in appendix B.1.

Stutter distribution

Our stutter distribution is chosen on the basis of data regarding stutters, kindly provided by Laurens Grol from the NFI. This data supports the assumption that the stutter distribution is lognormal, where the mean differs for each allele and locus, and where the standard deviation is 1. We will give one plot that shows the reported stutter percentages against their alleles. The plot in figure 5.6.2 shows this\(^2\).

---
\(^1\)The Hardy-Weinberg equilibrium is a state in which both allele and genotype frequencies in a population remain constant. That is, they are in equilibrium from generation to generation unless specific disturbing influences are introduced.

\(^2\)The data provided by Laurens Grol from the NFI.
information for the locus **D10**. We see that the average stutter percentage is higher for higher numbered alleles. The plot in figure 5.6.1 shows box plots of the stutter percentages of all loci. The box plots suggest a lognormal distribution and different parameters for different loci.

![Figure 5.6.1: Box plots of recorded stutter on different loci](image1)

![Figure 5.6.2: A plot of stutter percentages on the different alleles of locus D10](image2)

### 5.6.2 Assumptions

To work with our method we made some assumptions on the parameters and the distributions. The assumptions on the distributions of our variables are,

- The distribution for the noise $\epsilon$ is normal, with mean 0 and standard deviation 10.
- The distribution for the amplification factor $\lambda$ is normal, with mean 1 and standard deviation 0.2.
- The distribution for the stutters $\pi$ is lognormal, with a mean that is different for each allele and each locus, the standard deviation is 0.01 for each locus.
- The distribution for the contamination is heavy tailed and of the following form,

$$P(c > x) = 1 - \frac{1}{\sqrt{99 \cdot x + 1}}$$

We choose this distribution to have, $P(c > 1) = 0.1$.

### 5.6.3 Parameter estimation

In this section we shortly discuss how we estimated the values for our parameters. It is important to keep in mind that the method can work with different parameters and distributions. The decisions we made on distributions and parameters only represent how we think the distributions are shaped. If additional information comes available about the parameters or distributions, we can just change this in our method without harming it.

**The stutter distribution $\pi$**

As said, we received data on stutters on different loci and alleles. The box plot in Figure 5.6.1 suggests a lognormal or a normal distribution. The data provided us with estimates for the mean and standard deviation. The standard deviation seemed approximately constant over all loci and alleles and is set to 1%. The mean stutter percentage was different for different loci and even for different alleles on the same loci. We used the estimations from these data in our model. In the sensitivity analysis, we will see what the effect is of changing the distribution for the stutters from lognormal to normal.

**The amplification factor distribution $\Lambda$**

For the amplification factor $\Lambda$ we assumed a normal distribution with mean 1. Intuitively, this is correct. The exact amount of output DNA is unknown, but given the amount of input DNA and the number of amplification cycles and other experimental settings, a DNA specialist has an idea of the expected peak height. What remains is, what to choose for the standard deviation $\sigma$ of $\Lambda$. We choose it to be $\sigma = 0.2$, a decision based on analyzing DNA profiles we obtained and of which we looked at loci where there was a peak imbalance. We will examine the outcome of this decision in our sensitivity analysis.

**The expected peak height**

As said, a DNA specialist has an expectation of what the peak heights in the profile will be. We thought of three ways to determine the expected peak height (these are the values of $P$ at the alleles of interest). The first one is that a DNA specialist gives the peak heights that she expects. Another way of choosing the expected peak height is by taking the average peak height of the outcome profile $x$, i.e. if the DNA profile (on one locus) is

<table>
<thead>
<tr>
<th>allele</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>peak height (rfu)</td>
<td>10</td>
<td>30</td>
<td>15</td>
<td>280</td>
<td>300</td>
<td>12</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

That we could take the expected peak height as,

$$\text{Expected peak height} = \frac{1}{2} \cdot (10 + 30 + 15 + 280 + 300 + 12 + 2 + 7) = 328 \text{ rfu}$$

We could do this for each locus, and get a different value. We could also do this for the whole profile and use the same peak height on each locus. We will examine the results of choosing any of these methods over the other in the sensitivity analysis. Note that determining the expected peak height using this method will make it dependent of $x$. Therefore, we suggest additional research for this method when applied in practice.
The distribution of the noise $\epsilon$

We assumed a normal distribution for the noise $\epsilon$ with mean 0. This means that the noise can be negative, (which we will often use in our model to explain a profile), but a DNA profile never shows negative peak heights. We assume the standard deviation to be 10, but we will examine what happens with our results when we change this value in the sensitivity analysis. We think it would be wise to further investigate this noise in DNA profiles. Does the amount of noise depend on the amount of input DNA? Is the distribution of the noise the same on all loci?

The distribution of the contamination $c$

The distribution of the contamination $c$ is very hard to determine. A reasonable assumption would be that higher contaminations have lower probabilities of occurring, and that a ‘high’ contamination (i.e. one of the same height as the height of one of the alleles of the donor) is unlikely but not impossible. We choose the following distribution for the contamination,

$$P(c \leq a) = 1 - \frac{1}{\sqrt{1 + 99a}}$$

This distribution has the property that 90% of the outcomes is below 1, but it is heavy tailed so high values of $c$ are still possible. To give an idea how the distribution is shaped we will give a plot that shows the probability density function of the contamination, which is equal to,

$$f_c(x) = \frac{99}{2} \cdot \frac{2}{(1 + 99x)^{3/2}}$$

The plot is given in figure 5.6.3.

Figure 5.6.3: The probability density function of the contamination.

We will compare this distribution with a lognormal distribution for the contamination in the sensitivity analysis.

5.7 Sensitivity analysis

In this section we will do a sensitivity analysis on our parameters. We will examine which decisions on the values of different parameters will do with our results, i.e. how sensitive the results are for the values
of the parameters. We will do this on two different profiles; an ‘easy’ profile, where there are two clear peaks at each locus, and a ‘hard’ profile where it is harder to determine the profile of the donor. The sensitivity analysis of the more difficult profile can be found in appendix [5.3]. The ‘easy profile’ is given in table 5.7.1.

<table>
<thead>
<tr>
<th>locus</th>
<th>allele 1</th>
<th>height 1</th>
<th>allele 2</th>
<th>height 2</th>
<th>allele 3</th>
<th>height 3</th>
<th>allele 4</th>
<th>height 4</th>
<th>allele 5</th>
<th>height 5</th>
<th>allele 6</th>
<th>height 6</th>
<th>allele 7</th>
<th>height 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>vWA</td>
<td>10</td>
<td>21</td>
<td>13</td>
<td>18</td>
<td>16</td>
<td>1871</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D16</td>
<td>9</td>
<td>568</td>
<td>12</td>
<td>577</td>
<td>13</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td>16</td>
<td>625</td>
<td>19</td>
<td>973</td>
<td>21</td>
<td>4</td>
<td>23</td>
<td>96</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D8</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>1069</td>
<td>15</td>
<td>747</td>
<td>17</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D21</td>
<td>25.2</td>
<td>3</td>
<td>27.2</td>
<td>5</td>
<td>28</td>
<td>846</td>
<td>29</td>
<td>9</td>
<td>31</td>
<td>1071</td>
<td>32</td>
<td>9</td>
<td>35.2</td>
<td>18</td>
</tr>
<tr>
<td>D18</td>
<td>13</td>
<td>769</td>
<td>15</td>
<td>576</td>
<td>19</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D19</td>
<td>10</td>
<td>6</td>
<td>13</td>
<td>831</td>
<td>14</td>
<td>602</td>
<td>15.2</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TH0</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>703</td>
<td>7</td>
<td>889</td>
<td>9</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGA</td>
<td>20</td>
<td>555</td>
<td>21</td>
<td>32</td>
<td>25</td>
<td>338</td>
<td>26</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td>14</td>
<td>673</td>
<td>16</td>
<td>770</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.7.1: Alleles and peak heights of the different loci from our DNA profile

The table should be read in the following way: The rows represent the different loci. In the columns there are alternately the alleles (the gray bars) and their peak heights. From the table we see that not on each possible allele a peak was found. The columns bare the names allele 1, height 1, . . . . These represent the order in which the DNA profiling machine found the alleles. The first allele found on a locus is given the name allele 1, but at the locus vWA, this was the allele with number 10. We will assume that the donor of this profile bears the alleles on the different loci given in table 5.7.2; we will use these combinations of alleles to compare our results in the sensitivity analysis. In the sensitivity analysis, we will show these pairs in green.

The settings that we will change and what the possible choices are for each are given in table 5.7.3.

Intuitively, one would expect that a higher standard deviation for \( \Lambda \) or \( \epsilon \) will lead to more possible profiles. It becomes more likely that a combination that does not seem the contributor of the DNA profile gets a reasonable likelihood due to extreme values for \( \epsilon \) and \( \Lambda \). The same holds for \( \pi \). If we change

<table>
<thead>
<tr>
<th>( M )</th>
<th>the number of times we estimate the likelihood for each possible combination of alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi )</td>
<td>lognormal</td>
</tr>
<tr>
<td>( \Lambda )</td>
<td>normal</td>
</tr>
<tr>
<td>standard deviation ( \Lambda )</td>
<td>the standard deviation on the amplification factor</td>
</tr>
<tr>
<td>standard deviation ( \epsilon )</td>
<td>the standard deviation on the noise</td>
</tr>
<tr>
<td>expected peak height method</td>
<td>mean height per locus</td>
</tr>
<tr>
<td>contamination distribution</td>
<td>original</td>
</tr>
</tbody>
</table>

Table 5.7.3: The settings we will differ during the sensitivity analysis
the distribution of the stutters from lognormal to normal, it becomes less likely to find ‘extreme’ values, since the lognormal distribution has a heavier tail than the normal distribution has. In these cases, it is important to look at the difference in the likelihood ratio’s between two simulations with different settings. It is clear that when we make the standard deviation of $\epsilon$ higher, that the likelihoods of all the pairs will become higher, but what happens with the likelihood ratio’s of pairs on the same locus? Will the difference remain similar? Lastly, it will be interesting to see whether the pairs that we assumed to be the allelic pairs of the donor will give the highest likelihoods in all our simulations.

We did one simulation with the following settings,

$M$  | stutter distribution | amplification distribution | standard deviation $\epsilon$ | standard deviation $\Lambda$ | Expected peak height method | Contamination distribution
--- | --- | --- | --- | --- | --- | ---
10 000 | lognormal | normal | 10 | 0.2 | per locus | original

where, the expected peak height per locus will give us the expected peak heights as in table 5.7.4.

Table 5.7.4: The expected peak heights per locus

The results of this simulation are given in table 5.7.5 (we only give the 5 most likely combinations on each allele).

Table 5.7.5: The log$_{10}$ of the likelihoods of observing the DNA profile given different allele pairs under our standard assumptions.

We see that under our ‘standard’ assumptions on the parameters, the allelic pairs that we assumed to be of the donor have the highest likelihoods on each locus. On some loci, the likelihood ratios between the assumed pair and the other pairs are not that large. Especially at D19, the difference is almost negligible.

We will compare the outcome of this ‘reference’ simulation with all of the following simulations. In each of the following simulations we will change only one parameter and remain all others unchanged, i.e. in the simulation where we changed the distribution for $\Lambda$ from normal to uniform, the standard deviation for the noise is still 0.2, the distribution for the stutters is still lognormal, etc.

5.7.1 Stutter distribution

We assumed our stutter distribution to be lognormal, with standard deviation 1 and a mean that differs for each locus on each allele. The mean of the lognormal distribution was linear over the different alleles on the same locus. We will see what happens when we take a normal stutter distribution over all loci,
Normal distribution for the stutters

Here we give the results when a normally distributed stutter percentage is used, where the mean differs for each allele on each locus and where the standard deviation is set to be 0.01. We took $M = 1000$. The results of the simulation are given in table 5.7.6.

<table>
<thead>
<tr>
<th></th>
<th>vWA</th>
<th>D16</th>
<th>D2</th>
<th>D8</th>
<th>D21</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-16</td>
<td>-44</td>
<td>9-12</td>
<td>-30</td>
<td>16-19</td>
<td>-61</td>
</tr>
<tr>
<td>16-17</td>
<td>-138</td>
<td>9-9</td>
<td>-33</td>
<td>19-19</td>
<td>-64</td>
</tr>
<tr>
<td>11-16</td>
<td>-202</td>
<td>12-12</td>
<td>-36</td>
<td>16-16</td>
<td>-73</td>
</tr>
<tr>
<td>16-20</td>
<td>-230</td>
<td>9-11</td>
<td>-86</td>
<td>16-23</td>
<td>-84</td>
</tr>
<tr>
<td>16-18</td>
<td>-259</td>
<td>12-14</td>
<td>-97</td>
<td>19-23</td>
<td>-144</td>
</tr>
<tr>
<td>13-13</td>
<td>-47</td>
<td>13-14</td>
<td>-51</td>
<td>6-7</td>
<td>-25</td>
</tr>
<tr>
<td>13-15</td>
<td>-49</td>
<td>13-13</td>
<td>-52</td>
<td>7-7</td>
<td>-28</td>
</tr>
<tr>
<td>15-15</td>
<td>-58</td>
<td>14-14</td>
<td>-59</td>
<td>6-6</td>
<td>-30</td>
</tr>
<tr>
<td>13-18</td>
<td>-60</td>
<td>14-15</td>
<td>-101</td>
<td>7-8</td>
<td>-140</td>
</tr>
</tbody>
</table>

Table 5.7.6: The log$_{10}$ of the likelihoods of observing the DNA profile given different allele pairs with a normally distributed stutter percentage.

The first thing that we notice is that the assumed allelic pair 13-15 is not the pair with the highest likelihood on locus D18. Another pair, 13-13, has a slightly higher likelihood. If we look at the results as a whole, we see a lot of similarities with the results from our reference simulation. It is important to note that the approximated likelihood that we find in our simulations is usually due to these simulation rounds where the likelihood was quite big. Since we take the mean of $M$ (which was 1000 in this simulation) likelihoods, a sufficient part of the simulation rounds will lead to a likelihood that can be almost discarded, since it is a lot lower than the biggest ones.

5.7.2 Noise distribution

We assumed the noise $\epsilon$ to be normally distributed with mean 0 and standard deviation 10. What happens when we change the standard deviation of the noise. As said, we expect the likelihoods to become lower when the standard deviation decreases and we expect the likelihoods to become higher when the standard deviation increases.

Standard deviation of 5

We set the standard deviation of the noise to 5 and got the results given in table 5.7.7. The results from the simulations are in line with our expectations. The assumed allelic pair has the highest likelihood on all loci, but the likelihood ratio’s between the pair with the highest likelihood and the other pairs on that locus have increased. Especially at locus vWA, where the pair 16-16 is the only pair that has a non-zero likelihood. The loci where we might be uncertain about the profile of the donor are D2 and FGA where the difference in the approximated likelihoods is relatively small. This can be due to relatively low peaks at the locus or because there is a peak imbalance at this locus (one peak is significantly bigger than the other).

Standard deviation of 20

We will also investigate what happens when we change the standard deviation of the noise from 10 to 20. It is clear that it becomes easier to explain peaks in a profile when the standard deviation of the noise is higher. Therefore, we expect to find smaller likelihood ratio’s on all loci. There is no reason to expect
Table 5.7.7: The log_{10} of the likelihoods of observing the DNA profile given different allele pairs where the noise distribution was \( N(0,5) \).

Table 5.7.8: The log_{10} of the likelihoods of observing the DNA profile given different allele pairs where the noise distribution was \( N(0,20) \).

5.7.3 \( \Lambda \) distribution

\( \Lambda \) is the factor that represents the amplification process. Based on the input amount of DNA, an expert can give an estimation of the peak height it will produce. Due to the fact that the amplification process is a stochastic process, there is some variation around this peak height. We assumed it to be normally distributed with mean 1 and standard deviation 0.2. We will see what happens when we change the standard deviation to 0.1 and to 0.3. We will also look at the results where the distribution of \( \Lambda \) is uniform on (0.8,1.2).
Standard deviation 0.1

Here we will examine what happens when we set the standard deviation of Λ to 0.1. Intuitively, this means that the likelihood of profiles that are not immediately considered to be the donor of the profile becomes lower. Since the simulated vectors Λ will contain less ‘extreme’ values, it becomes more unlikely that the amplification factor is, for instance 0.5 or 1.5 when the standard deviation is higher. We expect similar results for the allelic pairs that we assume to be of the donor, but lower values for the likelihoods of different combinations. The results of this simulation are presented in table 5.7.9.

<table>
<thead>
<tr>
<th>pair</th>
<th>likelihood</th>
<th>pair</th>
<th>likelihood</th>
<th>pair</th>
<th>likelihood</th>
<th>pair</th>
<th>likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-16</td>
<td>−32</td>
<td>9-12</td>
<td>−27</td>
<td>16-19</td>
<td>−53</td>
<td>12-15</td>
<td>−31</td>
</tr>
<tr>
<td>16-17</td>
<td>−248</td>
<td>9-13</td>
<td>−84</td>
<td>19-19</td>
<td>−97</td>
<td>12-12</td>
<td>−87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12-12</td>
<td>−89</td>
<td>17-19</td>
<td>−145</td>
<td>13-15</td>
<td>−107</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9-9</td>
<td>−124</td>
<td>16-20</td>
<td>−177</td>
<td>12-16</td>
<td>−172</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10-12</td>
<td>−144</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D18</td>
<td></td>
<td>D19</td>
<td></td>
<td>TH0</td>
<td></td>
<td>FGA</td>
<td></td>
</tr>
<tr>
<td>13-15</td>
<td>−43</td>
<td>13-16</td>
<td>−47</td>
<td>6-7</td>
<td>−26</td>
<td>20-25</td>
<td>−52</td>
</tr>
<tr>
<td>14-15</td>
<td>−68</td>
<td>14-14</td>
<td>−57</td>
<td>6-8</td>
<td>−48</td>
<td>20-20</td>
<td>−53</td>
</tr>
<tr>
<td>13-13</td>
<td>−75</td>
<td>13-13</td>
<td>−103</td>
<td>7-7</td>
<td>−66</td>
<td>21-25</td>
<td>−90</td>
</tr>
<tr>
<td>13-16</td>
<td>−170</td>
<td>13-15</td>
<td>−294</td>
<td></td>
<td></td>
<td>20-26</td>
<td>−114</td>
</tr>
<tr>
<td>15-15</td>
<td>−225</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20-21</td>
<td>−152</td>
</tr>
</tbody>
</table>

Table 5.7.9: The log10 of the likelihoods of observing the DNA profile given different allele pairs where the standard deviation of the amplification factor Λ was 0.1.

On all loci, except for FGA, the results show a clear distinction between the assumed allelic pair and the other pairs. On FGA however, we see a peak imbalance in the DNA profile that could explain the relatively small difference in the likelihood.

Standard deviation 0.3

When the standard deviation of the amplification factor is higher, we can follow the same reasoning as in the part where the standard deviation was set to 0.1. We will expect similar results for the allelic pairs that we assumed to be of the donor and we expect the likelihoods for the other pairs to be higher, since more extreme values of Λ become more likely. The results of this simulation are in table 5.7.10.

The results show less distinction between the allelic pairs. On some of the loci, the allelic pair that we assumed to be of the donor does not even have the highest likelihood, although it is still of the same order as the allelic pair with the highest likelihood. Due to the fact that the values of Λ can get higher, as well as lower, this is not very remarkable.

Uniform distribution

In this subsection we will examine what happens with our results when we change the distribution of Λ. A normal distribution makes it possible to find extreme values (although with a very low probability) where the uniform distribution excludes a lot of values for Λ. We choose the boundaries to be 0.8 and 1.2 such that all the values for Λ are between these two boundaries. We can follow the same argument as we did before. Since a uniform distribution makes it less likely to see extreme values for Λ, we expect to see lower likelihoods for the allelic pairs that does not seem to fit in the profile. The results of our simulation can be found in table 5.7.11.

In this simulation we see the clearest distinction between the assumed allelic pairs and the other pairs on all loci. On each locus, the likelihood ratio between the assumed allelic pair and the others is at least 10^{14}.
Table 5.7.10: The \(\log_{10}\) of the likelihoods of observing the DNA profile given different allele pairs where the standard deviation of the amplification factor \(\Lambda\) was 0.3.

<table>
<thead>
<tr>
<th>vWA</th>
<th>D16</th>
<th>D2</th>
<th>D8</th>
<th>D21</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>16-16</td>
<td>-31</td>
<td>9-12</td>
<td>-28</td>
<td>16-16</td>
</tr>
<tr>
<td>15-16</td>
<td>-41</td>
<td>12-12</td>
<td>-32</td>
<td>19-19</td>
</tr>
<tr>
<td>16-19</td>
<td>-58</td>
<td>11-12</td>
<td>-38</td>
<td>16-17</td>
</tr>
</tbody>
</table>

Table 5.7.11: The \(\log_{10}\) of the likelihoods of observing the DNA profile given different allele pairs where the distribution of the amplification factor \(\Lambda\) was \(U(0.8, 1.2)\).

<table>
<thead>
<tr>
<th>vWA</th>
<th>D16</th>
<th>D2</th>
<th>D8</th>
<th>D21</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>13-15</td>
<td>-46</td>
<td>13-14</td>
<td>-46</td>
<td>6-7</td>
</tr>
<tr>
<td>13-13</td>
<td>-46</td>
<td>13-13</td>
<td>-51</td>
<td>7-7</td>
</tr>
<tr>
<td>15-15</td>
<td>-48</td>
<td>14-14</td>
<td>-54</td>
<td>6-6</td>
</tr>
<tr>
<td>15-23</td>
<td>-55</td>
<td>14-17</td>
<td>-56</td>
<td>6-9</td>
</tr>
</tbody>
</table>

5.7.4 Expected Peak Height

To estimate the peak height, we computed the sum of the peak heights and divided it by 2 for each locus. What happens when we change the way we estimate this peak height? We will now take the sum of all peak heights and divide it by 2 times the number of loci. This means that we will take the same peak height on each locus, 767.55. The results after simulation are presented in table 5.7.12.

If we compare the results from this simulation (table 5.7.12) with the results from our reference simulation (table 5.7.5), we see that the allelic pairs with the highest likelihood all have the approximate same likelihood where all the other pairs have a much lower likelihood. It becomes much more likely to observe the assumed allelic pairs when we change the way we choose the expected peak height.

5.7.5 Contamination distribution

We choose the contamination distribution to be a heavy-tailed. This is because you would expect (almost) no contamination in most cases, but it is possible to find very high contamination amounts. We could take different distributions to see what the consequences are for our results. Here we choose the contamination
**Table 5.7.12:** The log\(_{10}\) of the likelihoods of observing the DNA profile given different allele pairs where the expected peak height was taken equal on all loci.

<table>
<thead>
<tr>
<th>vWA</th>
<th>D16</th>
<th>D2</th>
<th>D8</th>
<th>D21</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>16-16</td>
<td>−31</td>
<td>9-12</td>
<td>−28</td>
<td>16-19</td>
</tr>
<tr>
<td>16-17</td>
<td>−40</td>
<td>12-12</td>
<td>−34</td>
<td>19-19</td>
</tr>
<tr>
<td>16-21</td>
<td>−76</td>
<td>9-9</td>
<td>−37</td>
<td>16-16</td>
</tr>
<tr>
<td>11-16</td>
<td>−96</td>
<td>11-12</td>
<td>−59</td>
<td>16-20</td>
</tr>
</tbody>
</table>

**Table 5.7.13:** The log\(_{10}\) of the likelihoods of observing the DNA profile when we use the original distribution for the contamination.

<table>
<thead>
<tr>
<th>D18</th>
<th>D19</th>
<th>TH0</th>
<th>FGA</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>13-15</td>
<td>−43</td>
<td>13-14</td>
<td>−46</td>
<td>6-7</td>
</tr>
<tr>
<td>13-13</td>
<td>−47</td>
<td>14-14</td>
<td>−51</td>
<td>7-7</td>
</tr>
<tr>
<td>15-15</td>
<td>−48</td>
<td>13-13</td>
<td>−52</td>
<td>6-6</td>
</tr>
<tr>
<td>14-15</td>
<td>−57</td>
<td>14-15</td>
<td>−57</td>
<td>6-8</td>
</tr>
<tr>
<td>9-15</td>
<td>−74</td>
<td>13-15</td>
<td>−69</td>
<td>7-9.3</td>
</tr>
</tbody>
</table>

The results are satisfying. We find approximately the same results under both distributions. Since the distributions have approximately the same shape, for both 90% of the mass is below 1 and 99% below 101, this is not a very surprising result. Therefore, we will also compare the results for two contamination distributions that have different shapes.

We will use our original distribution for \(c, f_c(x) = \frac{99}{2(99x+1)^{\frac{3}{2}}}, and a lognormal distribution where 90% is below 10, and 99% below 250. The results are given in table 5.7.15 and 5.7.16.

The most important difference between the results is that the results from the simulation where the contamination was lognormally distributed are shifted compared with the results from the simulation.
Table 5.7.14: The log\(_{10}\) of the likelihoods of observing the DNA profile when we use a lognormal distribution for the likelihood where 99% of the probability mass is below 100.

<table>
<thead>
<tr>
<th>vWA</th>
<th>D16</th>
<th>D2</th>
<th>D8</th>
<th>D21</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>12-16</td>
<td>-105</td>
<td>12-12</td>
<td>-33</td>
<td>19-19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D18</th>
<th>D19</th>
<th>TH0</th>
<th>FGA</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>13-13</td>
<td>-51</td>
<td>14-14</td>
<td>-51</td>
<td>7-7</td>
</tr>
</tbody>
</table>

Table 5.7.15: The log\(_{10}\) of the likelihoods of observing the DNA profile when we use the original distribution for the likelihood.

<table>
<thead>
<tr>
<th>vWA</th>
<th>D16</th>
<th>D2</th>
<th>D8</th>
<th>D21</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>12-16</td>
<td>-105</td>
<td>12-12</td>
<td>-32</td>
<td>19-19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D18</th>
<th>D19</th>
<th>TH0</th>
<th>FGA</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>13-15</td>
<td>-41</td>
<td>13-14</td>
<td>-45</td>
<td>6-7</td>
</tr>
<tr>
<td>13-13</td>
<td>-51</td>
<td>14-14</td>
<td>-51</td>
<td>7-7</td>
</tr>
</tbody>
</table>

Table 5.7.16: The log\(_{10}\) of the likelihoods of observing the DNA profile when we use the lognormal distribution for the likelihood where 99% of the probability mass is below 250.

<table>
<thead>
<tr>
<th>vWA</th>
<th>D16</th>
<th>D2</th>
<th>D8</th>
<th>D21</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>16-16</td>
<td>-29</td>
<td>9-12</td>
<td>-23</td>
<td>16-19</td>
</tr>
<tr>
<td>12-16</td>
<td>-99</td>
<td>12-12</td>
<td>-29</td>
<td>19-19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D18</th>
<th>D19</th>
<th>TH0</th>
<th>FGA</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>13-15</td>
<td>-36</td>
<td>13-14</td>
<td>-38</td>
<td>6-7</td>
</tr>
<tr>
<td>13-13</td>
<td>-43</td>
<td>14-14</td>
<td>-43</td>
<td>7-7</td>
</tr>
</tbody>
</table>

with the original contamination distribution. The results are in line with our expectations since under the lognormal distribution it is more likely to find high values for the contamination. The likelihood ratios on the loci are, however, approximately the same. This is a satisfying result, the shape of the contamination distribution does not seem to be able to radically change the results.
5.8 Consequences of using the direction of the MLE for finding the ‘new MLE’

In theory, we would like to use the maximum likelihood estimator \( \hat{c} \) in our simulation, but determining this value is an expensive procedure. Therefore, we decided to use the direction of the maximum likelihood estimator \( \hat{c} \) when \( \pi \) and \( \Lambda \) are their expected values to compute the vector \( c \). In this section we will see what happens with our results when we change the way we determine \( \hat{c} \). We will do this in one simulation, in order to have the exact same values for \( \Lambda \) and \( \pi \) each time. This way, we are in the best situation to compare the results. We used the profile from table \ref{tab:th0} and the following settings,

\[
\begin{array}{cccccc}
M & stutter distribution & amplification distribution & standard deviation \epsilon & standard deviation \Lambda & Expected peak height method & Contamination distribution \\
10 000 & lognormal & normal & 10 & 0.2 & per locus & original \\
\end{array}
\]

The results of this simulation can be found in tables \ref{tab:5.8.1} and \ref{tab:5.8.2}.

<table>
<thead>
<tr>
<th>( M )</th>
<th>( \text{stutter distribution} )</th>
<th>( \text{amplification distribution} )</th>
<th>( \text{standard deviation } \epsilon )</th>
<th>( \text{standard deviation } \Lambda )</th>
<th>( \text{Expected peak height method} )</th>
<th>( \text{Contamination distribution} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 000</td>
<td>lognormal</td>
<td>normal</td>
<td>10</td>
<td>0.2</td>
<td>per locus</td>
<td>original</td>
</tr>
</tbody>
</table>

\[\text{Table 5.8.1: The log}_{10} \text{ of the likelihoods of observing the DNA profile when we use the direction to compute } \hat{c}.\]

<table>
<thead>
<tr>
<th>( vWA )</th>
<th>( D16 )</th>
<th>( D2 )</th>
<th>( D8 )</th>
<th>( D21 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>16-17</td>
<td>-134.5448</td>
<td>9-9</td>
<td>-34.4995</td>
<td>16-19</td>
</tr>
<tr>
<td>15-16</td>
<td>-238.3434</td>
<td>9-13</td>
<td>-74.7937</td>
<td>19-23</td>
</tr>
</tbody>
</table>

\[\text{Table 5.8.2: The log}_{10} \text{ of the likelihoods of observing the DNA profile when we compute } \hat{c} \text{ each time.}\]

We see some remarkable differences between the two tables. On most loci, the results are comparable (\( vWA, D16, D8, FGA \) and \( D3 \)), but on some loci, we see less nice results. On \( D2 \) and \( TH0 \) the most
likely contributor is different. Apart from these pairs, the estimated likelihoods on the loci are the same, but these pairs that show up in the simulation where we use the direction of $\tilde{c}$, became higher.

5.9 Values for the simulated likelihoods

In this section we will examine the values for the likelihoods that we simulate. We will give a histogram of the simulated likelihoods for some fixed DNA profile and a fixed allele combination $P$. Since we deal with very low values for the likelihood we will use the log of the likelihood instead. This means that we will ignore the likelihood values that were equal to 0. We will show a histogram of the simulated likelihood values for 4 different simulation. In 2 cases we will give the histogram of the simulated likelihood where we assumed the allelic pairs that belong to the donor and in 2 cases we will give the histogram of the simulated likelihood where we assumed an allelic pair that did not belong to the donor. We used the standard settings for our simulation with $M = 1000$. As DNA profile we used the DNA profile from the sensitivity analysis, given in table 5.7.1.

Values for the simulated likelihoods for the allele combination 16-16 on vWA

In figure 5.9.1 we see a histogram of the simulated values for the likelihood. The allelic combination 16-16 is the same as the allele combination of the donor of the DNA profile. This can be seen from the fact that a relative big amount of the simulated values for the likelihood values is ‘high’.

![Figure 5.9.1: A histogram of the log of the simulated likelihoods for the allele combination 16-16 on vWA](image)

Values for the simulated likelihoods for the allele combination 7-7 on TH0

In figure 5.9.2 we see a histogram of the simulated values for the likelihood. The allelic combination 7-7 is different from the allele combination of the donor of the DNA profile.

Values for the simulated likelihoods for the allele combination 9-12 on D16

In figure 5.9.3 we see a histogram of the simulated values for the likelihood. The allelic combination 9-12 on the locus $D16$ is the same as the allele combination of the donor of the DNA profile. This can be seen from the fact that a relative big amount of the simulated values for the likelihood values is ‘high’.

Values for the simulated likelihoods for the allele combination 19-19 on D2S1338

In figure 5.9.4 we see a histogram of the simulated values for the likelihood. The allelic combination 19-19 is different from the allele combination of the donor of the DNA profile.
Figure 5.9.2: A histogram of the log of the simulated likelihoods for the allele combination 7-7 on TH0

Figure 5.9.3: A histogram of the log of the simulated likelihoods for the allele combination 9-12 on D16

Figure 5.9.4: A histogram of the log of the simulated likelihoods for the allele combination 9-12 on D2S1338

5.9.1 Conclusion

The results are satisfying. From the histograms we see that we take the mean of a ‘representative’ group, meaning that we don’t take the mean of a lot of zero valued likelihoods and one very big value of the likelihood.
5.10 Case Example

In this section we will use our method to come to a conclusion regarding 4 DNA profiles from the same stain. From a stain, one DNA samples was obtained. This sample was split into 4 new samples in the lab. These samples were amplified and 4 DNA profiles were made. So we have 4 profiles, each made from 31.25 picogram DNA, that came from the same stain. They were amplified separately (it is possible to amplify different samples at the same time with the same machine but separately, these samples were amplified separately but during the same run) and contain information on 15 loci. A reference profile was made as well. The profiles can be found in appendix C in tables C.1.1-C.1.3. The reference profile was made under ‘good’ conditions and therefore, the donor of that profile can be determined easily. So, (from table C.1.1) we conclude that the donor of the stain has the following allelic pairs on the different loci (table 5.10.1).

<table>
<thead>
<tr>
<th>Locus</th>
<th>Profile 1</th>
<th>Profile 2</th>
<th>Profile 3</th>
<th>Profile 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D10</td>
<td>102</td>
<td>95</td>
<td>110.5</td>
<td>184</td>
</tr>
<tr>
<td>vWA</td>
<td>131.5</td>
<td>81.5</td>
<td>207.5</td>
<td>159.5</td>
</tr>
<tr>
<td>D16</td>
<td>22.5</td>
<td>155</td>
<td>25.5</td>
<td>134</td>
</tr>
<tr>
<td>D2S</td>
<td>139</td>
<td>124</td>
<td>234</td>
<td>143</td>
</tr>
<tr>
<td>D8S</td>
<td>137.5</td>
<td>104</td>
<td>121.5</td>
<td>179.5</td>
</tr>
<tr>
<td>D21</td>
<td>62.5</td>
<td>217.5</td>
<td>88.5</td>
<td>156</td>
</tr>
<tr>
<td>D18</td>
<td>78</td>
<td>176</td>
<td>178</td>
<td>178</td>
</tr>
<tr>
<td>D22</td>
<td>176</td>
<td>103.5</td>
<td>122</td>
<td>128</td>
</tr>
<tr>
<td>D19</td>
<td>135</td>
<td>179.5</td>
<td>183.5</td>
<td>138.5</td>
</tr>
<tr>
<td>TH0</td>
<td>184</td>
<td>93</td>
<td>154</td>
<td>115</td>
</tr>
<tr>
<td>FGA</td>
<td>141.5</td>
<td>101</td>
<td>158</td>
<td>180</td>
</tr>
<tr>
<td>D2S</td>
<td>126</td>
<td>126.5</td>
<td>106</td>
<td>148.5</td>
</tr>
<tr>
<td>D3S</td>
<td>153.5</td>
<td>120.5</td>
<td>109</td>
<td>214</td>
</tr>
<tr>
<td>D1S</td>
<td>204</td>
<td>45.5</td>
<td>190</td>
<td>141.5</td>
</tr>
<tr>
<td>D12</td>
<td>115</td>
<td>87</td>
<td>80</td>
<td>157.5</td>
</tr>
</tbody>
</table>

Table 5.10.1: The alleles on the different loci of the donor of the crime stain of which we obtained four different DNA profiles

First we will determine the likelihood for each possible allele combination on each locus for the stain profiles separately. We will use the same settings for each simulation. The settings we used are,

<table>
<thead>
<tr>
<th>Setting</th>
<th>stutter distribution</th>
<th>amplification distribution</th>
<th>standard deviation $\epsilon$</th>
<th>standard deviation $\Lambda$</th>
<th>Expected peak height method</th>
<th>Contamination distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>lognormal</td>
<td>normal</td>
<td>10</td>
<td>0.2</td>
<td>per locus</td>
<td>original</td>
</tr>
</tbody>
</table>

The expected peak heights for the DNA profiles on all the loci are given in table 5.10.2.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Profile 1</th>
<th>Profile 2</th>
<th>Profile 3</th>
<th>Profile 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D10</td>
<td>102</td>
<td>95</td>
<td>110.5</td>
<td>184</td>
</tr>
<tr>
<td>vWA</td>
<td>131.5</td>
<td>81.5</td>
<td>207.5</td>
<td>159.5</td>
</tr>
<tr>
<td>D16</td>
<td>22.5</td>
<td>155</td>
<td>25.5</td>
<td>134</td>
</tr>
<tr>
<td>D2S</td>
<td>139</td>
<td>124</td>
<td>234</td>
<td>143</td>
</tr>
<tr>
<td>D8S</td>
<td>137.5</td>
<td>104</td>
<td>121.5</td>
<td>179.5</td>
</tr>
<tr>
<td>D21</td>
<td>62.5</td>
<td>217.5</td>
<td>88.5</td>
<td>156</td>
</tr>
<tr>
<td>D18</td>
<td>78</td>
<td>176</td>
<td>178</td>
<td>178</td>
</tr>
<tr>
<td>D22</td>
<td>176</td>
<td>103.5</td>
<td>122</td>
<td>128</td>
</tr>
<tr>
<td>D19</td>
<td>135</td>
<td>179.5</td>
<td>183.5</td>
<td>138.5</td>
</tr>
<tr>
<td>TH0</td>
<td>184</td>
<td>93</td>
<td>154</td>
<td>115</td>
</tr>
<tr>
<td>FGA</td>
<td>141.5</td>
<td>101</td>
<td>158</td>
<td>180</td>
</tr>
<tr>
<td>D2S</td>
<td>126</td>
<td>126.5</td>
<td>106</td>
<td>148.5</td>
</tr>
<tr>
<td>D3S</td>
<td>153.5</td>
<td>120.5</td>
<td>109</td>
<td>214</td>
</tr>
<tr>
<td>D1S</td>
<td>204</td>
<td>45.5</td>
<td>190</td>
<td>141.5</td>
</tr>
<tr>
<td>D12</td>
<td>115</td>
<td>87</td>
<td>80</td>
<td>157.5</td>
</tr>
</tbody>
</table>

Table 5.10.2: The expected peak heights on the different loci for the four profiles
First stain profile

<table>
<thead>
<tr>
<th>D10</th>
<th>D16</th>
<th>D2S</th>
<th>D8S</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>D21</th>
<th>D18</th>
<th>D22</th>
<th>D19</th>
<th>TH0</th>
</tr>
</thead>
<tbody>
<tr>
<td>28-28</td>
<td>-54</td>
<td>12-21</td>
<td>-43</td>
<td>11-17</td>
</tr>
<tr>
<td>28-31.2</td>
<td>-56</td>
<td>15-22</td>
<td>-43</td>
<td>16-16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FGA</th>
<th>D2S</th>
<th>D3S</th>
<th>D1S</th>
<th>D12</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>D2S</th>
<th>D3S</th>
<th>D1S</th>
<th>D12</th>
</tr>
</thead>
<tbody>
<tr>
<td>28-28</td>
<td>-54</td>
<td>12-21</td>
<td>-43</td>
</tr>
<tr>
<td>28-31.2</td>
<td>-56</td>
<td>15-22</td>
<td>-43</td>
</tr>
</tbody>
</table>

Table 5.10.3: The log\(_{10}\) of the likelihoods of observing the DNA profile of the first sample.

We see in table 5.10.3 that the assumed allelic pairs of the donor of the stain are always in the group of 5 most likely contributors. The overall differences in the likelihoods on the individual loci are much smaller than in the simulations we did in the sensitivity analysis. This is because the peak heights of the alleles in these profiles are generally much lower than the peak heights in the profile we used in the sensitivity analysis. For this profile we can compute the likelihood ratio between the following pair of hypotheses.

- **\(H_p\):** the suspect is the donor of the DNA profile
- **\(H_d\):** some unknown man is the donor of the DNA profile

The question is how to compute this likelihood ratio? The numerator is easy, if we assume independence between the loci, we can just multiply the likelihoods belonging to the allelic pairs of the suspect (the assumed allelic pairs). In the denominator we need to compute the likelihood associated with every possible profile to be the contributor of the DNA profile. If we assume independence between loci, we can compute this in the following way (we will denote the DNA profile with \(E\)).

\[
\frac{P(E|H_p)}{P(E|H_d)} = \prod_{\text{locus}} \frac{\text{likelihood of } E \text{ on locus, given donor is assumed allelic pair on locus}}{\sum_{\text{allelic pairs}} \text{prior probability}(\text{pair } | H_d) \cdot \text{likelihood}(E|\text{pair})} = 1.3877 \cdot 10^7
\]

Second stain profile

The results for this profile in table 5.10.4 show that on some loci (\(D10,D1S\) and \(D12\)) the assumed allele combination of the donor is not among the five with the highest likelihood. The likelihoods for the assumed allele combination on these loci were:

Table 5.10.4: The log\(_{10}\) of the likelihoods of observing the DNA profile of the second sample.
We assume independence between loci, and compute the likelihood ratio,

$$\frac{P(E|H_p)}{P(E|H_d)} = \prod_{\text{loci}} \frac{\text{likelihood of } E \text{ on locus, given donor is assumed allelic pair on locus}}{\sum_{\text{allelic pairs}} \text{prior probability}(\text{pair} | H_d) \cdot \text{likelihood}(E|\text{pair})}$$

$$= 16.8014$$

<table>
<thead>
<tr>
<th>D10</th>
<th>vWA</th>
<th>D16</th>
<th>D2S</th>
<th>D8S</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>15-15</td>
<td>−25</td>
<td>14-16</td>
<td>−30</td>
<td>6-10</td>
</tr>
<tr>
<td>15-16</td>
<td>−28</td>
<td>12-16</td>
<td>−32</td>
<td>10-10</td>
</tr>
<tr>
<td>15-17</td>
<td>−30</td>
<td>16-16</td>
<td>−33</td>
<td>9-11</td>
</tr>
<tr>
<td>15-19</td>
<td>−30</td>
<td>16-18</td>
<td>−33</td>
<td>8-9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D21</th>
<th>D18</th>
<th>D22</th>
<th>D19</th>
<th>TH0</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>28-28</td>
<td>−58</td>
<td>12-14</td>
<td>−42</td>
<td>16-16</td>
</tr>
<tr>
<td>28-32</td>
<td>−67</td>
<td>12-12</td>
<td>−46</td>
<td>11-11</td>
</tr>
<tr>
<td>28-35.2</td>
<td>−68</td>
<td>15-15</td>
<td>−46</td>
<td>16-18</td>
</tr>
<tr>
<td>31-31</td>
<td>−70</td>
<td>13-15</td>
<td>−50</td>
<td>16-19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FGA</th>
<th>D2S</th>
<th>D3S</th>
<th>D1S</th>
<th>D12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>24-26</td>
<td>−46</td>
<td>14-15</td>
<td>−27</td>
<td>15-16</td>
</tr>
<tr>
<td>20-24</td>
<td>−49</td>
<td>15-15</td>
<td>−33</td>
<td>15-18</td>
</tr>
<tr>
<td>19-24</td>
<td>−50</td>
<td>14-14</td>
<td>−35</td>
<td>15-15</td>
</tr>
<tr>
<td>24-24</td>
<td>−50</td>
<td>11-14</td>
<td>−39</td>
<td>16-20</td>
</tr>
</tbody>
</table>

Table 5.10.4: The log$_{10}$ of the likelihoods of observing the DNA profile of the second sample.
Table 5.10.5: The log_{10} of the likelihoods of observing the DNA profile of the third sample.

The results for this profile (see table 5.10.5) show that on one locus (D21) the assumed allele combination of the donor is not among the five with the highest likelihood. The likelihood for the assumed allele combination on this locus was,

\[
\log_{10}(\text{likelihood})_{\text{assumed combination}} = -62
\]

We assume independence between loci, and compute the likelihood ratio.

\[
\frac{P(E|H_d)}{P(E|H_a)} = \prod_{\text{locus}} \frac{\text{likelihood of } E \text{ on locus, given donor is assumed allelic pair on locus}}{\sum_{\text{allelic pairs}} \text{prior probability}(\text{pair} | H_d) \cdot \text{likelihood}(E | \text{pair})}
\]

\[
= 1028.66
\]
Fourth stain profile

<table>
<thead>
<tr>
<th>pair</th>
<th>likelihood</th>
<th>pair</th>
<th>likelihood</th>
<th>pair</th>
<th>likelihood</th>
<th>pair</th>
<th>likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>12-12</td>
<td>-32</td>
<td>16-16</td>
<td>-36</td>
<td>9-9</td>
<td>-28</td>
<td>23-27</td>
<td>-47</td>
</tr>
<tr>
<td>12-16</td>
<td>-32</td>
<td>13-16</td>
<td>-38</td>
<td>10-13</td>
<td>-29</td>
<td>20-27</td>
<td>-48</td>
</tr>
<tr>
<td>13-15</td>
<td>-33</td>
<td>14-14</td>
<td>-38</td>
<td>9-11</td>
<td>-32</td>
<td>20-20</td>
<td>-48</td>
</tr>
<tr>
<td>D21</td>
<td></td>
<td>D18</td>
<td></td>
<td>D22</td>
<td></td>
<td>D19</td>
<td></td>
</tr>
<tr>
<td>31-31</td>
<td>-55</td>
<td>12-12</td>
<td>-41</td>
<td>11-16</td>
<td>-27</td>
<td>14-14.2</td>
<td>-52</td>
</tr>
<tr>
<td>31-35</td>
<td>-62</td>
<td>12-23</td>
<td>-51</td>
<td>11-17</td>
<td>-34</td>
<td>14-16</td>
<td>-54</td>
</tr>
<tr>
<td>31-33.2</td>
<td>-62</td>
<td>9-12</td>
<td>-53</td>
<td>11-14</td>
<td>-34</td>
<td>14-14</td>
<td>-55</td>
</tr>
<tr>
<td>31-25.2</td>
<td>-65</td>
<td>12-21</td>
<td>-53</td>
<td>11-13</td>
<td>-35</td>
<td>10-14</td>
<td>-55</td>
</tr>
<tr>
<td>DFG</td>
<td></td>
<td>D2S</td>
<td></td>
<td>D3S</td>
<td></td>
<td>D1S</td>
<td></td>
</tr>
<tr>
<td>20-24</td>
<td>-53</td>
<td>14-14</td>
<td>-35</td>
<td>16-16</td>
<td>-32</td>
<td>16-17.3</td>
<td>-50</td>
</tr>
<tr>
<td>24-26</td>
<td>-54</td>
<td>14-15</td>
<td>-35</td>
<td>13-16</td>
<td>-32</td>
<td>16-16</td>
<td>-51</td>
</tr>
<tr>
<td>24-24</td>
<td>-56</td>
<td>14-16</td>
<td>-40</td>
<td>16-18</td>
<td>-34</td>
<td>16-18</td>
<td>-53</td>
</tr>
<tr>
<td>24-20.2</td>
<td>-57</td>
<td>13-14</td>
<td>-41</td>
<td>13-18</td>
<td>-39</td>
<td>16-19.3</td>
<td>-53</td>
</tr>
<tr>
<td>24-21.2</td>
<td>-60</td>
<td>11-14</td>
<td>-42</td>
<td>15-16</td>
<td>-40</td>
<td>15-16</td>
<td>-53</td>
</tr>
</tbody>
</table>

Table 5.10.6: The log_{10} of the likelihoods of observing the DNA profile of the fourth sample.

The results for this profile (see table 5.10.6) show that on some loci (D21 and D1S) the assumed allele combination of the donor is not among the five with the highest likelihood. The likelihoods for the assumed allele combination on these loci were,

We assume independence between loci, and compute the likelihood ratio

\[
\frac{P(E|H_d)}{P(E|H_a)} = \prod_{\text{loci}} \frac{\text{likelihood of } E \text{ on locus, given donor is assumed allelic pair on locus}}{\sum_{\text{allelic pairs}} \text{prior probability(pair } | H_d) \cdot \text{likelihood}(E | \text{pair})}
\]

Note that this profile supports the alternative hypothesis H_d.

Conclusion

When we analyze the DNA profiles separately; we see that especially the fourth profile has a very low likelihood ratio. It is more likely to observe the fourth profile when the donor is an unknown man than to observe it when the suspect is the donor. The other DNA profiles do have likelihoods that are greater than 1 and therefore it is more likely to observe the profile when the donor is the suspect.

5.10.1 Combined evidential value of the DNA profiles

We would like to give the combined evidential value of the four DNA profiles. We cannot simply take the product of the likelihood ratios for the individual profiles; (since the profiles all originated from the same stain, it is unlikely that the contamination on the profiles are independent of each other). However, when we deal with profiles where our model calculates tiny amounts of contamination, assuming independence
is an acceptable approximation. Another option is to use the sum of the peak heights on all alleles of different loci from different DNA profiles as input to our model. We will investigate this method.

5.10.2 Taking the sum peak heights of the individual DNA profiles as input

Here we will investigate the results when we take the sum of the individual DNA profiles as input. Since the profiles are made from the same stain, and under the assumption that the stain was left by one person (so the profiles are non-mixture profiles) we know that all profiles came from the same person. When profiles were made from different stains we would not be sure about this and it would be unwise to sum the different profiles in this model. The sum of the DNA profile on the first locus, D10, will give us the peak heights in table 5.10.7 on the different alleles.

<table>
<thead>
<tr>
<th>allele</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>peak height</td>
<td>20</td>
<td>20</td>
<td>35</td>
<td>302</td>
<td>26</td>
<td>48</td>
<td>472</td>
<td>19</td>
<td>15</td>
<td>20</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 5.10.7: The sum of the peak heights of the 4 DNA profiles on the locus D10

Under our standard settings, the expected peak heights on each locus are as in table 5.10.8.

<table>
<thead>
<tr>
<th>locus</th>
<th>D10S1248</th>
<th>vWA</th>
<th>D16S539</th>
<th>D2S1338</th>
<th>D8S1179</th>
<th>D21S11</th>
<th>D18S51</th>
<th>D22S1045</th>
</tr>
</thead>
<tbody>
<tr>
<td>peak height</td>
<td>491.5</td>
<td>640</td>
<td>337</td>
<td>640</td>
<td>542.5</td>
<td>524.5</td>
<td>610</td>
<td>529.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>locus</th>
<th>D19S433</th>
<th>TH01</th>
<th>FGA</th>
<th>D2S441</th>
<th>D3S1358</th>
<th>D1S1656</th>
<th>D12S391</th>
</tr>
</thead>
<tbody>
<tr>
<td>peak height</td>
<td>636.5</td>
<td>546</td>
<td>580.5</td>
<td>507</td>
<td>597</td>
<td>581</td>
<td>439.5</td>
</tr>
</tbody>
</table>

Table 5.10.8: The expected peak heights under our standard assumptions for the sum of 4 profiles

The results of this simulation (where we took \( M = 5000 \)), are presented in table 5.10.9.

<table>
<thead>
<tr>
<th>D10</th>
<th>vWA</th>
<th>D16</th>
<th>D2S</th>
<th>D8S</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>15-15</td>
<td>-37</td>
<td>14-14</td>
<td>-42</td>
<td>10-10</td>
</tr>
<tr>
<td>12-12</td>
<td>-43</td>
<td>16-16</td>
<td>-43</td>
<td>9-9</td>
</tr>
<tr>
<td>10-12</td>
<td>-52</td>
<td>15-16</td>
<td>-49</td>
<td>9-11</td>
</tr>
<tr>
<td>12-16</td>
<td>-54</td>
<td>14-17</td>
<td>-75</td>
<td>9-14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D21</th>
<th>D18</th>
<th>D22</th>
<th>D19</th>
<th>TH0</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>31-31</td>
<td>-70</td>
<td>12-15</td>
<td>-49</td>
<td>11-16</td>
</tr>
<tr>
<td>28-31</td>
<td>-70</td>
<td>15-15</td>
<td>-55</td>
<td>11-11</td>
</tr>
<tr>
<td>28-28</td>
<td>-78</td>
<td>12-12</td>
<td>-57</td>
<td>16-16</td>
</tr>
<tr>
<td>29-31</td>
<td>-79</td>
<td>11-15</td>
<td>-68</td>
<td>11-17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FGA</th>
<th>D2S</th>
<th>D3S</th>
<th>D1S</th>
<th>D12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>24-24</td>
<td>-72</td>
<td>14-14</td>
<td>-48</td>
<td>16-16</td>
</tr>
<tr>
<td>23-24</td>
<td>-95</td>
<td>15-13.3</td>
<td>-58</td>
<td>18-18</td>
</tr>
<tr>
<td>22-24</td>
<td>-97</td>
<td>14-13.3</td>
<td>-66</td>
<td>16-18</td>
</tr>
</tbody>
</table>

Table 5.10.9: The log_{10} of the likelihoods of observing the sum of the four DNA profile.
We see that on almost all loci the assumed allelic pair has the highest likelihood. On two loci, D12 and D21, the assumed allelic pair has the second highest likelihood. However, the difference between these pairs and the pair with the highest likelihood is quite small. If we take a closer look at the sum of the peak heights, we see that on D12 the peak height on allele 31 is 468 and that the peak height on allele 28 is only 275 rfu. There is a clear peak imbalance at this locus. The same occurs at D21, where the peak height on allele 19 is 403 rfu and the peak height on allele 18 is only 140 rfu. If we assume independence between loci, we can compute the likelihood ratio for the hypotheses

\[ H_p: \text{ the suspect is the donor of the DNA profile} \]
\[ H_d: \text{ some unknown man is the donor of the DNA profile}. \]

We get as our likelihood ratio

\[
\frac{\mathbb{P}(E|H_p)}{\mathbb{P}(E|H_d)} = \prod_{\text{loci}} \frac{\text{likelihood of } E \text{ on } \text{locus}, \text{ given donor is assumed allelic pair on } \text{locus}}{\sum_{\text{allelic pairs}} \text{prior probability}(\text{pair } | H_d) \cdot \text{likelihood}(E|\text{pair})} = 4.6481 \cdot 10^{18}
\]

This can be regarded as very strong evidence in favor of \( H_p \). The likelihood ratio is even bigger than the product of the likelihood ratio’s of the individual profiles. This can be explained easily, since ‘bad’ profiles won’t give any relevant information, the likelihood ratio will increase when we sum all profiles and obtain ‘good’ input, where the peak heights at the alleles of the donor distinct from the peak heights at the other alleles.

### 5.11 Conclusion and discussion

We presented a method to interpret non-mixture (low-template) DNA profiles. It can deal with stutters, contamination and drop-out. The method we created is meant to determine the likelihood of observing a DNA profile given an allelic combination on each locus. The sensitivity analysis showed that the choices for the distributions and parameters for the different stochastic variables leaves the order of likelihoods intact. Our method works by determining the maximum likelihood estimator for the contamination. Since the determination of this vector can be an expensive procedure, we proposed a method where the direction of the maximum likelihood estimator for the contamination where the variables \( \Lambda \) and \( \pi \) where equal to their expected values is used to find an approximation of the maximum likelihood estimator for the contamination when we draw random values for \( \Lambda \) and \( \pi \). If it is not necessary to do this, we would advise to compute the maximum likelihood estimator for each simulation to be more precise.

However, the likelihood ratios between pairs may differ when we change the distributions and parameters. Therefore, we advise to do more research on the distributions and parameters of the noise \( \epsilon \) and the amplification factor \( \lambda \). When combining results from different profiles we need to be careful. Simply multiplying the likelihoods for different profiles would mean that we assume independence of the contamination between the profiles. This is unrealistic. DNA samples from the same stain are obtained by the same person and are amplified in the same lab. A contamination in one DNA profile will make it more likely that there has been a contamination in one of the other DNA profiles. We suggested a method where you add the peak heights from different profiles and use that as input for our model. It is important to do this only on profiles of which you are reasonably sure that they were left by the same person (for instance DNA profiles made from the same stain).

A case example of four profiles from the same stain showed that the sum of these DNA profiles gave us a very discriminating likelihood ratio. The likelihood ratio stated that it is more likely to find the evidence when the suspect was the donor of the stain than when an unknown man was the donor.

Our method deviates from existing methods in that we make use of the peak heights and do not use thresholds for e.g. peak detection or estimation of drop-out probabilities.

Further research could concentrate on making the model able to work with mixtures as well as other methods to combine the output of the model for different DNA profiles into one likelihood ratio.
Chapter 6

Conclusions and Discussion

In this thesis, we concentrated on three problems that were all based on the combination of DNA profiles. We first developed a method that derives a lower bound for the likelihood ratio when we want to combine several Y-chromosomal DNA profiles. The lower bound doesn’t need the prior probabilities for the different hypotheses that are usually hard or impossible to determine. We expect that the method will only need small adaptations to make it applicable to other sorts of evidence.

We made a simulation model that investigates the dependence between autosomal and Y-chromosomal DNA profiles. All the model assumptions are made in such a way that they are not disadvantageous to the defending party. The results of the simulation model pointed towards independence between autosomal and Y-chromosomal DNA profiles when we neglected the people that are in the close family (people from the same grandfather as our suspect). Further research could concentrate on using the model on way larger/smaller populations. The simulations we did only used autosomal profiles on up to 5 loci. It is important to note that it won’t be necessary to use the model with autosomal profiles on too much loci, since the profile frequency of the autosomal profile will be discriminating enough.

In the last part of our thesis we presented a model to determine the likelihood of a DNA profile given any combination of alleles. The model uses simulation to find an approximation of this likelihood. We assumed some parameters and distributions for the stochastic variables in the model, but the model will be able to work with any other distributions and parameters. We advise to do further research on the distributions/parameters of the amplification factor $\Lambda$ and the noise $\epsilon$. We proposed a method to combining several low template DNA profiles that were made from the same stain that is based on summing the peak height information of the different profiles and use the outcome as new input of our model. Further research could concentrate on finding alternative methods of combining low template DNA profiles using the described method.
Appendix A

DNA

In this chapter we will give background information regarding DNA, which will help the reader understand why certain model assumptions were made or what the different terms mean. We first discuss some of the important terms and give a glossary afterwards. Some of the other chapters will give some additional DNA-based information that is only needed to understand the decisions and assumptions that are made in that particular chapter.

A.1 Introduction

DNA (Deoxyribonucleic acid) is the genetic code that is contained in the nuclei of all human cells. In each nuclei the DNA is distributed over 46 chromosomes, that occur in pairs. From each pair of chromosomes, one is hereditary from your father, the other from your mother. These 23 chromosomal pairs can be divided into two groups; there are 22 ‘autosomal’ pairs and one pair of sex chromosomes.

The sex chromosomes are notated with either two X’s or one X and one Y, where two XX represents a woman, and XY represents a man. The remaining autosomal DNA doesn’t contain information on whether the source of the DNA is a man or a woman.

Only 2% of the DNA provides for the genetic characteristics. The other 98% are non-coding repetitive sequences, in other words, they do not contribute in how you look and do not say anything about how you are in general. These hypervariabel area’s contain repetitive parts of small parts of DNA code, (ATCG-ATCG-ATCG-ATCG-. . . ) For unrelated people we can be almost certain that they have a different number of repeating units. These features make them very attractive in matching DNA from a person to a trace. These repetitive parts are called ‘STRs’, Short Tandem Repeats.

Short Tandem Repeats

A short tandem repeat in DNA occurs when a pattern of two or more nucleotides are repeated and the repeated sequences are directly adjacent to each other. The pattern can range in length from 2 to 5, for instance ATCG is of length 4. The location of the hypervariable area’s where these short tandem repeats occurs is called the locus (pl. loci).

Loci

A locus is the specific location of a gene or DNA sequence on a chromosome. A variant of the DNA sequence at a given locus is called an allele.

Alleles

Each hypervariable area is characterized by the number of times that the repetitive part is occurring. This is called the DNA-characteristic or allele (pl. alleles). The allele is represented with a number, the number of times that the repetitive part is occurring. For example, when we have ATCG-ATCG-ATCG-ATCG, we would represent it with the number 4.
The appearance of a locus is represented by the alleles of the corresponding hypervariable areas on the DNA molecules of the chromosomal pair. This is indicated with two (different or equal) numbers, for example (6/8, 7/7, 9/10).

A.2 DNA profiles

DNA profiles are a representation of some of alleles on some loci. Their are several different DNA profiles.

A.2.1 Autosomal DNA profiles

An autosomal DNA profile is a DNA profile over all the chromosomes. Currently, a complete DNA profile contains information on which alleles a person carries on 15 loci, as well as information on whether the donor is a man or a woman. An autosomal DNA profile of a person could be represented with fifteen pairs of numbers and two X's (when the source is a woman) or an X and a Y (when the source is a man). The fifteen number pairs represent the alleles on the 15 loci.

A.2.2 Y-chromosomal DNA profiles

A Y-chromosomal DNA profile is a profile only on the Y-chromosome. Therefore, it is only possible to obtain a Y-chromosomal DNA profile from a man.

A.3 Heritage

The DNA profile of each person is a result of the combination of the DNA profiles of his/her parents. On each locus, one allele is heritated from your father, and one is heritated
### Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>allele</td>
<td>This is a number that represent the number of repeats of some small part of code on a certain locus.</td>
</tr>
<tr>
<td>amplification cycles</td>
<td>An amplification cycle is a procedure to double the amount of DNA obtained from a stain.</td>
</tr>
<tr>
<td>artifact</td>
<td>This is the collective term for stutters, drop-ins and drop-outs.</td>
</tr>
<tr>
<td>autosomal DNA profile</td>
<td>A DNA profile on 15 loci from different chromosomes and the information on the sex chromosome.</td>
</tr>
<tr>
<td>DNA-mixture</td>
<td>This is a DNA profile of a stain where there were multiple donors. Therefore it is possible to see more than 2 peaks in the DNA profile.</td>
</tr>
<tr>
<td>DNA-profile</td>
<td>A representation of the alleles on the loci from a certain donor or stain.</td>
</tr>
<tr>
<td>drop-in</td>
<td>An allele that has a peak height on the DNA profile that comes over a certain threshold, although the actual allele doesn’t appear in the actual profile, since the donor doesn’t bear this allele, i.e. an allele that should not be in the DNA profile that is in the DNA profile.</td>
</tr>
<tr>
<td>drop-out</td>
<td>An allele has been dropped out when we the peak height at that allele is below a certain threshold, but where the donor does bear that allele, i.e. it should be in the DNA profile, but it is not.</td>
</tr>
<tr>
<td>heterozygote</td>
<td>Two different alleles on a locus.</td>
</tr>
<tr>
<td>homozygote</td>
<td>Two equal alleles on a locus.</td>
</tr>
<tr>
<td>locus</td>
<td>One of the locations of the DNA of which the alleles are checked.</td>
</tr>
<tr>
<td>locus drop-out</td>
<td>If both alleles on a locus drop-out, we speak of a locus drop-out.</td>
</tr>
<tr>
<td>low template DNA profiles</td>
<td>A DNA profile that was made from a very low quantity of DNA.</td>
</tr>
<tr>
<td>PCR</td>
<td>This is the process of amplifying the DNA. The PCR, or polymerase chain reaction can be regarded as a molecular photocopier.</td>
</tr>
<tr>
<td>STR</td>
<td>Short Tandem Repeats, regions with short repeat units (usually 2-6 base pairs in length) are called Short Tandem Repeats (STR). These regions are used to make the DNA profile.</td>
</tr>
<tr>
<td>stutter</td>
<td>A peak on a DNA profile just before a peak that belongs to an allele.</td>
</tr>
<tr>
<td>Y-chromosomal DNA profile</td>
<td>A DNA profile only on the Y-chromosome.</td>
</tr>
<tr>
<td>Y-chromosome</td>
<td>One of the two possible sex-chromosomes. Only men have an Y-chromosome. Women have two X-chromosomes.</td>
</tr>
</tbody>
</table>
Appendix B

Additional results

B.1 Possible alleles on the different loci

Here we will give a table containing all the possible alleles on the different loci, that a person can have. This table is not necessarily a representation of all the alleles that exist in real life, it is a representation of the alleles that were observed in a sample from the Dutch population. Therefore this table can be regarded as the alleles we used in our simulations. Notice that on some alleles there are several possible ‘sorts’ of alleles. A ‘standard’ allele is a sequence that consists of small repetitive parts with ‘length’ 4 (ATCG-ATCG...+), however, it is also possible that these repetitive parts are of length 1, 2 or 3. These alleles are represented with the number of times the part is repeated and the length of the part. For example, 14.2, means that it is an allele where the repetitive part is repeated 14 times and is of length 2. The ‘standard’ alleles of length 4 could be represented by $x.4$, where $x$ is the number of times the part is repeated, but it is usually represented by just $x$. If on a locus the possible alleles are 10, 11, 12, 13, 14, 15, 16, 17, 18 and 19, we will write $10 \rightarrow 19$.

<table>
<thead>
<tr>
<th>locus</th>
<th>x.1</th>
<th>x.2</th>
<th>x.3</th>
<th>x.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D10S1248</td>
<td></td>
<td></td>
<td>10→19</td>
<td></td>
</tr>
<tr>
<td>D16S539</td>
<td></td>
<td></td>
<td>7→16</td>
<td></td>
</tr>
<tr>
<td>D2S1338</td>
<td></td>
<td>12→28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D8S1179</td>
<td>12→8</td>
<td>8→18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D21S11</td>
<td>26→35</td>
<td>24→35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D18S51</td>
<td></td>
<td>9→24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D22S1045</td>
<td></td>
<td>10→19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D19S433</td>
<td>12→19</td>
<td>10→18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TH01</td>
<td>8→9</td>
<td>5→10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGA</td>
<td>19→24</td>
<td>17→28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2S441</td>
<td>11→13</td>
<td>9→16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3S1358</td>
<td></td>
<td>11→20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D18S1656</td>
<td>14→20</td>
<td>10→18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D12</td>
<td>17→20</td>
<td>14→28</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table B.1.1: The possible alleles on the different loci
B.2 Results for the combination of autosomal and Y-chromosomal DNA profiles

B.2.1 A profile on four loci

In this section we present the results we got by simulating with a suspect profile on 4 loci. The suspect profile we used is given in table B.2.1.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>D22S1045</td>
<td>15</td>
</tr>
<tr>
<td>D10S1248</td>
<td>16</td>
</tr>
<tr>
<td>D3S1358</td>
<td>19</td>
</tr>
<tr>
<td>VWA</td>
<td>16</td>
</tr>
</tbody>
</table>

Table B.2.1: The profile on 4 loci assigned to our suspect

The random match probability of this profile is 5.1814 \cdot 10^{-6}.

We first look at the results for a tree that relates 1000 living men. A tree which relates 100 living men will show almost no matching profiles because the profile probability is very low.

Family trees relating 1000 living men

We did 100 profile simulations over 50 trees linking 1000 men. The approximate probabilities of the number of matching profiles for all the different trees are given in figure B.2.1.

Figure B.2.1: The estimated probability for the number of matching profiles for 20 trees without close family

From the figure it is clear that is almost never happened that there was a living person in the family tree (not from the close family) that had the same autosomal DNA-profile as our suspect. This will be the same when we look at the approximate probabilities over all the trees (table B.2.2). In the table we also give the probabilities for a random unrelated population.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Max</th>
<th>Min</th>
<th>random unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(no matching profiles)</td>
<td>0.9942</td>
<td>1</td>
<td>0.97</td>
<td>0.9949</td>
</tr>
<tr>
<td>P(1 matching profile)</td>
<td>0.0058</td>
<td>0.03</td>
<td>0</td>
<td>0.0051</td>
</tr>
</tbody>
</table>

Table B.2.2: The mean, maximum and minimum computed over 50 trees
The difference with a profile of 3 loci is that we had less simulations that showed matching profiles. This is what we would expect, since the number matching profiles should be lower when the profile frequency is lower. Adding another locus to a profile lowers the profile frequency. On the other hand, if we compare the mean frequency over all the trees with the probability we expect in a random unrelated population we see almost no difference. This is a satisfying result.

**Family trees relating 10 000 living men**

We also did the profile simulation over 20 family trees relating 10 000 living men. The profile simulation was done 100 times on each tree. The results can be found in figure B.2.2.

![Figure B.2.2: The estimated probability for the number of matching profiles for 10 trees without close relatives](image)

And once again, the mean, maximum and minimum can be found in the following table (table B.2.3).

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Max</th>
<th>Min</th>
<th>random unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(no matching profiles)</td>
<td>0.9609</td>
<td>0.98</td>
<td>0.93</td>
<td>0.9495</td>
</tr>
<tr>
<td>P(1 matching profile)</td>
<td>0.0355</td>
<td>0.06</td>
<td>0.02</td>
<td>0.0492</td>
</tr>
<tr>
<td>P(2 matching profiles)</td>
<td>0.0036</td>
<td>0.02</td>
<td>0</td>
<td>0.0013</td>
</tr>
</tbody>
</table>

Table B.2.3: The mean, maximum and minimum computed over 20 trees

The difference with the trees containing 1000 men is that there were more simulations that gave a positive number of matching profiles. This is what we would expect, since more people should increase the probability of a match. Just as in the simulation with 1000 living men, the number of matching profiles is almost always 0. The difference between the frequency over the family trees and the probability we would expect in a random unrelated population is similar, the probability is even slightly higher for a random unrelated population.
B.2.2 A profile on five loci

The last profile length we used to examine the dependence was a profile on 5 loci, with profile probability $1.5722 \cdot 10^{-5}$, as in table B.2.4

<table>
<thead>
<tr>
<th>Locus</th>
<th>Allele 1</th>
<th>Allele 2</th>
<th>Allele 3</th>
<th>Allele 4</th>
<th>Allele 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>D22S1045</td>
<td>15</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D10S1248</td>
<td>13</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3S1358</td>
<td>17</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VWA</td>
<td>17</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D16S539</td>
<td>11</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table B.2.4: The profile on 5 loci assigned to our suspect

We only did the 100 profile simulations on 20 trees that linked 10 000 living men. The results are in the following figure (B.2.3) and table (B.2.5)

Figure B.2.3: The estimated probability for the number of matching profiles for 20 trees without close relatives

<table>
<thead>
<tr>
<th>Number of Matching Profiles</th>
<th>Mean</th>
<th>Max</th>
<th>Min</th>
<th>Random Unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.8535</td>
<td>0.92</td>
<td>0.78</td>
<td>0.8546</td>
</tr>
<tr>
<td>1</td>
<td>0.1265</td>
<td>0.19</td>
<td>0.06</td>
<td>0.1343</td>
</tr>
<tr>
<td>2</td>
<td>0.0185</td>
<td>0.03</td>
<td>0</td>
<td>0.0106</td>
</tr>
<tr>
<td>3</td>
<td>0.0015</td>
<td>0.01</td>
<td>0</td>
<td>0.0006</td>
</tr>
</tbody>
</table>

Table B.2.5: The mean, maximum and minimum computed over 50 trees

We see that the expected number of matching profiles is bigger than with the profile on 4 loci. This can be explained by the higher profile probability. The profile that we used on 4 loci had a profile probability of $5.1825 \cdot 10^{-6}$, and the profile probability of this profile is only $1.5728 \cdot 10^{-5}$. Similar as in the previous results we have that the majority of the results show that there were no matching profiles in the family tree without the close relatives. Once again, we see that the difference between the frequencies over the trees and the probabilities for a random unrelated population are approximately the same.
B.3 Sensitivity analysis of a ‘hard’ profile

The ‘easy’ profile on which we did a sensitivity analysis had clear peaks on each locus. Therefore, we also performed a sensitivity analysis on the ‘hard’ partial profile given in Table B.3.1. This profile is a selection of some of the loci of the ‘first stain DNA profile’ which can be found in Table C.1.2. We choose this loci because there was either no clear highest peak (D16,D21) or a big peak imbalance (TH0,FGA,D12)

<table>
<thead>
<tr>
<th>allele 1</th>
<th>D16</th>
<th>D21</th>
<th>TH0</th>
<th>FGA</th>
<th>D12</th>
</tr>
</thead>
<tbody>
<tr>
<td>height 1</td>
<td>7</td>
<td>24.2</td>
<td>5</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>height 2</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>allele 2</td>
<td>8</td>
<td>25</td>
<td>6</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>height 2</td>
<td>5</td>
<td>5</td>
<td>12</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>allele 3</td>
<td>9</td>
<td>26.2</td>
<td>7</td>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td>height 3</td>
<td>7</td>
<td>4</td>
<td>103</td>
<td>7</td>
<td>45</td>
</tr>
<tr>
<td>allele 4</td>
<td>10</td>
<td>27</td>
<td>8</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>height 4</td>
<td>22</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>93</td>
</tr>
<tr>
<td>allele 5</td>
<td>13</td>
<td>28</td>
<td>9.3</td>
<td>22</td>
<td>19.3</td>
</tr>
<tr>
<td>height 5</td>
<td>4</td>
<td>56</td>
<td>238</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>allele 6</td>
<td>14</td>
<td>29</td>
<td>23</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>height 6</td>
<td>3</td>
<td>5</td>
<td>19</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>allele 7</td>
<td>30</td>
<td>24</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>height 7</td>
<td>6</td>
<td>142</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allele 8</td>
<td>30.2</td>
<td>25</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>height 8</td>
<td>7</td>
<td>11</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allele 9</td>
<td>31</td>
<td>26</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>height 9</td>
<td>6</td>
<td>37</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allele 10</td>
<td>31.2</td>
<td>27</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>height 10</td>
<td>4</td>
<td>9</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>allele 11</td>
<td>32</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>height 11</td>
<td>6</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>allele 12</td>
<td>32.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>height 12</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>allele 13</td>
<td>35.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>height 13</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table B.3.1: DNA profile on 5 loci for our sensitivity analysis

Just as before, we will do one simulation under our standard settings (see Table B.3.2). Selecting the peak height at each locus results in the following expected peak heights

<table>
<thead>
<tr>
<th>M</th>
<th>stutter distribution</th>
<th>amplification distribution</th>
<th>standard deviation $\epsilon$</th>
<th>standard deviation $\Lambda$</th>
<th>Expected peak height method</th>
<th>Contamination distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>lognormal</td>
<td>normal</td>
<td>10</td>
<td>0.2</td>
<td>per locus</td>
<td>original</td>
</tr>
</tbody>
</table>

Table B.3.2: Settings

During the sensitivity analysis, we will use the same settings as in this simulation except from the one or two parameters we vary. The settings are as in Table B.3.2 except for those that we say we change for that simulation. The results of our simulation under the standard settings are given in table B.3.3.
This will be our reference simulation. We won’t assume some allele pairs to be the donor of the profile, however we will give extra attention to the top 2 pairs on all the loci. We only do this to be able to compare the results more easily.

Universal peak height

The following simulation was done with a universal expected peak height. On all the loci we used the same peak heights, and we choose this to be the sum of all peak heights in our DNA profile divided by 2 times the number of loci. The expected peak height was 105.1. Since the peak heights at some loci are nowhere near 105.1 we expect the results on these loci to be almost worthless. We expect the allele combinations on these loci (D16 D21) to be much lower. The results of this simulation are in table B.3.4.

Standard deviation of the noise

In this simulation we changed the standard deviation of the noise from 10 to 5. Since this will only change the distribution in such a way that high amounts of noise become less likely, we expect to find roughly the same results as in our reference simulation but with lower values for the likelihood. The results of the simulation are in table B.3.5.

Table B.3.3: The \( \log_{10} \) of the likelihoods of observing the DNA profile given different allele pairs under our standard settings.

<table>
<thead>
<tr>
<th>D16</th>
<th>D21</th>
<th>TH0</th>
<th>FGA</th>
<th>D12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>9-10</td>
<td>-20</td>
<td>28-32.2</td>
<td>-54</td>
<td>9.3-9.3</td>
</tr>
<tr>
<td>8-10</td>
<td>-22</td>
<td>28-31.2</td>
<td>-55</td>
<td>9.3-9.3</td>
</tr>
<tr>
<td>10-14</td>
<td>-22</td>
<td>28-34.2</td>
<td>-55</td>
<td>7-7</td>
</tr>
</tbody>
</table>

Table B.3.4: The \( \log_{10} \) of the likelihoods of observing the DNA profile given different allele pairs where the expected peak height method was universal.

<table>
<thead>
<tr>
<th>D16</th>
<th>D21</th>
<th>TH0</th>
<th>FGA</th>
<th>D12</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
<td>likelihood</td>
<td>pair</td>
</tr>
<tr>
<td>9-10</td>
<td>-34</td>
<td>28-31.2</td>
<td>-60</td>
<td>7-9.3</td>
</tr>
<tr>
<td>8-10</td>
<td>-37</td>
<td>28-24.2</td>
<td>-62</td>
<td>6-7</td>
</tr>
</tbody>
</table>

Table B.3.5: The \( \log_{10} \) of the likelihoods of observing the DNA profile given different allele pairs where the standard deviation of the noise was 5.
Normal stutter and uniform amplification factor

Here we changed the distribution of the stutters from normal to lognormal and the distribution for the amplification factor from normal to uniform. This gives us the results given in table B.3.6.

<table>
<thead>
<tr>
<th></th>
<th>D16</th>
<th></th>
<th>D21</th>
<th></th>
<th>TH0</th>
<th></th>
<th>FGA</th>
<th></th>
<th>D12</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>pair</td>
<td></td>
<td>pair</td>
<td></td>
<td>likelihood</td>
<td></td>
<td>pair</td>
<td></td>
<td>likelihood</td>
<td></td>
<td>pair</td>
</tr>
<tr>
<td>9-10</td>
<td>−21</td>
<td>28-30</td>
<td>−54</td>
<td>7-9.3</td>
<td>−28</td>
<td>24-26</td>
<td>−54</td>
<td>18-19</td>
<td>−53</td>
<td></td>
</tr>
<tr>
<td>7-10</td>
<td>−23</td>
<td>28-31</td>
<td>−55</td>
<td>9.3-9.3</td>
<td>−29</td>
<td>24-24</td>
<td>−60</td>
<td>19-26</td>
<td>−59</td>
<td></td>
</tr>
<tr>
<td>10-10</td>
<td>−23</td>
<td>28-35.2</td>
<td>−56</td>
<td>6-9.3</td>
<td>−66</td>
<td>24-28</td>
<td>−67</td>
<td>19-25</td>
<td>−61</td>
<td></td>
</tr>
<tr>
<td>10-12</td>
<td>−23</td>
<td>28-24.2</td>
<td>−56</td>
<td>8-9.3</td>
<td>−69</td>
<td>17-24</td>
<td>−69</td>
<td>19-19.3</td>
<td>−61</td>
<td></td>
</tr>
</tbody>
</table>

Table B.3.6: The log$_{10}$ of the likelihoods of observing the DNA profile given different allele pairs where the standard deviation of the noise was 5.

Conclusion

The results that we found in this sensitivity analysis are more distinct than the results in the sensitivity analysis of the ‘easy’ profile. On the loci where there was a peak imbalance, the method still gives the same allelic pairs show up in the top 5 of highest likelihoods. The loci where there were no alleles with outstanding peaks the top 5 of highest likelihoods differ between the different simulations very often. However, the relative difference between their likelihoods is small when we compare them with the relative differences between likelihoods of allelic pairs on loci with a peak imbalance. On most loci, the change in the likelihood when we change the parameters act like we expect them to do.
Appendix C

Low template DNA profiles

Here we give the low template DNA profiles that we used in our case from section 5.10.
C.1 DNA profiles for case example

C.1.1 Reference profile

| Locus | Allele 1 | Height 1 | Allele 2 | Height 2 | Allele 3 | Height 3 | Allele 4 | Height 4 | Allele 5 | Height 5 | Allele 6 | Height 6 | Allele 7 | Height 7 | Allele 8 | Height 8 | Allele 9 | Height 9 | Allele 10 | Height 10 | Allele 11 | Height 11 | Allele 12 | Height 12 | Allele 13 | Height 13 | Allele 14 | Height 14 | Allele 15 | Height 15 |
|-------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| D10   | 10       | 11       | 11       | 158      | 12       | 202      | 13       | 21       | 14       | 192      | 15       | 223      | 16       | 7        | 17       | 12       | 18       | 4        | 19       | 4        |
| vWA   | 10       | 21       | 12       | 8        | 13       | 79       | 14       | 243      | 15       | 100      | 16       | 165      | 17       | 8        | 18       | 7        | 19       | 8        | 20       | 4        |
| D16   | 7        | 7        | 8        | 76       | 9        | 200      | 10       | 224      | 11       | 15       | 12       | 4        | 13       | 2        | 14       | 4        |
| D2S   | 14       | 9        | 15       | 4        | 16       | 12       | 17       | 23       | 18       | 15       | 19       | 160      | 20       | 191      | 21       | 12       | 22       | 147      | 23       | 176      | 24       | 6        | 25       | 6        | 27       | 44       | 28       | 4        |
| D8S   | 8        | 17       | 9        | 22       | 10       | 20       | 11       | 152      | 12       | 225      | 13       | 192      | 14       | 58       | 15       | 7        | 16       | 7        | 17       | 7        |
| D21   | 24, 21   | 25       | 6        | 25, 4    | 15       | 57       | 28       | 145      | 28, 4    | 9        | 29       | 31       | 29        | 4        | 30       | 187      | 30        | 5        | 31       | 241      | 31, 5    | 32, 30    | 34, 6    | 4        |
| D18   | 8        | 17       | 9        | 4        | 11       | 94       | 12       | 157      | 13       | 15        | 14       | 145      | 15       | 175      | 16       | 23       | 17       | 6        | 19       | 9        | 26       | 5        |
| D22   | 14       | 19       | 10       | 49       | 11       | 156      | 12       | 12       | 13       | 20       | 14       | 24       | 15       | 123      | 16       | 122      | 17       | 63       | 18       | 10       |
| D19   | 9        | 9        | 10       | 13       | 11       | 13       | 12       | 13       | 13       | 167      | 14       | 217      | 15       | 179      | 16       | 10       | 16, 2     | 13       | 17       | 6        | 17, 2     | 12       |
| TH0   | 4        | 12       | 5        | 14       | 6        | 31       | 6, 3     | 31       | 7        | 189      | 8        | 9        | 8, 3      | 29       | 9        | 65       | 9, 3      | 206      | 11       | 13       | 12       | 61, 18     | 13, 5    | 5        |
| FGA   | 17       | 11       | 18       | 14       | 20       | 34       | 20, 2    | 57       | 22       | 13       | 23       | 14        | 24       | 128      | 25       | 117      | 26       | 121      | 27       | 16       | 28       | 9        |
| D2S   | 9        | 12       | 10       | 9        | 11       | 14       | 11       | 13       | 90       | 14       | 183      | 15       | 160      | 16       | 29       |
| D5S   | 12       | 21       | 13       | 38       | 14       | 184      | 15       | 196      | 15, 4    | 28       | 16       | 197      | 17        | 10       | 18       | 46       | 19       | 16       |
| D5S   | 9        | 12       | 10       | 11       | 11       | 13       | 12       | 109      | 13       | 166      | 14       | 18       | 15       | 206      | 16       | 175      | 17        | 24       | 18       | 20       | 18, 16     | 19, 7     | 28, 10    |
| D12   | 14       | 12       | 15       | 25       | 16       | 21       | 17       | 146      | 18       | 191      | 19       | 113      | 20       | 9        | 21       | 15       | 22       | 11       | 23       | 11       | 26       | 10       | 27       | 10       |

Table C.1.1: Reference DNA profile for the case example

C.1.2 Stain profiles 1-4
<table>
<thead>
<tr>
<th>Sample</th>
<th>vWA</th>
<th>D16</th>
<th>D18</th>
<th>D21</th>
<th>D19</th>
<th>TH0</th>
<th>FGA</th>
<th>D2S</th>
<th>D3S</th>
<th>D1S</th>
<th>D12</th>
<th>D13</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table C.1.2: DNA profile 1 and 2 for the case example
| locus | allele 1 | height 1 | allele 2 | height 2 | allele 3 | height 3 | allele 4 | height 4 | allele 5 | height 5 | allele 6 | height 6 | allele 7 | height 7 | allele 8 | height 8 | allele 9 | height 9 | allele 10 | height 10 | allele 11 | height 11 | allele 12 | height 12 | allele 13 | height 13 | allele 14 | height 14 |
|-------|---------|----------|---------|----------|---------|----------|---------|----------|---------|----------|---------|----------|---------|----------|---------|---------|---------|----------|----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| D2S010 | 9 | 4 | 10 | 6 | 11 | 10 | 12 | 20 | 13 | 6 | 14 | 7 | 15 | 92 | 16 | 3 | 17 | 4 | 18 | 5 |
| vWA | 10 | 9 | 11 | 6 | 12 | 5 | 13 | 9 | 14 | 7 | 15 | 24 | 9 | 15 | 13 | 16 | 23 | 18 | 4 | 19 | 7 | 20 | 5 |
| D16 | 9 | 6 | 10 | 26 | 11 | 7 | 12 | 6 | 13 | 6 |
| D2S15 | 15 | 3 | 16 | 4 | 17 | 8 | 18 | 5 | 19 | 15 | 20 | 15 | 24 | 4 | 22 | 16 | 23 | 27 | 27 | 28 | 4 |
| D8S8 | 8 | 7 | 10 | 7 | 11 | 11 | 12 | 99 | 13 | 80 | 14 | 32 | 17 | 7 |
| D21 | 24 | 6 | 25 | 6 | 26 | 6 | 28 | 6 | 29 | 6 | 30 | 8 | 30 | 2 | 7 | 31 | 94 | 31 | 7 | 32 | 8 | 33 | 6 | 33 | 2 | 34 | 6 | 35 | 5 |
| D18 | 9 | 9 | 10 | 4 | 11 | 9 | 12 | 96 | 13 | 8 | 14 | 13 | 15 | 17 | 17 | 5 | 18 | 6 | 19 | 7 | 21 | 7 | 23 | 8 | 24 | 7 |
| D22 | 9 | 15 | 10 | 8 | 11 | 46 | 12 | 12 | 13 | 13 | 14 | 9 | 15 | 18 | 16 | 16 | 18 | 7 | 19 | 16 |
| D19 | 11 | 5 | 12 | 13 | 12 | 7 | 13 | 22 | 14 | 17 | 15 | 109 | 15 | 10 | 17 | 14 | 17 | 2 | 10 |
| TH0 | 4 | 16 | 6 | 11 | 7 | 112 | 9,3 | 169 |
| FGA | 17 | 9 | 18 | 9 | 19 | 9 | 20 | 9 | 20 | 2 | 11 | 21 | 13 | 21 | 12 | 22 | 13 | 23 | 21 | 24 | 119 | 25 | 33 | 26 | 37 | 27 | 9 | 28 | 12 |
| D9S9 | 9 | 11 | 10 | 10 | 11 | 11 | 14 | 17 | 33 | 39 | 14 | 48 | 15 | 67 | 16 | 9 |
| D3S13 | 13 | 51 | 14 | 9 | 15 | 35 | 16 | 65 | 18 | 51 | 19 | 7 |
| D1S9 | 9 | 12 | 10 | 15 | 11 | 9 | 12 | 21 | 13 | 166 | 14 | 10 | 14 | 19 | 15 | 21 | 16 | 46 | 16 | 14 | 17 | 3 | 13 | 18 | 11 | 19 | 5 | 6 | 20 | 3 | 17 |
| D12 | 15 | 8 | 16 | 8 | 17 | 4 | 18 | 17 | 19 | 82 | 19 | 8 | 20 | 10 | 23 | 9 | 27 | 14 |

Table C.1.3: DNA profile 3 and 4 for the case example
Bibliography

   The Sign of Four
   *Spencer Blackett*
   February 1890.

[2] Corina Benschop
   On the effects of sampling, analysis and interpretation strategies for complex forensic DNA research
   - with focus on sexual assault cases, PhD thesis
   *University of Amsterdam*
   ISBN 978-94-91407-03-1

[3] Roberto Puch-Solis, Lauren Rodgers, Anjali Mazumder, Susan Pope, Ian Evett, James Curran and David Balding
   Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters
   *Biometrics*
   Submitted in 2011.

   Least Square Deconvolution: A Framework for Interpreting Short Tandem Repeat Mixtures
   *Journal of Forensic Science*
   2006; Vol. 51, No. 6.

[5] Perlin MW, Szabady B.
   Linear mixture analysis: a mathematical approach to resolving mixed DNA samples
   *Journal of Forensic Science*
   2001; 46(6) 1372-1378.

   Probabilistic expert systems for handling artifacts in complex DNA mixtures.
   *Forensic Science International: Genetics*
   2011; No. 5, 202-209.

   Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out.
   *Forensic Science International: Genetics*
   2012; No. 4, 97-101.

[8] Balding DJ, Buckleton J
   Interpreting low template DNA profiles
   *Forensic Science International: Genetics*
   2009; No. 4, 1-10.

   Analysis of complex DNA mixtures using the Forensim package
   *Forensic Science International: Genetics*
   Suppl. (2011)


[12] Balding, D.J. and Nichols, R.A.
DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands
*Forensic Science International*
1994; 64: 125-140.

A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity
*Genetica*
1995; 96: 3-12.

Forensic DNA Evidence Interpretation
*CRC Press*
2005;

It was one of my brothers
*Int J Leg Med*
2005;

[16] Taroni F., Aitken C., Garbolino P., Biedermann A.
Bayesian Networks and Probabilistic Inference in Forensic Science
*John Wiley & Sons Ltd*
2006; 150-152

[17] Taroni F., Aitken C.
Statistics and the Evaluation of Evidence for Forensic Scientists
*John Wiley & Sons Ltd*
2004; 248-253

[18] Taroni F., Aitken C., Garbolino P., Biedermann A.
Bayesian Networks and Probabilistic Inference in Forensic Science
*John Wiley & Sons Ltd*
2006; 147

[19] Evett I.W., Jackson G., Lindly D.V., Meuwly D.
Logical evaluation of evidence when a person is suspected of committing two separate offenses
*Science & justice Volume 46 No.1*
2006; 25-31
