Rarities of genotype profiles in a normal Swedish population

Ronny Hedell

LiTH - MAT - EX - - 2010 / 25 - - SE
Rarities of genotype profiles in a normal Swedish population

Department of Mathematics, Linköpings Universitet

Ronny Hedell

LiTH - MAT - EX - - 2010 / 25 - - SE

Examensarbete: 30 hp

Level: D

Supervisors: Anders Nordgaard,
Department of Computer and Information Science, Linköpings Universitet
Statens Kriminaltekniska Laboratorium
Ricky Ansell,
Department of Physics, Chemistry and Biology, Linköpings Universitet
Statens Kriminaltekniska Laboratorium, The Biology Unit

Examiner: Torkel Erhardsson,
Department of Mathematics, Linköpings Universitet

Linköping: September 2010
Rarities of genotype profiles in a normal Swedish population

Investigation of stains from crime scenes are commonly used in the search for criminals. At The National Laboratory of Forensic Science, where these stains are examined, a number of questions of theoretical and practical interest regarding the databases of DNA profiles and the strength of DNA evidence against a suspect in a trial are not fully investigated. The first part of this thesis deals with how a sample of DNA profiles from a population is used in the process of estimating the strength of DNA evidence in a trial, taking population genetic factors into account. We then consider how to combine hypotheses regarding the relationship between a suspect and other possible donors of the stain from the crime scene by two applications of Bayes’ theorem. After that we assess the DNA profiles that minimize the strength of DNA evidence against a suspect, and investigate how the strength is affected by sampling error using the bootstrap method and a Bayesian method. In the last part of the thesis we examine discrepancies between different databases of DNA profiles by both descriptive and inferential statistics, including likelihood ratio tests and Bayes factor tests. Little evidence of major differences is found.

DNA profiles, Likelihood ratio, Multiple hypotheses, Minimum likelihood ratio, Database discrepancies.
Abstract

Investigation of stains from crime scenes are commonly used in the search for criminals. At The National Laboratory of Forensic Science, where these stains are examined, a number of questions of theoretical and practical interest regarding the databases of DNA profiles and the strength of DNA evidence against a suspect in a trial are not fully investigated. The first part of this thesis deals with how a sample of DNA profiles from a population is used in the process of estimating the strength of DNA evidence in a trial, taking population genetic factors into account. We then consider how to combine hypotheses regarding the relationship between a suspect and other possible donors of the stain from the crime scene by two applications of Bayes’ theorem. After that we assess the DNA profiles that minimize the strength of DNA evidence against a suspect, and investigate how the strength is affected by sampling error using the bootstrap method and a Bayesian method. In the last part of the thesis we examine discrepancies between different databases of DNA profiles by both descriptive and inferential statistics, including likelihood ratio tests and Bayes factor tests. Little evidence of major differences is found.

Keywords: DNA profiles, Likelihood ratio, Multiple hypotheses, Minimum likelihood ratio, Database discrepancies.
Acknowledgements

Many thanks to my supervisors Anders Nordgaard and Ricky Ansell for all your help and for introducing me to the exciting area of forensic science and forensic statistics. Thanks to my examiner Torkel Erhardsson for many suggestions that have improved the report. Also thanks to my opponent Martin Fagerlund for reading the report and giving ideas of improvement. The staff at SKL also deserves thanks for being so kind and polite. Finally big thanks to my family and to Margje for all your support.
Nomenclature

Abbreviations

DNA  Deoxyribonucleic acid
LR   Likelihood ratio
mpmp most probable matching profile
SKL  Statens kriminaltekniska laboratorium (The National Laboratory of Forensic Science)
Contents

1 Introduction 1
   1.1 Background ................................. 1
   1.2 Chapter outline ............................ 2

2 DNA biology 3

3 Population genetics 7
   3.1 Population allele proportions .................... 7
   3.1.1 Bayes estimators ............................. 11
   3.2 Hardy-Weinberg’s law and linkage equilibrium ........ 11
   3.2.1 Test of Hardy-Weinberg’s law and linkage equilibrium .. 12
   3.3 Distribution of profile proportions .................. 16
   3.4 Inbreeding coefficients ........................... 17
   3.5 Summary ...................................... 20

4 Likelihood ratios and match probabilities 21
   4.1 Framework ..................................... 21
   4.1.1 Bayes factors ................................. 22
   4.2 General formula .................................. 23
   4.3 SKL scale of conclusions .......................... 25
   4.4 Combining hypotheses ............................. 26
   4.5 Summary ...................................... 32

5 Most probable matching profile 33
   5.1 Tool for Excel .................................. 33
   5.1.1 Full siblings and the highest level of conclusion .. 35
   5.2 Assessing sampling error .......................... 35
   5.3 Summary ...................................... 39

6 Database discrepancies 41
   6.1 Allele discrepancies ............................. 41
   6.1.1 Likelihood ratio test ........................... 45
   6.1.2 Interval estimation ............................. 49
   6.1.3 Bayes factor test ............................... 50
   6.2 Likelihood ratio discrepancies ....................... 53
   6.3 Most probable matching profile discrepancies ........ 56
   6.4 Summary ...................................... 58

Hedell, 2010. xiii
7 Conclusions and discussion

7.1 More loci .................................................. 63
Chapter 1

Introduction

1.1 Background

Sometimes when a crime is committed a DNA trace from the offender is left behind at the crime scene. This piece of DNA evidence can be used in the search for the offender. In Sweden it is The National Laboratory of Forensic Science, SKL (Statens Kriminaltekniska Laboratorium), which analyzes and compares DNA traces from crime scenes and from suspects. The analysis of a DNA trace results in a "fingerprint", a DNA profile, that can be compared to other DNA profiles in order to tell if they match or not.

A number of questions of theoretical and practical interest regarding the databases of DNA profiles and the strength of DNA evidence against a suspect in a trial are not fully investigated at SKL. The issues that this thesis deals with are presented below:

• How are population genetic factors taken into account in the process of estimating the strength of DNA evidence? We will see how a sample of DNA profiles from a population can be used in the process of estimating the strength of DNA evidence in a trial. This includes recommendations of statistical tests for validation of the sample of DNA profiles; a variant of Fisher’s exact test, conducted by simulation. A survey of parameters that are of interest in the process of estimating the strength of DNA evidence is also given.

• How is the strength of DNA evidence affected when the suspect puts the blame on some of his or her close relatives? We will see how different hypotheses regarding the relationship between the suspect and other possible donors of the stain from the crime scene can be combined by two different applications of Bayes’ theorem.

• What DNA profile minimizes the strength of DNA evidence against a suspect under different scenarios? How low is the strength and how is it affected by sampling errors in the estimation procedure? By studying the formulas for strength of evidence under different scenarios we are able to find the minimum values. The impact of sampling error will be investigated by finding the intervals of likely values of the strength of evidence.
for these profiles using two different methods; the bootstrap method and a Bayesian method. In fact, the impact of the sampling error is relevant to all calculations of the strength of evidence but in this thesis we only discuss it in the context of the DNA profile that minimizes the strength of evidence.

• What similarities and differences are there between the different databases of DNA profiles at SKL? We try to answer this question by regarding the data in the databases as samples from different populations and investigate if these populations are similar in aspect of their DNA profiles. This is done by both descriptive and inferential statistics, including likelihood ratio tests and Bayes factor tests.

1.2 Chapter outline

Chapter 2: DNA biology This chapter introduces much of the DNA terminology that will be used throughout the thesis.

Chapter 3: Population genetics In this chapter a brief introduction to some population genetic concepts are given and how these are related to a sample of DNA profiles from a population.

Chapter 4: Likelihood ratios and match probabilities Here we look at how to estimate the strength of DNA evidence against a suspect and how to combine multiple hypotheses given in a trial.

Chapter 5: Most probable matching profile In this chapter we examine the DNA profile that minimizes the strength of DNA evidence under different scenarios.

Chapter 6: Database discrepancies This chapter deals with all the comparisons of the databases of DNA profiles at SKL.

Chapter 7: Conclusion and discussion Finally, a chapter devoted to summary, conclusion and discussion of the results.

Chapter 3 to 6 ends with a summary that is intended to describe the main results of the chapter in a less mathematical fashion.

All simulations and implementations of algorithms have been done with the statistical program R [1], except for the tool for finding the most probable matching profiles, introduced in chapter 5, and their corresponding LRs which has been done using Microsoft Excel.
Chapter 2

DNA biology

This chapter introduces much of the DNA terminology that will be used throughout the thesis. The contents follow mainly from Butler [2].

All human cells, except red blood cells, contain DNA which is divided into 46 chromosomes. Each chromosome comes in two sets, pairs: one inherited from the mother and one from the father. A specific area of the chromosome is usually referred to as a locus (plural: loci) or more correct a genetic locus. A large proportion of the human genome does not carry any genetic information. Still, some of these non-coding regions show variation between individuals and can be utilized for forensic purposes. So called short tandem repeats (STR), express a variation seen as number of repeat units of 2, 3, 4, 5 etc nucleotide bases (the structural units of the DNA). The four base repeats have been found most useful for forensic analysis. Thousands of STRs loci are scattered around the non-coding regions. Without going into further details, the DNA sequence at a four base repeat STR locus is characterized by a number - an allele that takes a value from the set \{0,0.1,0.2,0.3,1,1.1,1.2,1.3,2,2.1,2.2,2.3,3,...\}. The n.1, n.2 and n.3 values refer to so-called micro variants, as the loci might contain non-complete 4 base repeats. For example at SKL the locus named D16S539 is examined, typed, and the result is a genotype (a,b), where a is the allele number from the chromosome half inherited from one of the parents and b is from the other parent. (9,15), (11,11) and (8,13) are three examples of genotypes that may result from the analyze of locus D16S539. If a and b are the same numbers, such as (11,11), the individual is said to have a homozygous genotype at that locus. Otherwise he or she is said to have a heterozygous genotype at that locus. The combination of several genotypes over multiple loci is called a DNA profile. Another name for DNA profile is genotype profile. The DNA profiling at SKL involves eleven genotypes at this time, one of which tells whether the individual is a male or a female. A partial DNA profile is a DNA profile where the genotypes are only known at some of the loci due to inhibited or degraded DNA, in contrast to a full DNA profile where all genotypes of interest are successfully typed.

A full DNA profile is shown in Figure 2.1. In the upper left corner is the visualization of locus D3S1358. The graph indicates that the allele numbers for this individual at this locus is 15 and 16, i.e. the genotype is (15,16). In the
Figure 2.1: A DNA profile.
bottom left corner is the visualization of locus D19S433. There is only one peak with allele number 14, that is approximately twice as high as expected for a peak from the typing of one chromosome half, so the genotype is (14,14). A partial profile would have none or unexpected small peaks at some locus.

The idea behind DNA profiling is that all humans have a unique DNA sequence, except for identical twins. But since only very small portions of the DNA sequences are compared, the DNA profiles cannot be considered as unique. Instead you have to calculate how big the chance is that two persons will have the same DNA profile. This is a useful approach in criminal investigation when the offender has left a stain at the crime scene containing his or her DNA, a saliva stain for example. The forensic experts collect the DNA sample from the crime scene, as well as DNA samples from all suspects (if possible) and compare their DNA profiles in order to tell if any of them match the DNA profile from the crime scene. If they do not match, the individual is generally no longer considered as the donor of the stain. However, if they do match the suspect may put the blame on someone else and a calculation will indicate how much more likely the match between the stain and the suspect is if the stain came from the suspect than if it came from someone else.
Chapter 3

Population genetics

In this chapter we will discuss some necessary basics in population genetics. We will see in chapter 4 how many of these concepts are used in the estimation of the strength of DNA evidence against a suspect.

For the following we define a population as a large group of people that are genetically related, such as "Swedish Caucasians" or "US Hispanics". A sub-population is a division of a population with people that are mutually even more genetically related such as "North East Swedish Caucasians". Distinct and well defined populations and sub-populations are hard to find in the real world but these two terms are still useful as a part of our genetic model.

3.1 Population allele proportions

The allele proportions, i.e. the relative frequencies of the alleles, in a population are not known exactly but have to be estimated in some way. A number of individuals are drawn from our population of interest and their DNA profiles are scored in a reference database. The allele proportions are then estimated from this database of DNA profiles. The reference database at SKL that has been used in this study has DNA profiles from 205 Swedish blood donors.

Now, let us define the database population as the population from which the reference database is drawn. Denote allele \( j \) at locus \( i \) as \( a_{ij} \) and define \( p_{ij} \) as the true proportion of \( a_{ij} \) in the database population. We are at this point not interested in the gender of the individuals so we will not make use of the locus that tells whether the person is a male or a female. The estimated proportion of \( a_{ij} \) in the database population is denoted \( \hat{p}_{ij} \). Strictly, if our database population is made up by several sub-populations having different values of \( p_{ij} \) then \( p_{ij} \) is the allele proportion averaged over all sub-populations [3].

Following Weir [4], assume that our database population consists of \( N \) individuals and that we have taken a random sample of \( n \) individuals without replacement and scored their DNA profiles. Let \( (x_{i1}, x_{i2}, \ldots, x_{ik_i}) \) be the observed allele counts of alleles \( (a_{i1}, a_{i2}, \ldots, a_{ik_i}) \) for locus \( i = 1, \ldots, Q \) where \( Q \) is the number of typed loci. If \( N \) is much larger than \( n \) then it is reasonable to state that the sampling without replacement is approximately the same as sampling
with replacement because, for example, the probability of observing allele \( a_{ij} \) is practically the same before and after it has been sampled. Therefore we consider \( x_i = (x_{i1}, x_{i2}, \ldots, x_{ik_i}) \) to be an observation of the random vector of allele counts \( X_i = (X_{i1}, X_{i2}, \ldots, X_{ik_i}) \) with multinomial distribution, i.e. with probability mass function:

\[
f(x_{i1}, x_{i2}, \ldots, x_{ik_i} | p_{i1}, p_{i2}, \ldots, p_{ik_i}) = \frac{n!}{x_{i1}! \cdots x_{ik_i}!} \prod_{j=1}^{k_i} p_{ij}^{x_{ij}} \tag{3.1}
\]

where \( n = \sum_{j=1}^{k_i} x_{ij}, 0 \leq p_{ij} \leq 1 \) for all \( i, j \) and \( \sum_{j=1}^{k_i} p_{ij} = 1 \) \[5\]. We now want to find an estimator of each allele proportion \( p_{ij} \). Following Weir \[4\], we first introduce the likelihood function \( L \) where the parameters of the probability mass function are treated as variables and vice versa:

\[
L = L(p_{i1}, p_{i2}, \ldots, p_{ik_i} | x_{i1}, x_{i2}, \ldots, x_{ik_i}) = f(x_{i1}, x_{i2}, \ldots, x_{ik_i} | p_{i1}, p_{i2}, \ldots, p_{ik_i}) \tag{3.2}
\]

An estimate of \( p_{ij} \) is found by maximizing the likelihood function. This is accomplished by first taking the logarithm of the likelihood function and then finding its extreme value. Note that we must not forget the constraint \( \sum_{j=1}^{k_i} p_{ij} = 1 \Leftrightarrow p_{ik_i} = 1 - \sum_{j=1}^{k_i-1} p_{ij} \).

\[
\ln(L) = \ln\left( \frac{n!}{x_{i1}! \cdots x_{ik_i}!} (1 - p_{i1} - \ldots - p_{i(k_i-1)})^{x_{ik_i}} \right) = \text{Constant} + x_{i1} \ln(p_{i1}) + \ldots + x_{ik_i} \ln(1 - p_{i1} - \ldots - p_{i(k_i-1)}) \tag{3.3}
\]

Differentiation gives:

\[
\frac{\partial \ln(L)}{\partial p_{ij}} = \frac{x_{ij}}{p_{ij}} - \frac{x_{ik_i}}{1 - p_{i1} - \ldots - p_{i(k_i-1)}} \tag{3.4}
\]

By manipulating the \( k_i - 1 \) equations \( \frac{\partial \ln(L)}{\partial p_{ij}} = 0, j = 1, \ldots, k_i - 1 \), we find the extreme value

\[
\hat{p}_{ij} = \frac{x_{ij}}{n} \tag{3.5}
\]

which is just the sampling proportion. This is the global maximum point \[4\] and is called the maximum likelihood estimator of \( p_{ij} \). Estimation of a parameter by maximizing the likelihood function is one of the most popular techniques. More on this topic is given by Casella and Berger \[5\].

As an example, in table 3.1, this formula has been applied on the reference database at SKL for locus D16S539. The first column gives the allele number, the second tells how many times a specific allele is observed in the sample of 205 individuals (which gives 410 alleles since all individuals carries one allele inherited from each parent). The third column is the estimated population allele proportion. E.g. for allele 8; \( 4/410 = 0.009756098 \).
3.1. Population allele proportions

<table>
<thead>
<tr>
<th>Allele number &lt;br&gt;((a_{ij}))</th>
<th>Allele count &lt;br&gt;((x_{ij}))</th>
<th>Estimated population proportion ((\hat{p}_{ij}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4</td>
<td>0.009756098</td>
</tr>
<tr>
<td>9</td>
<td>51</td>
<td>0.124390244</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>0.039024390</td>
</tr>
<tr>
<td>11</td>
<td>118</td>
<td>0.287804878</td>
</tr>
<tr>
<td>12</td>
<td>133</td>
<td>0.324390244</td>
</tr>
<tr>
<td>13</td>
<td>73</td>
<td>0.178048780</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>0.034146341</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>0.002439024</td>
</tr>
</tbody>
</table>

Table 3.1: Allele proportions for locus D16S539

A different approach, see Lange [6], known as the Bayesian approach in estimating \(p_{ij}\) begins with a prior distribution for the random vector \(p_i = (p_{i1}, p_{i2}, \ldots, p_{ik_i})\). The prior distribution is the distribution we believe \(p_i\) has before we have seen the data from the reference database. The distribution is often chosen as Dirichlet(\(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik_i}\)) [6, 7, 8], which has the probability density function:

\[
f(p_{i1}, p_{i2}, \ldots, p_{ik_i}|\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik_i}) = \frac{\Gamma(a_i)}{\prod_{j=1}^{k_i} \Gamma(\alpha_{ij})} \prod_{j=1}^{k_i} p_{ij}^{\alpha_{ij} - 1}
\]

with \(a_i = \sum_{j=1}^{k_i} \alpha_{ij}\) and \(\alpha_{ij} > 0\) for all \(i, j\) [6]. As before \(0 \leq p_{ij} \leq 1\) for all \(i, j\) and \(\sum_{j=1}^{k_i} p_{ij} = 1\). \(\Gamma(\cdot)\) is known as the Gamma function. For \(z > 0\), [9]:

\[
\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt
\]

If \(z\) is an integer then

\[
\Gamma(z) = (z - 1)!
\]

When the allele counts are multinomially distributed, as we have assumed, the posterior distribution for \(p_i\) is Dirichlet(\(x_{i1} + \alpha_{i1}, x_{i2} + \alpha_{i2}, \ldots, x_{ik_i} + \alpha_{ik_i}\)) \(\cdot\) The posterior distribution is the distribution \(p_i\) has when we combine the data from the reference database with our prior distribution. One estimator of \(p_{ij}\) is then given by the mean of the posterior distribution of \(p_{ij}\):

\[
\hat{p}_{ij} = \frac{x_{ij} + \alpha_{ij}}{n + a_i}
\]

If we do not know anything about the allele proportions we may chose the prior distribution at locus \(i\) as Dirichlet(1, 1, \ldots, 1). In this case all values for the random vector \(p_i\) are equally likely, representing a complete ignorance about the parameter values \(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik_i}\). Again, if the allele counts \((x_{i1}, x_{i2}, \ldots, x_{ik_i})\) for locus \(i\) in the reference database is an observation of a multinomial distribution then the posterior distribution for \(p_i\) is Dirichlet(\(x_{i1} + 1, x_{i2} + 1, \ldots, x_{ik_i} + 1\)).

Applying this approach on the reference database at SKL together with formula (3.9) yields the results in Table 3.2 for locus D16S539. We see that slightly
different results are obtained but these differences approaches zero when the sample size increases since the formula will be dominated by $x_{ij}/n$.

<table>
<thead>
<tr>
<th>Allele number $(a_{ij})$</th>
<th>Estimated population proportion $(\hat{p}_{ij})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.011961722</td>
</tr>
<tr>
<td>9</td>
<td>0.124401914</td>
</tr>
<tr>
<td>10</td>
<td>0.04069856</td>
</tr>
<tr>
<td>11</td>
<td>0.28468995</td>
</tr>
<tr>
<td>12</td>
<td>0.32054163</td>
</tr>
<tr>
<td>13</td>
<td>0.17703493</td>
</tr>
<tr>
<td>14</td>
<td>0.035885167</td>
</tr>
<tr>
<td>15</td>
<td>0.00478469</td>
</tr>
</tbody>
</table>

Table 3.2: Allele proportions for locus D16S539 using a Bayesian approach

Next consider the following scenario; we do not know anything about the allele proportions so we begin with a Dirichlet$(1, 1, \ldots, 1)$ prior. Then we observe data from another database population that we believe is similar to the database population we will sample data from. In our case we have data from a Norwegian population, see Andreassen et al. [10]. Combining these data with our prior distribution yields a posterior distribution that we will use as a prior distribution in combination with the data from the reference database at SKL. By doing so we get a new posterior distribution which contains information from our first prior, the Norwegian population and the Swedish population.

Applying formula (3.9) on this distribution yields the results in Table 3.3 for locus D16S539.

<table>
<thead>
<tr>
<th>Allele number $(a_{ij})$</th>
<th>Estimated population proportion $(\hat{p}_{ij})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.011993383</td>
</tr>
<tr>
<td>9</td>
<td>0.140612076</td>
</tr>
<tr>
<td>10</td>
<td>0.052522746</td>
</tr>
<tr>
<td>11</td>
<td>0.297353184</td>
</tr>
<tr>
<td>12</td>
<td>0.298593879</td>
</tr>
<tr>
<td>13</td>
<td>0.167089232</td>
</tr>
<tr>
<td>14</td>
<td>0.028535980</td>
</tr>
<tr>
<td>15</td>
<td>0.003308519</td>
</tr>
</tbody>
</table>

Table 3.3: Allele proportions for locus D16S539 using a Bayesian approach including both Norwegian and Swedish data.

If we believe that the Norwegian population is genetically similar to the Swedish population then these numbers may be more accurate than those in Table 3.2.
3.1.1 Bayes estimators

In the previous section we used a Bayesian method to estimate $p_{ij}$. The general Bayesian approach for estimating a quantity $\delta$, known as the prior distribution, i.e. $\delta$ is treated as a random variable. Next we observe a random sample $x$ with probability distribution $f(x|\delta)$. The updating of the prior distribution with this new information is done by a version of Bayes’ theorem [5]:

$$\pi(\delta|x) = \frac{f(x|\delta)\pi(\delta)}{\int f(x|\delta)\pi(\delta) d\delta} \tag{3.10}$$

The posterior distribution $\pi(\delta|x)$ is then used to make statements about $\delta$.

3.2 Hardy-Weinberg’s law and linkage equilibrium

Two important concepts in population genetics are introduced in this section; Hardy-Weinberg’s law and linkage equilibrium.

We define the following random variables

- $G_i^x$ = the genotype of individual $x$ at locus $i$.
- $G_x^x$ = $x$’s DNA profile = $(G_1^x; \ldots; G_Q^x)$

Since, a DNA profile is a vector of allele numbers we may write $G_x = g$ when we want to state that a specific DNA profile $g$ is observed when we examine the unknown DNA profile $G_x$.

If allele $j$ at locus $i$ is denoted by $a_{ij}$ then the observation of individual $x$’s DNA profile $G_x$ could we written as $g = ((a_{1j}, a_{1k}), (a_{2j}, a_{2k}), \ldots, (a_{Qj}, a_{Qk}))$, where the $js$ may be different from each other, the same holds for $k$. In this case the genotype at the first locus is $(a_{1j}, a_{1k})$, at the second locus $(a_{2j}, a_{2k})$ etc. Returning to Figure 2.1; the DNA profile of the typed individual $y$ is $g_y = ((15,16), (17,19), (9,12), (20,25), (12,14), (14,18), (14,14), (7,7),(20,22))$. In addition, the $(X,Y)$ genotype tells that the individual is male. A female would have genotype $(X,X)$. Remember that if $a_{ij}$ and $a_{ik}$ are the same then the genotype is called homozygous. Otherwise it is called heterozygous. The probability of $x$ having a specific genotype $G_x^i$ at locus $i$ is, under some assumptions, given by Hardy-Weinberg’s law [2, 11]:

$$P(G_x^i = (a_{ij}, a_{ik})) = \begin{cases} p_{ij}^2 & \text{if homozygous genotype} \\ 2p_{ij}p_{ik} & \text{if heterozygous genotype} \end{cases} \tag{3.11}$$

The value of $P(G_x^i = (a_{ij}, a_{ik}))$ may be interpreted as the proportion of the population that has this genotype. Hardy-Weinberg’s law is a statement of independence between having different alleles at a locus. Knowing one allele does not increase our knowledge about the other allele.

If linkage equilibrium [11] holds then the event of $x$ having a specific genotype at a locus is considered as independent of the event of having any genotype at a different locus. Hence,
\[ P(G_x = ((a_{1j}, a_{1k}), \ldots, (a_{Qj}, a_{Qk}))) = \prod_{i=1}^{Q} P(G_x^{i} = (a_{ij}, a_{ik})) \]

This probability is interpreted as the proportion of the population that has this combination of genotypes. In other words: the proportion of the population that has this DNA profile.

These equations are useful in DNA profiling but they come with a number of assumptions, of which none fully applies in a real population. The assumptions are [11]: infinite population, no mutation or natural selection at loci of interest, no migration into or away from the population, random mating and that an infinite number of generations have passed. Still, in practice Hardy-Weinberg’s law and linkage equilibrium hold approximately [7].

### 3.2.1 Test of Hardy-Weinberg’s law and linkage equilibrium

In this section a test for the validity of Hardy-Weinberg’s law and a test for linkage equilibrium with respect to the reference database are introduced. These specific tests have not been applied to the reference database at SKL earlier, instead other tests have been used but these are less appropriate for the data at hand. The two new tests may also be applied in investigations of future reference databases.

When taking a random sample of DNA profiles from the database population we expect the genotype proportions to be reasonably similar to those given by Hardy-Weinberg’s law. Extreme departures may indicate DNA typing or data entry errors or that the sample is highly unrepresentative to the population. A common statistical test for this type of data is a variant of Fisher’s exact test [7, 11]: Assuming that Hardy-Weinberg’s law holds (the null hypothesis) we estimate the probability \( P_{F_i} \) of obtaining our set of genotypes or less probable sets of genotypes given the allele counts for locus \( i \). This probability is known as a \( P \)-value; the probability that the test statistic is at least as extreme as the value observed given that the null hypothesis is true. If Hardy-Weinberg’s law holds for our data then \( P_{F_i} \) is likely to be high (\( \geq \alpha \in [0, 1] \)). A typical choice is \( \alpha = 0.05 \) because the combination of alleles would be completely random, so it is likely that our observed data is a “typical” set of genotypes, or even one of the most probable ones. Let \( x_{ij}, n_{ijk}, h_i \) be observations of the random variables

\[
X_{ij} = \text{count of allele } a_{ij} \\
N_{ijk} = \text{count of genotype } (a_{ij}, a_{ik}) \\
H_i = \text{total number of heterozygous genotypes}
\]

The probability of obtaining a specific set of genotypes given the allele counts is, under the null hypothesis of Hardy-Weinberg’s law [12]:

\[
P_i = P(\cap_{jk} N_{ijk} = n_{ijk} | \cap_{j} X_{ij} = x_{ij}) = \frac{(\sum_{jk} n_{ijk})!^2}{(2 \sum_{jk} n_{ijk})!} \prod_{j} x_{ij}!}
\]

(3.13)
Two problems arise; first, for example, \( (\sum_{jk} n_{ijk})! \) will in our case be larger than the computer can handle, so instead we calculate \( e^{\ln(P_i)} \) to avoid direct calculations of the factorials, e.g., \( \ln((\sum_{jk} n_{ijk})!) = \ln(\sum_{jk} n_{ijk}) + \ln((\sum_{jk} n_{ijk}) - 1) \ldots + \ln(1), \) etc. The next problem is that we need to find all possible sets of genotypes in order to tell how rare the observed genotype data is. The reference database at SKL has a total of 410 alleles for each locus so the total number of genotype data sets, i.e., the number of ways we can combine the alleles two and two is \( \prod_{i=1}^{205} (2i - 1) \) which is too large to enumerate. Guo and Thompson [13] proposes a simulation strategy to obtain \( P_F \) as follows:

1. Set \( K = 0 \) and calculate \( P' = P_i \) for the observed genotype data.
2. Generate a new set of genotypes by random pairing of the observed alleles.
3. Calculate \( P_i \). If \( P_i \leq P' \) then \( K = K + 1 \)
4. Repeat step 2 and 3 \( N \) times.
5. \( P_F \approx K/N \)

Note that we perform \( N \) comparisons between \( P' \) and the \( N \) \( P_i \)'s. The number of comparisons, \( K \), for which \( P_i \) is less than or equal to \( P' \) is binomially distributed \( \text{Bin}(N, P) \) for an unknown value \( P \). The maximum likelihood estimator of \( P \) is \( K/N \), hence \( P_F \approx K/N \).

When applied to the reference database at SKL with number of runs \( N \) set to 5000 and 10000 we obtain the results in Table 3.4:

<table>
<thead>
<tr>
<th>Locus (i)</th>
<th>( P_F ) 5000 runs</th>
<th>( P_F ) 10000 runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3S1358</td>
<td>0.6212</td>
<td>0.6157</td>
</tr>
<tr>
<td>vWA</td>
<td>0.9924</td>
<td>0.9899</td>
</tr>
<tr>
<td>D16S539</td>
<td>0.8432</td>
<td>0.8467</td>
</tr>
<tr>
<td>D2S1338</td>
<td>0.1036</td>
<td>0.1013</td>
</tr>
<tr>
<td>D8S1179</td>
<td>0.4848</td>
<td>0.4955</td>
</tr>
<tr>
<td>D21S11</td>
<td>0.5462</td>
<td>0.5416</td>
</tr>
<tr>
<td>D18S51</td>
<td>0.7784</td>
<td>0.7754</td>
</tr>
<tr>
<td>D19S433</td>
<td>0.8580</td>
<td>0.8539</td>
</tr>
<tr>
<td>TH01</td>
<td>0.8622</td>
<td>0.8685</td>
</tr>
<tr>
<td>FGA</td>
<td>0.4830</td>
<td>0.4705</td>
</tr>
</tbody>
</table>

Table 3.4: P-values for all loci

The values for 5000 and 10000 are practically the same so we conclude that the algorithm has "converged". The probability of erroneously rejecting the null hypothesis for a test is \( \alpha \). When we perform several tests as above, the probability that we erroneously reject at least one of the hypotheses tested is greater than \( \alpha \). Let us call this probability \( \alpha_{TOT} \). Bonferroni’s inequality [5] says that for \( n \) events \( A_1, A_2, \ldots, A_n \)

\[
P(\cap_{i=1}^{n} A_i) \geq 1 - \sum_{i=1}^{n} (1 - P(A_i)) \quad (3.14)
\]
In our case set \( A_i = \text{test } i \) is not erroneously rejected. If we want

\[
P(\cap_{i=1}^{n} A_i) = 1 - \alpha_{\text{TOT}} \geq 0.95
\]

then we may set \( \alpha = 0.05/n \) since

\[
P(\cap_{i=1}^{n} A_i) \geq 1 - \sum_{i=1}^{n} (1 - P(A_i)) = 1 - n \alpha = 1 - n \frac{0.05}{n} = 0.95
\]

(3.16)

So if we want \( \alpha_{\text{TOT}} \leq 0.05 \) then \( \alpha = 0.05/n = 0.005 \) since we have ten tests for Hardy-Weinberg’s law. When a P-value is below this limit we say that the test is “significant”, i.e. the data is unlikely under the assumption of the null hypothesis. In Table 3.4 all values are high (\( \geq 0.005 \)) and there is no evidence against the null hypothesis of Hardy-Weinberg’s law, hence no evidence of gross DNA typing or data entry errors or that the sample is highly unrepresentative to the population.

If the null hypothesis is incorrect, then the power of a test is its ability to correctly detect this departure from the null hypothesis. Fisher’s exact test has better power than alternative tests for this type of data but the power is still quite low for our sample size of 205 DNA profiles, see e.g Buckleton et al. [11], so it is unlikely that the test will correctly detect any minor departures from Hardy-Weinberg’s law.

The test for linkage equilibrium is similar to the test for Hardy-Weinberg’s law, but now we ask how rare our set of DNA profiles is, if linkage equilibrium holds (the null hypothesis). If the null hypothesis is correct then we expect that our data would, on average, be similar to a set of DNA profiles where all genotypes are combined completely by random. In order to resolve this issue we once again utilize a version of Fisher’s exact test that is closely related to the test for Hardy-Weinberg’s law. By Zaykin et al. [14] we get the test for linkage equilibrium between two loci: Let \( n_{ijk}, n_{lmn} \) and \( n_{ijklmn} \) be observations of the random variables

\[
N_{ijk} = \text{count of genotype } (a_{ij}, a_{ik})
\]

\[
N_{lmn} = \text{count of genotype } (a_{lm}, a_{ln})
\]

\[
N_{ijklmn} = \text{count of profile } ((a_{ij}, a_{ik}), (a_{lm}, a_{ln}))
\]

Then calculate

\[
P_{sl} = P(\cap_{jkmn} N_{ijklmn} = n_{ijklmn} | \cap_{jkmn} N_{ijkl} = n_{ijkl}, \cap_{lmn} N_{lmn} = n_{lmn}) = \frac{\prod_{jkmn} n_{ijklm}! \prod_{lmn} n_{lmn}!}{(\sum_{jkmn} n_{ijklmn})! \prod_{jkmn} n_{ijklmn}!}
\]

(3.17)

We then estimate how many sets of DNA profiles that are equally or less probable under the null hypothesis than our given set of DNA profiles, similar to the test for Hardy-Weinberg’s law. The difference is now that we permute the genotypes and not the alleles. We will then obtain a P-value for each pairwise
test. But the question we really would like to answer is if linkage equilibrium holds for the whole dataset, i.e. if the genotype counts are independent between all loci. However, due to lack of sufficient amount of data we will only be able to test pairs of loci as we have described above. Buckleton et al. [11] proposes two strategies for taking the multiple comparison into account other than using the Bonferroni inequality which gives a conservative limit that may be unnecessary low [15]. Both methods require that the P-values, regarded as random variables, should be independent. As pointed out by [11] the approaches may be useful even if the independent assumptions are not fully met as in our case where, for example, the tests between TH01/FGA and TH01/vWA give some information about FGA/vWA.

One way to combine the test results is to plot the observed P-values from the tests against the quantiles of the expected distribution of the P-values. The quantiles are points taken at regular intervals from the cumulative distribution function of the P-value regarded as a random variable. Under the null hypothesis the n P-values are observations from a $U(0, 1)$ distribution, i.e. the values are expected to be evenly distributed between 0 and 1 [16]. We apply the test to the 45 pairs of loci in the reference database with the number of runs set to 5000. The plot is shown in Figure 3.1. Here the expected values are the quantiles of a $U(0, 1)$ distribution. The points are expected to lie on the diagonal line in the figure if linkage equilibrium holds. A 95% confidence region is also included. For a large number of similar tests, the confidence region is expected to include 95% of the points. This is found by considering the distribution of the order statistics $P(1), \ldots, P(n)$ of the (assumed) independent P-values $P_1, \ldots, P_n$ taken as random variables. The order statistics are the P-values placed in ascending order, so $P(1) \leq \ldots \leq P(n)$. When the P-values follow a $U(0, 1)$ distribution then $P(k)$ has a Dirichlet($k, n - k + 1$) distribution [5, 17]. From this we can find the 95% confidence region for each $P(k)$ by taking the 2.5th and the 97.5th percentile of the cumulative Dirichlet($k, n - k + 1$) distribution, i.e. the points $z_1$ and $z_2$ where $P(Z \leq z_1) = 0.025$ and $P(Z \leq z_2) = 0.975$ for a Dirichlet($k, n - k + 1$) random variable $Z$.

As we can see, all points are within the confidence region and we find no evidence against the null hypothesis. This way of summarizing the P-values are known as a quantile-quantile plot.

A second method of combining the n P-values, that is recommended by [11], is by consider the statistic

$$T = -2 \sum_{i=1}^{n} \ln P_i$$

(3.18)

When all n P-values are treated as independent random variables, T is approximately $\chi^2$ distributed with $2n$ degrees of freedom [18]. Applying this technique to the reference database and the tests for linkage equilibrium yields

$$-2 \sum_{i=1}^{n} \ln P_i \approx 87.7429$$

\footnote{In the thesis we use the word conservative as in the meaning of pessimistic or on the safe side.}
The probability for this or more extreme observations under the null hypothesis is

\[ P(T \geq 87.7429) \approx 0.5477 \]

which is much larger than the standard 0.05 limit. Hence, we find no evidence against the null hypothesis of linkage equilibrium.

### 3.3 Distribution of profile proportions

The commonness of a DNA profile in our population can be assessed using the reference database. To do that we have to use some kind of measurement of the profile and then compare it to the other profiles in our population. A reasonable choice of measurement in this case would be the probability (3.12);

\[ P(\cap_{i=1}^{Q}(a_{ij}, a_{ik})) \]

since it can be interpreted in terms of population proportion of a DNA profile. Secondly, we do not have access to the DNA profiles of all individuals in our population for comparison but by the reference database we may simulate a population that is to some extent similar to the required population. The generated population may be used in different situations but we will restrict the use of it to section 4.4 as a quick way of stating the rarity of a profile over ten and five loci.

The simulation of DNA profiles is done in the following way:

1. For each locus generate two alleles according to their estimated population proportion, i.e. an allele \( a_{ij} \) with estimated population proportion \( \hat{p}_{ij} \) will be generated with probability \( \hat{p}_{ij} \). The two alleles could be the same, in
3.4. Inbreeding coefficients

that case we have generated a homozygous genotype, otherwise they are different and we have generated a heterozygous genotype.

2. When genotypes for all ten loci have been generated (we are not interested of the gender indicating locus) then calculate \( P(\bigcap_{i=1}^{Q} (a_{ij}, a_{ik})) \) using formula (3.11) and (3.12).

3. Repeat these steps \( n \) times in order to find the distribution of DNA profile proportions.

A histogram of 50000 simulated DNA profile proportions on a log\(_{10}\) scale is given in Figure 3.2. The highest profile proportions are found at \( 10^{-10} \). There are only a few combinations of genotypes that yield such high proportions. This is reflected in the histogram by the low density at that segment. We can expect that a typical DNA profile has mixture of common and some less common alleles since there are a large number of combinations that will result in such a profile. Hence, we can expect that the histogram will have a peak of profile proportions at some interval. In our case this seems to be between \( 10^{-15} \) and \( 10^{-12} \). Low profile proportions are those below approximately \( 10^{-17} \).

In Figure 3.3 the distribution of 50000 simulated DNA profile proportions over five loci is given. Five loci is the minimum number of typed loci that are required to record the profile in some of the databases at SKL [19]. For each of the simulated profiles the five loci are selected randomly.

By these means we are able to state the rarity of a DNA profile over ten and five loci in the Swedish population.

3.4 Inbreeding coefficients

Since all populations and subpopulations are finite and the mating is not at random they all show some degree of inbreeding. There are three important measures of this phenomenon; \( \theta \), \( f \) and \( F \). We will begin by investigating some interpretations of \( \theta \):

1. The probability that any two individuals from a subpopulation share a specific allele because it was inherited from a common ancestor in the subpopulation [20].

2. Our degree of uncertainty that the allele proportions from the database population are the same as in the suspect’s subpopulation [21, 22].


If \( \theta = 0 \) then there is no division of the population into subpopulations. By the first interpretation above, two individuals will not share an allele because it was inherited from a common ancestor in the subpopulation. By the second interpretation we are 100 percent sure that our database population, the one that we sampled from to build the reference database, is the same as the suspect’s subpopulation, that is; the subpopulation to which the individual under investigation belong. And by the third interpretation the genetic distance is zero between the subpopulations. In real life \( \theta > 0 \). The more segregated the
Figure 3.2: Simulation of profile proportions over all ten loci.

Figure 3.3: Simulation of profile proportions over five loci.
population is into subpopulations the higher value of $\theta$ is expected. Knowing $\theta$ will affect the estimation of how likely it is that two individuals share a DNA profile, as we will see in the next chapter.

While $\theta$ can be interpreted in terms of inbreeding of the population into subpopulations, $f$ is interpreted in terms of inbreeding within the subpopulation due to non random mating [23]. Some ways of estimating $f$ is given by Ayres and Overall [24].

$F$ is the total inbreeding coefficient [3] and is related to $\theta$ and $f$ as

$$F = \theta + f(1 - \theta)$$ (3.19)

If there is random mating within subpopulations then $f = 0$ and $F = \theta$. In this case we may estimate $\theta$ as described below. Estimations of $\theta$ are given since $\theta$ will be used later on. The technique has earlier been applied to two Norwegian subpopulations [10]. The resulting value of $\theta$ was 0.002. The method for estimating $\theta$ is given by Weir [25, 26]; Assume that we have data from $H$ subpopulations and let $x_{hij}$ denote the sample count of allele $j$ at locus $i$ for subpopulation $h$. We denote the allele proportion as $p_{hij}$. Now set

$$x_{hi} = \sum_j x_{hij}$$ (3.20)

$$x_{ci} = \frac{1}{H - 1} \left( \sum_h x_{hi} - \frac{\sum_h x_{hi}^2}{\sum_h x_{hi}} \right)$$ (3.21)

$$MSW_{ij} = \frac{\sum_h x_{hij}p_{hij}(1 - p_{hij})}{\sum_h(x_{hi} - 1)}$$ (3.22)

$$MSA_{ij} = \frac{\sum_h x_{hij}(p_{hij} - \bar{p}_{ij})^2}{H - 1}$$ (3.23)

where

$$\bar{p}_{ij} = \frac{\sum_h x_{hij}p_{hij}}{\sum_h x_{hij}}$$ (3.24)

then an estimator of $\theta$ is given by

$$\hat{\theta} = \frac{\sum_{ij}(MSA_{ij} - MSW_{ij})}{\sum_{ij}(MSA_{ij} + (x_{ci} - 1)MSW_{ij})}$$ (3.25)

A different approach is used to estimate $\theta$ in the UK Caucasian population, see Foreman et al. [22]. This method utilizes a Bayesian reasoning and $\theta$ is assigned a prior Dirichlet(1.5, 50) distribution. Estimates of allele proportions are made by formula (3.9) with a Dirichlet(1, 1, . . . , 1) prior distribution. Further, this method gives an estimate of $\theta$ between a reference database and each of the sampled subpopulations. The estimator is taken as the posterior mode of the posterior distribution of $\theta$, the point where the posterior distribution of $\theta$ takes its largest value. Using two different reference databases, posterior modes for $\theta$ varies between 0.0001 and 0.0036. These results will be used in calculations later on.
3.5 Summary

The allele proportions in a population are not known exactly but have to be estimated based on a sample of DNA profiles. The sample that has been used consists of profiles from 205 Swedish blood donors which are scored in a so called reference database. Different ways of estimating the allele proportions are then demonstrated.

It is of big importance that we can trust the estimated values of the allele proportions. We therefore performed tests for Hardy-Weinberg’s law and linkage equilibrium in order to find indications of major DNA typing or data entry errors or indications of a highly unrepresentative sample of individuals. The tests did not indicate any significant deviations from Hardy-Weinberg’s law or linkage equilibrium. The tests are more appropriate than earlier applied tests at SKL and may also be applied for validation studies of future reference databases.

By simulating a population that is similar to our population we have a quick way of stating the commonness of a DNA profile by comparison to the simulated population. We will restrict the applications of these means to some specific situations given in the next chapter.

All populations and subpopulations show some degree of inbreeding. There are different measures of this phenomenon: $\theta$, $f$ and $F$. Some interpretations of these are presented in the chapter. Estimations of $\theta$ are given since $\theta$ will be used later on. For a Norwegian population the estimated value of $\theta$ was 0.002, and for a UK Caucasian population the estimated value of $\theta$ was between 0.0001 and 0.0036.
Chapter 4

Likelihood ratios and match probabilities

Assume that a crime is committed and a stain from the offender containing his or her DNA is left behind at the crime scene. Further suppose that a suspect’s DNA profile is found to match the DNA profile from the stain. We will now take a look on how to measure the strength of the DNA evidence against the suspect. The approach and results in section 4.1 and 4.2 follow from Balding [20], Buckleton et al. [11] and from Evett and Weir [3].

4.1 Framework

Define the following events:

\[ H_p = \text{the stain is from the suspect.} \]
\[ H_d = \text{the stain is from someone else than the suspect.} \]

and let \( g \) be an observation of the random variable

\[ G = \text{the common DNA profile of the stain from the crime scene and the suspect.} \]

In a trial there is a prosecution that strives to prove that the suspect is guilty to the crime. The defense on the other hand will do their best in order to show that the suspect is not guilty. \( H_p \) is the hypothesis put forward by the prosecution and \( H_d \) is the hypothesis put forward by the defense. Of interest to the court would be the probability of \( H_p \) in comparison with the probability of \( H_d \), both given the evidence \( G = g \), i.e. if \( P(H_p|G = g)/P(H_d|G = g) \) is greater than one then \( H_p \) would be the more likely event given \( G = g \) and vice versa. Bayes’ theorem states that

\[
P(H_p|G = g) = \frac{P(G = g|H_p)P(H_p)}{P(G = g)} \tag{4.1}
\]

and

\[
P(H_d|G = g) = \frac{P(G = g|H_d)P(H_d)}{P(G = g)} \tag{4.2}
\]
The last step is true because if the suspect left the stain and has profile $g$, the stain will also have profile $g$. Hence $P(G = g|H_p) = P(G = g|H_d) = 1$. Further, the actual profile of the suspect does not depend on who left the stain so $P(G_s = g|H_p) = P(G_s = g|H_d)$.

The LR could of course be estimated and presented alone without the use of (4.3) as given above. But the use of (4.3) provides a framework for the entire case when $P(H_p)/P(H_d)$ is given by the judge.

Next, let us define the random variables:

$G_c = \text{the DNA profile of the stain from the crime scene.}$

$G_s = \text{the DNA profile of the suspect.}$

By noting that $G = g$ is the same event as $(G_c = g) \cap (G_s = g)$ (Here expressed "$G_c = g, G_s = g$") we expand the LR:

$$LR = \frac{P(G = g|H_p)}{P(G = g|H_d)} = \frac{P(G_c = g, G_s = g|H_p)}{P(G_c = g, G_s = g|H_d)} = \frac{P(G_c = g|G_s = g, H_p)P(G_s = g|H_p)}{P(G_c = g|G_s = g, H_d)P(G_s = g|H_d)} = \frac{1}{P(G_c = g|G_s = g, H_d)}$$

(4.4)

The last step is true because if the suspect left the stain and has profile $g$ then, obviously, the stain will also have profile $g$. Hence $P(G_c = g|G_s = g, H_p) = 1$. Further, the actual profile of the suspect does not depend on who left the stain so $P(G_s = g|H_p) = P(G_s = g|H_d)$.

The denominator, $P(G_c = g|G_s = g, H_d)$, is called a match probability because it measures how likely it is that a second person will match the suspect’s profile. Since the DNA profile tells whether the individual is a male or a female we will assume from here on that the suspect and all other possible donors of the stain are of the same sex.

### 4.1.1 Bayes factors

The LR presented in the previous section is related to Bayes factors which are used in hypothesis testing of one null hypothesis $H_0$ versus $p$ competing hypotheses $H_1, H_2, \ldots, H_p$, see e.g. Kass [27]. The observed data $d$ of the random variable $D$ is assumed to have arisen under one of these exclusive hypotheses so $\sum_{i=0}^{p} P(H_i|D = d) = 1$. As in the previous section assume that we only have two competing hypotheses, now denoted $H_0$ and $H_1$. We also assume that $D$ is a discrete random variable. Then

$$\frac{P(H_0|D = d)}{P(H_1|D = d)} = \frac{P(H_0|D = d)}{1 - P(H_0|D = d)}$$

(4.5)
4.2. General formula

The right hand term is by definition an odds. Further

\[
P(H_0) = \frac{P(H_0)}{1 - P(H_0)}
\]  

(4.6)

The Bayes factor is

\[
\frac{P(D = d|H_0)}{P(D = d|H_1)}
\]  

(4.7)

so we may express equation (4.3) as

posterior odds of \(H_0\) = Bayes factor \(\cdot\) prior odds of \(H_0\)  

(4.8)

The Bayes factor is then the ratio between the posterior odds and the prior odds of \(H_0\). If the Bayes factor is greater than one, then the observation of data \(d\) have increased our belief in \(H_0\) versus \(H_1\) in comparison to our prior belief.

If there are free parameters \(\lambda = (\lambda_1, \ldots, \lambda_n)\) in the model with prior density \(\pi(\lambda|H_k)\) for \(k = 0, 1\) then the Bayes factor is obtained by integration over the parameter space:

\[
\frac{P(D = d|H_0)}{P(D = d|H_1)} = \int \cdots \int P(D = d|\lambda, H_0)\pi(\lambda|H_0)d\lambda_1 \ldots d\lambda_n
\]

\[
\int \cdots \int P(D = d|\lambda, H_1)\pi(\lambda|H_1)d\lambda_1 \ldots d\lambda_n
\]  

(4.9)

\(P(G = g|H_p)/P(G = g|H_d)\) from the previous section is called a likelihood ratio because, strictly it is the ratio between two likelihood functions of \(\hat{\lambda} = (\theta, f, p_{11}, \ldots, p_{Qkq})\): \(L(\hat{\lambda}|d, H_k) = P(D = d|\lambda, H_k), k \in 0, 1\). Obviously, it can also be regarded as a Bayes factor if the parameters are eliminated by integration. A common approach in evaluation of DNA evidence is not to integrate over the parameters \(\theta, f, p_{11}, \ldots, p_{Qkq}\) but to “plug in” estimates such as those obtained from formula (3.5) and (3.9). One argument for adopting this simpler approach is that the integration can be computationally cumbersome for little gain [7].

4.2 General formula

Let \(\alpha_l\) denote the event that any two individuals will have \(l\) identical alleles at a locus because they were inherited from their recent known common ancestors. A common ancestor may also be one of themselves. Values for \(P(\alpha_l)\) under some relationships between two individuals are given in Table 4.1 [7, 11]. For example, identical twins will for sure inherit the same two alleles at a locus from their parents, hence \(P(\alpha_2) = 1\). The probability that two individuals will share a DNA profile will depend on their relationship so knowing \(P(\alpha_l)\) will be necessary.

From the previous section we saw that the LR is the reciprocal of the match probability \(P(G_e = g|G_s = g, H_d)\). Now let

- \(H_p\) = the stain is from the suspect.
- \(H_d\) = the stain is from a person with relationship \(r\) to the suspect.
Table 4.1: Relationship coefficients

<table>
<thead>
<tr>
<th>Relationship</th>
<th>$P(\alpha_0)$</th>
<th>$P(\alpha_1)$</th>
<th>$P(\alpha_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical twin</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Full sibling</td>
<td>0.25</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>Parent/child</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Half sibling</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>Grandparent/grandchild</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>Uncle/nephew</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>First cousin</td>
<td>0.75</td>
<td>0.25</td>
<td>0</td>
</tr>
<tr>
<td>Unrelated</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

where $r$ may, for example, be one of the relationships given in the table above. A formula \[11, 20\] for the match probability at locus $i$ when both individuals have a homozygous genotype is:

$$
P(G_c = (a_{ij}, a_{ik}) | G_s = (a_{ij}, a_{ik}), H_d) = 
\frac{P(\alpha_2) + P(\alpha_1) \left( 2\theta + (1 - \theta)p_{ij}\right)}{1 + \theta} + P(\alpha_0) \frac{(2\theta + (1 - \theta)p_{ij})(3\theta + (1 - \theta)p_{ij})}{(1 + \theta)(1 + 2\theta)}$$

(4.10)

For a heterozygous genotype the match probability is

$$
P(G_c = (a_{ij}, a_{ik}) | G_s = (a_{ij}, a_{ik}), H_d) = 
\frac{P(\alpha_2) + P(\alpha_1) \theta + (1 - \theta)(p_{ij} + p_{ik})/2}{1 + \theta} + P(\alpha_0) \frac{(2\theta + (1 - \theta)p_{ij})(\theta + (1 - \theta)p_{ik})}{(1 + \theta)(1 + 2\theta)}$$

(4.11)

$p_{ij}$ is the proportion of allele $a_{ij}$ in the population of the donor of the stain. The usual approach is to plug in an estimate $\hat{p}_{ij}$ into these formulas. Two estimates of $p_{ij}$ was given in section 3.1. It is of big importance that we can trust the estimated values of $p_{ij}$ since ultimately they decide how strong the DNA evidence against the suspect is. We therefore performed tests for Hardy-Weinberg’s law and linkage equilibrium in order to find indications of major DNA typing or data entry errors or indications of a highly unrepresentative sample of individuals. Formula (4.10) and (4.11) is derived under the assumption that the suspect and the true donor of the stain belong to the same subpopulation [7, 11]. Reasons for this will be given soon. Linkage equilibrium is assumed at the subpopulation level so the match probability for the full profile is found by multiplication over all loci:

$$
P(G_c = g | G_s = g, H_d) = \prod_{i=1}^{Q} P(G_c^i = (a_{ij}, a_{ik}) | G_s^i = (a_{ij}, a_{ik}), H_d)$$

(4.12)

The LR which is to be presented in the court is, again

$$
\text{LR} = \frac{1}{P(G_c = g | G_s = g, H_d)}
$$

(4.13)
From this we see that the LR is low when the match probability is high. A low LR is in favor of the suspect while a high LR is not.

One of the reasons for assuming that the suspect and the true donor of the stain belong to the same subpopulation is that an innocent suspect is many times similar to the true offender regarding physical appearance or living area, increasing the chance that they belong to the same subpopulation. A somewhat better argument is that the suspect’s alleles are expected to be more similar to those found in the same subpopulation than those from a different subpopulation. Ignoring this fact will be in disfavor of the suspect because his or her alleles will likely be recognized as more unusual if compared to a different subpopulation, making the LR too high which may result in criticism from the court.

The derivation of the match probabilities assumes that there is no inbreeding due to non-random mating within subpopulations, so that \( f = 0 \). In this case the total inbreeding coefficient is equal to \( \theta \). By the second interpretation of \( \theta \) in section 3.4 there will always be some uncertainty regarding the difference in allele proportions between the database population and the suspect’s subpopulation since we generally do not have genotype data at the subpopulation level. Hence \( \theta \) is nonzero. Further, for homozygous genotypes \( P(G_c = g|G_s = g; H_d) \) is increasing with \( \theta \). For heterozygous genotypes \( P(G_c = g|G_s = g; H_d) \) is increasing with \( \theta \) for allele proportions that are present in daily casework [21]. Hence, in practice larger values of \( \theta \) yields lower values of the LR giving raise to conservative interpretation of the DNA evidence, which is in favor of the suspect. Therefore, in order to not disfavor the suspect we should adopt a reasonably large value of \( \theta \). Note that for unrelated individuals formula (4.10) and (4.11) coincide with Hardy-Weinberg’s law (3.11) when \( \theta = 0 \). Although we did not find any evidence of deviations from Hardy-Weinberg’s law in section 3.2.1 we still apply a nonzero value of \( \theta \) since we know that Hardy-Weinberg’s law do not apply exactly in the broader population and also by the reasons stated above.

In section 3.4 we saw that the investigation of the Norwegian and the UK Caucasian population indicated that \( \theta < 0.01 \). If we conclude that the Swedish population is similar to those in aspect of subdivision of the population we may set \( \theta = 0.01 \) and most likely be on the safe side regarding the calculation of the strength of evidence against the suspect.

### 4.3 SKL scale of conclusions

SKL uses a scale of conclusions from -4 to +4 for reporting the strength of the DNA evidence [28]. If the suspect’s profile and the profile of the crime stain do not match then the conclusion reported is the lowest one of the scale -4. In practice this means that the stain does not originate from the suspect. If they do match, then the LR determines the outcome, as shown in Table 4.2. LRs reaching levels -1 to -3 are very rare and only appear under special circumstances [19]. Currently the scale is only used when the defense’s hypothesis, \( H_d \), is that the stain is from a person that is unrelated to the suspect [19]. But if we conclude that the scale measures the strength of the DNA evidence given \( H_p \) in comparison with \( H_d \), as given in section 4.2, there are no statistical reason for
not using the same scale when $H_d$ considers some other relationship between the suspect and the true donor of the stain, as long as only one and not several possible relationships at once are considered. That situation will be investigated in the next section.

<table>
<thead>
<tr>
<th>Level</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>+4</td>
<td>$&gt; 10^6$</td>
</tr>
<tr>
<td>+3</td>
<td>$6000 - 10^6$</td>
</tr>
<tr>
<td>+2</td>
<td>$100 - 6000$</td>
</tr>
<tr>
<td>+1</td>
<td>$6 - 100$</td>
</tr>
<tr>
<td>0</td>
<td>$1/6 - 6$</td>
</tr>
<tr>
<td>-1</td>
<td>$1/100 - 1/6$</td>
</tr>
<tr>
<td>-2</td>
<td>$1/6000 - 1/100$</td>
</tr>
<tr>
<td>-3</td>
<td>$1/10^6 - 1/6000$</td>
</tr>
<tr>
<td>-4</td>
<td>$&lt; 1/10^6$</td>
</tr>
</tbody>
</table>

Table 4.2: SKL scale of conclusions

### 4.4 Combining hypotheses

Assume that there are $N$ possible donors of the stain, other than the suspect. The hypotheses put forward would now be:

- $H_p = \text{the stain is from the suspect.}$
- $H_{d_1} = \text{the stain is from person 1.}$
- $H_{d_2} = \text{the stain is from person 2.}$
- $\vdots$
- $H_{d_N} = \text{the stain is from person } N.$

So, $H_{d_1}, \ldots, H_{d_N}$ is an exclusive and exhaustive partition of $H_d = \text{the stain is from someone else than the suspect.}$

An application of Bayes’ theorem [11] gives:

$$P(H_p|G_c = g, G_s = g) = \frac{P(G_c = g, G_s = g|H_p)P(H_p)}{P(G_c = g, G_s = g|H_d)P(H_d) + P(G_c = g, G_s = g|H_p)P(H_p)}$$

$$= \frac{1}{1 + \sum_{i=1}^{N} P(G_c = g, G_s = g|H_{d_i})P(H_{d_i})/P(H_p)}$$

(4.14)
Where we used that \( P(G_c = g|G_s = g, H_p) = 1 \) because if the suspect left the stain and has profile \( g \) then, obviously, the stain will also have profile \( g \). Further, the actual profile of the suspect does not depend on who left the stain so \( P(G_s = g|H_p) = P(G_s = g|H_{d_i}) \) for all \( i \).

We may group together the \( N \) individuals according to their relationship with the suspect. Suppose that there are \( N_r \) individuals with relationship \( r \) to the suspect, \( r = 1, \ldots, R \). Then \( N = \sum_{r=1}^R N_r \). If we assume that \( P(G_c = g|G_s = g, H_{d_i}) \) is the same for all individuals \( i \) with relationship \( r \) to the suspect then (4.14) may be written as:

\[
P(H_p|G_c = g, G_s = g) = \frac{1}{1 + \sum_{r=1}^R N_r P(G_c = g|G_s = g, H_{d_i}) P(H_{d_i})/P(H_p)}
\]

(4.15)

\( H_{d_i} \) now means that the stain is from a person with relationship \( r \) to the suspect.

The problem with this approach is that all \( P(H_{d_i})/P(H_p) \) must be given numerical values but this is unlikely to be supplied by the court.

Let us take a look at an example using this method: suppose that the defense’s hypotheses are:

\( H_{d_1} = \) the stain is from a full sibling to the suspect.
\( H_{d_2} = \) the stain is from a person who is unrelated to the suspect and from the same subpopulation

Assume that the suspect has 2 full siblings and that there are 1000000 unrelated persons from the same subpopulation. If we set \( P(H_{d_2})/P(H_p) = 1 \) and \( P(H_{d_1})/P(H_p) = x \) then

\[
P(H_p|G_c = g, G_s = g) = \frac{1}{1 + 2P(G_c = g|G_s = g, H_{d_1})x + 1000000P(G_c = g|G_s = g, H_{d_2})}
\]

(4.16)

Next we need a DNA profile to calculate the match probabilities in the expression above. In section 3.3 we saw that profile proportions ranged from approximately \( 10^{-19} \) to \( 10^{-10} \). Let us investigate the behavior of (4.16) for different values on \( x \) using one DNA profile with a high proportion and one DNA profile with a low proportion:

In Figure 4.1 a DNA profile with proportion \( 1.96 \cdot 10^{-11} \) is used. We see that high prior ratios \( P(H_{d_1})/P(H_p) \) are required if we look for very low values of \( P(H_p|G_c = g, G_s = g) \). In Figure 4.2 a DNA profile with proportion \( 1.90 \cdot 10^{-18} \) is used. As expected a low profile proportion will make the case against the suspect stronger and a sibling has to be very much more likely to be the donor of the stain than the suspect if we seek a low posterior probability.

A second approach in the quest of combining hypotheses is shortly mentioned by Buckleton et al. [11]. We will make a more thoroughly investigation of it
Figure 4.1: Posterior probabilities for a high profile proportion.

Figure 4.2: Posterior probabilities for a low profile proportion.
Here. Using the same techniques as earlier in this section we expand the LR:

\[
\text{LR} = \frac{P(G_c = g, G_s = g | H_p)}{P(G_c = g, G_s = g | H_d)} = \frac{P(G_c = g, G_s = g | H_p)}{P(G_c = g, G_s = g, H_d) / P(H_d)} = \\
\sum_{i=1}^{N} P(G_c = g, G_s = g, H_{d_i}, H_d) / P(H_d)
\]

\[
\sum_{i=1}^{N} P(G_c = g, G_s = g, H_{d_i}, H_d) P(H_d) / P(H_d)
\]

\[
\sum_{i=1}^{N} P(G_c = g) P(G_s = g | H_p) P(G_s = g | H_p) = \frac{1}{1}
\]

\[
\sum_{r=1}^{R} N_r P(G_c = g | G_s = g, H_{d_r}, H_d) P(H_{d_r} | H_d)
\]

In the last step all individuals with relationship \( r \) to the suspect are grouped together. Further:

\[
\sum_{r=1}^{R} N_r P(H_{d_r} | H_d) = \sum_{r=1}^{R} N_r P(H_{d_r}, H_d) / P(H_d) = \\
\sum_{r=1}^{R} N_r P(H_{d_r}) / P(H_d) = \sum_{i=1}^{N} P(H_{d_i}) / P(H_d) = 1
\]

\( P(H_{d_r}, H_d) = P(H_{d_r}) \) because \( H_{d_1}, \ldots, H_{d_R} \) is a partition of \( H_d \). If we let \( R \) be the most distant relationship between the suspect and a possible true donor of the stain and set

\[
\frac{P(H_{d_r} | H_d)}{P(H_{d_r}) | H_d)} = k_r; \quad r = 1, \ldots, R; \quad k_r > 0
\]

then

\[
1 = \sum_{r=1}^{R} N_r P(H_{d_r} | H_d) = \sum_{r=1}^{R} N_r k_r P(H_{d_r} | H_d)
\]

\[
\implies P(H_{d_r} | H_d) = \frac{1}{\sum_{r=1}^{R} N_r k_r}
\]

\[
\implies P(H_{d_r} | H_d) = \frac{k_r}{\sum_{r=1}^{R} N_r k_r}
\]

so

\[
\text{LR} = \frac{\sum_{r=1}^{R} N_r k_r}{\sum_{r=1}^{R} N_r k_r P(G_c = g | G_s = g, H_{d_r}, H_d)} = \\
\frac{\sum_{r=1}^{R} N_r k_r}{\sum_{r=1}^{R} N_r k_r P(G_c = g | G_s = g, H_{d_r})}
\]

This approach solves two problems; first, the result is not a posterior probability but a LR which is the common choice for assessing the strength of DNA evidence.
Secondly we only need to consider prior probabilities $P(H_d|H_a)$ and not the suspect’s prior $P(H_p)$ which is perhaps more acceptable to a court.

Again let us assume that the suspect has 2 full siblings and that there are 1000000 unrelated persons from the same subpopulation, and

- $H_{d_1}$ = the stain is from a full sibling to the suspect.
- $H_{d_2}$ = the stain is from a person who is unrelated to the suspect and from the same subpopulation.

Set $P(H_{d_1}|H_d)/P(H_{d_2}|H_d) = x$. Then

$$LR = \frac{2x + 1000000}{2xP(G_c = g|G_s = g, H_{d_1}) + 1000000P(G_c = g|G_s = g, H_{d_2})} \quad (4.22)$$

As before we need a DNA profile to calculate the match probabilities in the expression above. Let us investigate the behavior of (4.22) for different values on $x$ using one DNA profile with a high population proportion and one DNA profile with a low population proportion. These proportions are given by the results from section 3.3. In Figure 4.3 a DNA profile with proportion $1.6 \cdot 10^{-11}$ is used. The LR is measured on a log10 scale. We see that as long as $P(H_{d_1}|H_d)/P(H_{d_2}|H_d)$ is below approximately 5300 then the LR is above $10^6$, i.e. at level +4 on the scale of conclusions. Minimum values of $P(H_{d_1}|H_d)/P(H_{d_2}|H_d)$ required to drop down on the scale of conclusions under different scenarios is given in Tables 4.3, 4.4, 4.5 and 4.6. In Table 4.3 a full profile with a high estimated profile proportion ($1.6 \cdot 10^{-11}$) is used. The scenario varies from one to five full siblings of the same sex and with $10^4$, $10^5$ or $10^6$ unrelated possible donors of the stain. As we can see, when more siblings are considered the minimum prior ratio required to drop from +4 to +3 is decreased, while adding more unrelated possible donors of the stain increases the minimum prior ratio due to the lower impact of the siblings presence. In the next table, 4.4, a low profile proportion is used ($5.4 \cdot 10^{-18}$). This makes the case against the suspect stronger since his or her DNA profile is estimated as more unlikely to find among the siblings or the unrelated individuals and we see that larger minimum prior ratios are needed.

Since partial profiles appear, two analogous tables, 4.5 and 4.6 for five loci are also given. As mentioned earlier, five loci is the minimum number of typed loci that are required to record the profile in some of the databases at SKL. Naturally, information about the profile of the donor of the stain for only five loci gives lower strength of evidence against the suspect. In Table 4.5 the LR is always below the +4 level and the values gives the minimum values of $P(H_{d_1}|H_d)/P(H_{d_2}|H_d)$ required to drop down from +3 to +2 on the scale of conclusions. The profile used has an estimated profile proportion of $1.55 \cdot 10^{-5}$, while in Table 4.6 the profile has an estimated profile proportion of $4.49 \cdot 10^{-11}$. 
Figure 4.3: LR$s$ for sibling versus unrelated. 2 siblings and 1000000 unrelated individuals and a full profile with a high profile proportion.

<table>
<thead>
<tr>
<th>Siblings</th>
<th>$10^4$ unrelated</th>
<th>$10^5$ unrelated</th>
<th>$10^6$ unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>106.8</td>
<td>1068.0</td>
<td>10680.3</td>
</tr>
<tr>
<td>2</td>
<td>53.4</td>
<td>534.0</td>
<td>5340.1</td>
</tr>
<tr>
<td>3</td>
<td>35.6</td>
<td>356.0</td>
<td>3560.1</td>
</tr>
<tr>
<td>4</td>
<td>26.7</td>
<td>267.0</td>
<td>2670.0</td>
</tr>
<tr>
<td>5</td>
<td>21.3</td>
<td>213.6</td>
<td>2136.1</td>
</tr>
</tbody>
</table>

Table 4.3: Minimum prior ratios for sibling versus unrelated to drop from +4 to +3 on the scale of conclusions. Full profile with a high profile proportion.

<table>
<thead>
<tr>
<th>Siblings</th>
<th>$10^4$ unrelated</th>
<th>$10^5$ unrelated</th>
<th>$10^6$ unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>587.2</td>
<td>5872.0</td>
<td>58719.9</td>
</tr>
<tr>
<td>2</td>
<td>293.6</td>
<td>2936.0</td>
<td>29360.0</td>
</tr>
<tr>
<td>3</td>
<td>195.7</td>
<td>1957.3</td>
<td>19573.3</td>
</tr>
<tr>
<td>4</td>
<td>146.8</td>
<td>1468.0</td>
<td>14680.0</td>
</tr>
<tr>
<td>5</td>
<td>117.4</td>
<td>1174.4</td>
<td>11744.0</td>
</tr>
</tbody>
</table>

Table 4.4: Minimum prior ratios for sibling versus unrelated to drop from +4 to +3 on the scale of conclusions. Full profile with a low profile proportion.
Table 4.5: Minimum prior ratios for sibling versus unrelated to drop from +3 to +2 on the scale of conclusions. Partial profile over five loci with a high profile proportion.

<table>
<thead>
<tr>
<th>Siblings</th>
<th>$10^4$ unrelated</th>
<th>$10^5$ unrelated</th>
<th>$10^6$ unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>124.2</td>
<td>1242.3</td>
<td>12422.6</td>
</tr>
<tr>
<td>2</td>
<td>62.1</td>
<td>621.1</td>
<td>6211.4</td>
</tr>
<tr>
<td>3</td>
<td>41.4</td>
<td>414.1</td>
<td>4141.0</td>
</tr>
<tr>
<td>4</td>
<td>31.1</td>
<td>310.6</td>
<td>3105.7</td>
</tr>
<tr>
<td>5</td>
<td>24.8</td>
<td>248.6</td>
<td>2484.6</td>
</tr>
</tbody>
</table>

Table 4.6: Minimum prior ratios for sibling versus unrelated to drop from +4 to +3 on the scale of conclusions. Partial profile over five loci with a low profile proportion.

<table>
<thead>
<tr>
<th>Siblings</th>
<th>$10^4$ unrelated</th>
<th>$10^5$ unrelated</th>
<th>$10^6$ unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.0</td>
<td>19.9</td>
<td>198.6</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>9.9</td>
<td>99.3</td>
</tr>
<tr>
<td>3</td>
<td>0.7</td>
<td>6.6</td>
<td>66.2</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
<td>5.0</td>
<td>49.6</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
<td>4.0</td>
<td>39.7</td>
</tr>
</tbody>
</table>

4.5 Summary

Formulas for calculating match probabilities and likelihood ratios (LR) are given. LR's are the common choice for assessing the strength of DNA evidence against a suspect. It is interpreted as how many times more likely the match between the DNA profile of the stain from the crime scene and the suspect’s profile is if the suspect left the stain than if someone else left the stain. The strongest DNA evidence at SKL is for those profiles that yield a LR above $10^6$.

It is then discussed how the strength of DNA evidence could be assessed if the defense has multiple hypotheses about the relationship between the suspect and a possible true donor of the stain. One method that we investigate that is only shortly mentioned in the references that have been used, requires that we estimate how many possible donors of the stain there might be for each relationship, and also a prior assessment $x$ of how many times more probable an individual with a certain relationship to the suspect is of being the donor of the stain in comparison with an individual with the most distant relationship considered. Given these numbers we may calculate a LR which may be used as a measure of the strength of DNA evidence. If specific values for the different $x$ cannot be given we may still calculate the values for which the LR drops down from one level on the scale of conclusions to another level. This can be helpful in the process of reaching a final conclusion about the strength of evidence. Perhaps even better we may draw graphs to illustrate the effect of $x$ on the LR. These techniques can be used in a straightforward way, as demonstrated in the chapter.
Chapter 5

Most probable matching profile

A match between two DNA profiles is generally followed by a calculation of the LR. Knowing the lowest possible LR may be of interest by the following reason; a match for a full or partial profile may result in a LR that is always above some level of the scale of conclusions at SKL, which allows for preliminary reporting of the DNA evidence or even make a calculation of LR unnecessary if the lowest possible LR is above the highest level of the scale. This chapter continues and extends parts of the work by Ansell [29] by allowing for different relationships, any level of $\theta$ correction and all possible sets of partial profiles as well as consideration of the sampling error.

We define the most probable matching profile, $g_{\text{mpmp}}$, as the DNA profile that minimizes the LR under the hypothesis $H_d$ given by the defense. The prosecutor’s hypothesis $H_p$ will in our cases always be that the crime stain is from the suspect. In order to minimize the LR we should maximize the match probability $P(G_c = g | G_s = g, H_d)$. And because $P(G_c = g | G_s = g, H_d)$ is the product of the match probabilities at all loci we should maximize $P(G_{ci} = (a_{ij}, a_{ik}) | G_{si} = (a_{ij}, a_{ik}), H_d)$ for all $i$. This could either be the match probability for a homozygous genotype or a heterozygous genotype, so we need to maximize both of these for comparison. A short inspection of (4.10) and (4.11) tells that we will maximize $P(G_{ci} = (a_{ij}, a_{ik}) | G_{si} = (a_{ij}, a_{ik}), H_d)$ if we choose $p_{ij}$ and $p_{ik}$ as high as possible. With these steps undertaken we will obtain $g_{\text{mpmp}}$ and its corresponding LR.

5.1 Tool for Excel

Using the ideas above we may construct and implement a simple algorithm in Excel for obtaining the most probable matching profiles and their corresponding LR$s$ under different hypothesis $H_d = “the stain is from a person with relationship r to the suspect” and with any level of $\theta$ correction. The implemented Excel tool handles both full and partial DNA profiles. Most of these results have earlier not been easily available at SKL. The tool works as follows:
1. The user selects which loci that are part of the DNA profile and what value of $\theta$ correction that should be used.

2. The tool then performs the following steps for all different relationships:

   2.i For each locus it calculates the match probability for a homozygous genotype using the highest allele proportion and formula (4.10), and for a heterozygous genotype using the two highest allele proportions and formula (4.11). It then continues with the higher of these two values.

   2.ii It multiplies the match probabilities that were selected in step 2.i, but only for those loci that the user has chosen (since the user may chose to investigate a partial profile).

   2.iii It calculates the LR by taking the reciprocal of the result from step 2.ii.

The resulting LRs with a full DNA profile, a $\theta$ value of 0.01 and allele proportions estimated using (3.5) are given in Table 5.1.

<table>
<thead>
<tr>
<th>Relationship</th>
<th>min LR</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical twin</td>
<td>$1.00 \cdot 10^0$</td>
<td>0</td>
</tr>
<tr>
<td>Full sibling</td>
<td>$5.81 \cdot 10^3$</td>
<td>+2</td>
</tr>
<tr>
<td>Parent/child</td>
<td>$2.68 \cdot 10^5$</td>
<td>+3</td>
</tr>
<tr>
<td>Half sibling</td>
<td>$1.39 \cdot 10^7$</td>
<td>+4</td>
</tr>
<tr>
<td>Grandparent/grandchild</td>
<td>$1.39 \cdot 10^7$</td>
<td>+4</td>
</tr>
<tr>
<td>Uncle/nephew</td>
<td>$1.39 \cdot 10^7$</td>
<td>+4</td>
</tr>
<tr>
<td>First cousin</td>
<td>$1.47 \cdot 10^8$</td>
<td>+4</td>
</tr>
<tr>
<td>Unrelated</td>
<td>$2.19 \cdot 10^9$</td>
<td>+4</td>
</tr>
</tbody>
</table>

Table 5.1: LRs for $g_{mpmp}$ with a full DNA profile

One conclusion is, under these settings, that for the relationships full sibling and parent/child we cannot for sure say that a match between two full DNA profiles will result in a LR above $10^6$ which is at the highest level, +4, on the scale of conclusions at SKL. Since the suspect and his or her twin have the same DNA profile the LR will always be equal to one and the level of conclusion 0. $g_{mpmp}$ over all loci for the different relationships are given in Table 5.2.

Homozygous genotypes seem to be more common for closely related individuals than for unrelated individuals. For closely related individuals the second term in (4.10) and (4.11) will have greater impact on the match probability than the third term due to the $P(\alpha_i)$ coefficients. The second term favors homozygous profiles when we want to maximize the match probability because, if we ignore $\theta$, $P(\alpha_1)p_{ij} > P(\alpha_1)(p_{ij} + p_{ik})/2$. For unrelated individuals the only contribution comes from the third term which tends to favor heterozygous genotypes because, if we ignore $\theta$, $2p_{ij}p_{ik}$ is in most cases greater than $p_{ij}^2$.

We may also be interested in partial profiles and in that case we should discard the loci that increase the value on the LR the most, since we require the
5.2. Assessing sampling error

The value of the LR for different \( g_{mpmp} \) is affected by sampling error due to the fact that we take a random sample of DNA profiles from the population instead of examining the profiles of all individuals in the population. If we take another or a larger sample we will most probably get a different value of the LR. It is therefore a good idea to assess this sampling error somehow.

The theoretical distribution of the LR is unknown so we need to utilize approximate methods for assessing the sampling error of the LR. Some methods are given by Buckleton et al. [11] and by Curran et al. [30]. Two of these
Table 5.3: LRs for $g_{\text{mpmp}}$ over 9 loci

<table>
<thead>
<tr>
<th>Relationship</th>
<th>min LR</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical twin</td>
<td>$1.00 \cdot 10^6$</td>
<td>0</td>
</tr>
<tr>
<td>Full sibling</td>
<td>$2.09 \cdot 10^3$</td>
<td>+2</td>
</tr>
<tr>
<td>Parent/child</td>
<td>$5.33 \cdot 10^4$</td>
<td>+3</td>
</tr>
<tr>
<td>Half sibling</td>
<td>$1.67 \cdot 10^6$</td>
<td>+4</td>
</tr>
<tr>
<td>Grandparent/grandchild</td>
<td>$1.67 \cdot 10^6$</td>
<td>+4</td>
</tr>
<tr>
<td>Uncle/nephew</td>
<td>$1.67 \cdot 10^6$</td>
<td>+4</td>
</tr>
<tr>
<td>First cousin</td>
<td>$1.25 \cdot 10^7$</td>
<td>+4</td>
</tr>
<tr>
<td>Unrelated</td>
<td>$1.24 \cdot 10^8$</td>
<td>+4</td>
</tr>
</tbody>
</table>

Table 5.4: LRs for $g_{\text{mpmp}}$ over 8 loci

<table>
<thead>
<tr>
<th>Relationship</th>
<th>min LR</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical twin</td>
<td>$1.00 \cdot 10^6$</td>
<td>0</td>
</tr>
<tr>
<td>Full sibling</td>
<td>$7.51 \cdot 10^2$</td>
<td>+2</td>
</tr>
<tr>
<td>Parent/child</td>
<td>$1.05 \cdot 10^4$</td>
<td>+3</td>
</tr>
<tr>
<td>Half sibling</td>
<td>$2.11 \cdot 10^5$</td>
<td>+3</td>
</tr>
<tr>
<td>Grandparent/grandchild</td>
<td>$2.11 \cdot 10^5$</td>
<td>+3</td>
</tr>
<tr>
<td>Uncle/nephew</td>
<td>$2.11 \cdot 10^5$</td>
<td>+3</td>
</tr>
<tr>
<td>First cousin</td>
<td>$1.21 \cdot 10^6$</td>
<td>+4</td>
</tr>
<tr>
<td>Unrelated</td>
<td>$8.36 \cdot 10^6$</td>
<td>+4</td>
</tr>
</tbody>
</table>
5.2. Assessing sampling error

<table>
<thead>
<tr>
<th>Relationship</th>
<th>min LR</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical twin</td>
<td>1.00 · 10^9</td>
<td>0</td>
</tr>
<tr>
<td>Full sibling</td>
<td>6.25 · 10^4</td>
<td>+3</td>
</tr>
<tr>
<td>Parent/child</td>
<td>2.99 · 10^5</td>
<td>+3</td>
</tr>
<tr>
<td>Half sibling</td>
<td>1.74 · 10^7</td>
<td>+4</td>
</tr>
<tr>
<td>Grandparent/grandchild</td>
<td>1.74 · 10^7</td>
<td>+4</td>
</tr>
<tr>
<td>Uncle/nephew</td>
<td>1.74 · 10^7</td>
<td>+4</td>
</tr>
<tr>
<td>First cousin</td>
<td>2.07 · 10^8</td>
<td>+4</td>
</tr>
<tr>
<td>Unrelated</td>
<td>2.81 · 10^9</td>
<td>+4</td>
</tr>
</tbody>
</table>

Table 5.5: LRs for \( g_{mpmp} \) with a full DNA profile using Bayesian estimation with Norwegian data as prior.

methods are implemented as described below. The first one is a resampling method, known as the bootstrap method, see e.g Efron and Tibshirani [31], that calculates an approximate confidence interval for the lowest possible LR over \( k \) loci as follows:

1. Generate a new database by sampling profiles from the reference database with replacement until the size of the new database equals the reference database.
2. Estimate allele proportions using formula (3.5).
3. Calculate the highest match probability for each locus using either formula (4.10) with the highest allele proportion as estimated in step 2 or formula (4.11) with the two highest allele proportions as estimated in step 2. Proceed with the highest of these two match probabilities.
4. Compare the 10 resulting match probabilities and discard the \( 10 - k \) lowest values. We discard the lowest values since the highest values are those yielding the lowest likelihood ratio.
5. Calculate the LR by taking the reciprocal of the product of the remaining match probabilities as in formula (4.13).
6. Save the LR and repeat step 1-5 \( n \) times.
7. We now have \( n \) values of the LR that are all the lowest possible ones with regard to the \( n \) databases that were sampled in step 1. Sort the \( n \) LRs. The 2.5 and the 97.5 percentiles of the sorted LRs now yield an approximate 95% confidence interval for the true LR.

In a large number of similar experiments, approximately 95% of the confidence intervals will cover the true value of the LR.

The second method, as given by Curran [8], takes advantage of the Bayesian approach that we briefly examined in section 3.1; the observed allele proportions are assumed to constitute an observation of a Dirichlet distribution. By repeatedly taking observations from that distribution and calculating the LR for \( g_{mpmp} \) we build up a sample of LRs from which we may construct a Bayesian
credible interval for the true LR. The interpretation of a 95% credible interval for
the LR is that the probability of LR being inside the credible interval is
95%. This is different from the interpretation of a classical confidence inter-
val in which the probability assessment regards the interval itself and not the
parameter.

As in chapter 3 we shall use two different Dirichlet distributions. One con-
structed with a Dirichlet(1,1,\ldots,1) prior distribution for all loci and data from
the reference database at SKL, and one that also includes data from the Nor-
wegian population. The 95% approximate credible interval is found by the
following algorithm:

1. Create a Dirichlet posterior distribution for each locus by updating the
   Dirichlet(1,1,\ldots,1) prior with the desired data.
2. Take a random observation from that posterior distribution for each locus.
3. Calculate the lowest possible LR as before.
4. Save the LR and repeat step 2-3 \(n\) times.
5. Sort the \(n\) LRs. The 2.5 and the 97.5 percentiles now yield an approximate
   95% credible interval for the LR.

By looking at Table 5.3 in the previous section we might wonder if the relation-
ships half sibling, grandparent/grandchild and uncle/nephew for sure gives a LR
for \(g_{mpmp}\) at +4 for a partial profile over 9 loci if we also include the sampling
error. Another question is if the relationships first cousin and unrelated always
results in a LR at +4 for \(g_{mpmp}\) if we investigate a partial profile for 8 loci as
in Table 5.4.

In the examination of these scenarios the number of runs \(n\) is set to 20000. The
results from the analysis of the reference database data are presented in Tables
5.6, 5.7 and 5.8. In Table 5.6 confidence and credible intervals for the LR of
\(g_{mpmp}\) with a partial profile over 9 loci for half sibling, grandparent/grandchild
and uncle/nephew are given. The confidence interval obtained using the boot-
strap method indicates that the strength of evidence might be below +4 for
\(g_{mpmp}\), while the credible intervals, first using a Dirichlet(1,1,\ldots,1) prior only
and then a Dirichlet(1,1,\ldots,1) prior and also the Norwegian data included,
suggest that it remains at +4.

Next we compute confidence and credible intervals for the LR of \(g_{mpmp}\) with a
partial profile over 8 loci between first cousins. The result is shown in Table 5.7.
The bootstrap method and the Bayesian method with a Dirichlet(1,1,\ldots,1) prior
only suggest that the strength of evidence might be below +4 while the
Bayesian method with the Norwegian data included indicates that the strength
of evidence is at +4. As an illustration, Figure 5.1 show a histogram of the
20000 LRs generated using the Dirichlet(1,1,\ldots,1) prior.

In Table 5.8 we once again consider the LR of \(g_{mpmp}\) with a partial profile over
8 loci, but this time for unrelated individuals. All intervals suggest that the
strength of evidence is at +4.
5.3. Summary

We define the most probable matching profile as the DNA profile that mini-
mizes the strength of DNA evidence, the LR, under the hypothesis given by the
defense. Knowing the lowest possible LR may be of interest by the following
reason; a match for a full or partial profile may result in a LR that is always
above some level of the scale of conclusions at SKL, which allows for preliminary
reporting of the DNA evidence or even make a calculation of LRs unnecessary
if the lowest possible LR is above the highest level on the scale.

A tool in Excel is developed for obtaining the most probable matching profiles
and their corresponding LRs under different hypothesis from the defense regard-
ing the relationship between the suspect and a possible true donor of the stain.
The tool handles both full and partial DNA profiles. The results are shown in
the tables in this chapter.

From the calculations leading to these results we assess the minimum number
of required loci a DNA profile should consist of for the LR of the most probable
matching profile to be above $10^6$ between a suspect and his or her full siblings.
The minimum number of loci is estimated as 17.
Figure 5.1: Histogram of the 20000 generated LRs of $g_{mpmp}$ for cousins over 8 loci using a Dirichlet(1, 1, \ldots, 1) prior distribution.

The value of the LR is affected by sampling error due to the fact that we only take a random sample of DNA profiles from the population. If we take another or a larger sample we will most probably get a different value of the LR. Hence, it may be a good idea to investigate the effect of the sampling error on the LR. Here we do that for the LR of the most probable matching profile. Two methods for doing this are examined and these are then applied to the most interesting profiles; those that are just above the highest level on the scale of conclusions according to the earlier results. For example, we find that the LR is likely to remain at the highest level on the scale of conclusions when we assess the LR for the most probable matching profile over 8 loci between unrelated individuals, even when taking the sampling error into account.
Chapter 6

Database discrepancies

In this chapter we will investigate discrepancies and similarities of some different DNA databases. These issues have not yet been considered at SKL.

We will work with three different databases: the reference database that we have used earlier in this thesis, and two casework databases of DNA profiles at SKL that we will refer to as database A and database B. Database A consists of 14983 full profiles and database B consist of 77055 full profiles. As mentioned earlier the reference database consists of 205 full profiles. The three databases are regarded as samples from three populations and our task is to find out if these three populations are similar in aspect of their DNA profiles.

6.1 Allele discrepancies

In this section we will investigate the differences in allele proportions between the different databases. Remember that the allele proportions are simply a vector of values between zero and one that sum to one for each locus. Let \( p_{dij} \) denote the population allele proportion of allele \( j \) at locus \( i \) in population \( d \). To measure the distance between two populations in terms of the allele proportions \((p_{1i1}, \ldots, p_{1ik})\) and \((p_{2i1}, \ldots, p_{2ik})\) for locus \( i \) we utilize two common distance measures; the \( L_1 \) distance also known as the Manhattan distance;

\[
\sum_{j=1}^{k_i} |p_{1ij} - p_{2ij}|
\]

and the \( L_2 \) distance also known as the Euclidean distance;

\[
\sqrt{\sum_{j=1}^{k_i} (p_{1ij} - p_{2ij})^2}
\]

In practice we are using the estimated allele proportions \( \hat{p}_{dij} \) to estimate the distances. The results are shown in Table 6.1, 6.2 and 6.3. The largest and the smallest distances are indicated by bold numbers. Table 6.1 shows the estimated distances between the reference database population and the database A population. The largest distance seems to be that for locus D19S433, and
the smallest for locus D2S1338. Bar charts of the estimated allele proportions for these loci are presented in Figure 6.1 and 6.2.

<table>
<thead>
<tr>
<th>Locus</th>
<th>$L_1$ distance</th>
<th>$L_2$ distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3S1358</td>
<td>0.1281</td>
<td>0.0572</td>
</tr>
<tr>
<td>VWA</td>
<td>0.1123</td>
<td>0.0612</td>
</tr>
<tr>
<td>D16S539</td>
<td>0.0756</td>
<td>0.0334</td>
</tr>
<tr>
<td>D2S1338</td>
<td>0.0524</td>
<td>0.0204</td>
</tr>
<tr>
<td>D8S1179</td>
<td>0.1711</td>
<td>0.0861</td>
</tr>
<tr>
<td>D21S11</td>
<td>0.0865</td>
<td>0.0284</td>
</tr>
<tr>
<td>D18S51</td>
<td>0.1181</td>
<td>0.0421</td>
</tr>
<tr>
<td>D19S433</td>
<td>0.1958</td>
<td>0.0975</td>
</tr>
<tr>
<td>TH01</td>
<td>0.1000</td>
<td>0.0514</td>
</tr>
<tr>
<td>FGA</td>
<td>0.1307</td>
<td>0.0479</td>
</tr>
</tbody>
</table>

Table 6.1: Distances between loci from the reference database and DNA database A

The comparison of the reference database and database B is given in Table 6.2. Still, it is locus D19S433 and D8S1179 that shows the largest discrepancies while locus D2S1338 shows the smallest discrepancy. The ordering of discrepancies is not identical for the other loci. For example in the comparison of the reference database and database A the third largest distance is for FGA when using the $L_1$ distance and vWA when using the $L_2$ distance. In the comparison with database B the third largest distance is for TH01 using either the $L_1$ distance or the $L_2$ distance.

<table>
<thead>
<tr>
<th>Locus</th>
<th>$L_1$ distance</th>
<th>$L_2$ distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3S1358</td>
<td>0.1097</td>
<td>0.0507</td>
</tr>
<tr>
<td>VWA</td>
<td>0.1112</td>
<td>0.0606</td>
</tr>
<tr>
<td>D16S539</td>
<td>0.1027</td>
<td>0.0437</td>
</tr>
<tr>
<td>D2S1338</td>
<td>0.0520</td>
<td>0.0193</td>
</tr>
<tr>
<td>D8S1179</td>
<td>0.1663</td>
<td>0.0893</td>
</tr>
<tr>
<td>D21S11</td>
<td>0.1191</td>
<td>0.0381</td>
</tr>
<tr>
<td>D18S51</td>
<td>0.1251</td>
<td>0.0442</td>
</tr>
<tr>
<td>D19S433</td>
<td>0.2125</td>
<td>0.1059</td>
</tr>
<tr>
<td>TH01</td>
<td>0.1381</td>
<td>0.0767</td>
</tr>
<tr>
<td>FGA</td>
<td>0.1378</td>
<td>0.0537</td>
</tr>
</tbody>
</table>

Table 6.2: Distances between loci from the reference database and DNA database B

When comparing database A with database B, see Table 6.3, we see that the distances are much smaller than the comparisons of the reference database. A bar chart of the largest discrepancy, that for TH01, is given in Figure 6.3.

Given the small differences between database A and B we will mainly pay attention to further investigation of discrepancies between the reference database and database A, and between the reference database and database B.
6.1. Allele discrepancies

Figure 6.1: Estimated allele proportions for locus D19S433 for the reference database population and the database A population.

Figure 6.2: Estimated allele proportions for locus D2S1338 for the reference database population and the database A population.
### Table 6.3: Distances between loci from DNA database A and DNA database B

<table>
<thead>
<tr>
<th>Locus</th>
<th>$L_1$ distance</th>
<th>$L_2$ distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3S1358</td>
<td>0.0316</td>
<td>0.0155</td>
</tr>
<tr>
<td>vWA</td>
<td>0.0172</td>
<td>0.0092</td>
</tr>
<tr>
<td>D16S539</td>
<td>0.0306</td>
<td>0.0127</td>
</tr>
<tr>
<td>D2S1338</td>
<td>0.0399</td>
<td>0.0138</td>
</tr>
<tr>
<td>D8S1179</td>
<td>0.0387</td>
<td>0.0158</td>
</tr>
<tr>
<td>D21S11</td>
<td>0.0393</td>
<td>0.0140</td>
</tr>
<tr>
<td>D18S51</td>
<td>0.0288</td>
<td>0.0108</td>
</tr>
<tr>
<td>D19S433</td>
<td>0.0392</td>
<td>0.0145</td>
</tr>
<tr>
<td>TH01</td>
<td><strong>0.0520</strong></td>
<td><strong>0.0292</strong></td>
</tr>
<tr>
<td>FGA</td>
<td>0.0394</td>
<td>0.0133</td>
</tr>
</tbody>
</table>

Figure 6.3: Estimated allele proportions for locus TH01 for the database A population and the database B population.
6.1. Allele discrepancies

6.1.1 Likelihood ratio test

We saw in the previous section that there seems to be some differences between the reference database and the other two databases that may require some more investigation. In this section we will examine if the data is consistent with equal allele proportions.

We begin with independent random variables $X_{mi} = \text{allele counts for DNA database } m \text{ at locus } i$. $X_{mi}$ is multinomially distributed with probability mass function $f(x_{mi}|p_{mi})$, where $p_{mi}$ is the vector of allele proportions. For database $d$ and $c$ the null hypothesis and the alternative hypothesis is

$$H_0 : p_{di} = p_{ci}$$
$$H_1 : p_{di} \neq p_{ci}$$

where $p_{ci}$ is some vector of allele proportions. We may refer to these kind of hypotheses as point null hypotheses since we test if the parameters are exactly the same, namely equal to the point $p_{ci}$. A common way of testing this point null hypothesis is by performing a likelihood ratio test [5]. Under the assumptions we have made we consider the test statistic

$$\lambda(x_{di}, x_{ci}) = \frac{\sup_{p_{di}=p_{ci}} L(p_{di}, p_{ci}|x_{di}, x_{ci})}{\sup_{p_{di}, p_{ci}} L(p_{di}, p_{ci}|x_{di}, x_{ci})} = \sup_{p_{di}=p_{ci}} \frac{f(x_{di}, x_{ci}|p_{di}, p_{ci})}{f(x_{di}, x_{ci}|p_{di}, p_{ci})} \sup_{p_{ci}} f(x_{ci}|p_{ci})$$

(6.3)

where the independence assumption between $X_{di}$ and $X_{ci}$ is used. The numerator of the test statistic is the supremum of the likelihood function $L(p_{di}, p_{ci}|x_{di}, x_{ci})$ under the null hypothesis. In contrast, the denominator is the supremum of the likelihood function when we consider the whole parameter space for $p_{di}$ and $p_{ci}$. As in section 3.1, the maximum points of the multinomial distributions are the sampling proportions, i.e. $(x_{di} + x_{ci})/\sum_i (x_{di} + x_{ci})$ for the numerator and $x_{di}/\sum_i x_{di}$ and $x_{ci}/\sum_i x_{ci}$ for the denominator. If the null hypothesis is true then we expect the numerator and the denominator to be similar, that is we expect $\lambda(x_{di}, x_{ci})$ relatively close to one. Exactly how small the observed value of the test statistic should be in order for us to reject the null hypothesis depends on the probability distribution of $\lambda(X_{di}, X_{ci})$. We have to search for a constant $c$ such that $P(\lambda(X_{di}, X_{ci}) \leq c) \leq \alpha$ if the null hypothesis is true. $\alpha$ is the probability of rejecting the null hypothesis when it is in fact true. A typical choice is $\alpha = 0.05$. A way of finding an approximate probability distribution of $\lambda(X_{di}, X_{ci})$ is by consulting asymptotic theory. One theorem [5] says that when the sample size approaches infinity then $-2\ln\lambda(X_{di}, X_{ci})$ converges to a $\chi^2$ distribution with degrees of freedom equal to the difference between the number of free parameters specified by $H_0$ and the number of free parameters specified by $H_1$. In our case we have $k_i$ parameters for both $X_{di}$ and $X_{ci}$. Those are $p_{di} = (p_{di1}, \ldots, p_{dik_i})$ and $p_{ci} = (p_{ci1}, \ldots, p_{cik_i})$. Since $\sum_{j=1}^{k_i} p_{mij} = 1$ for $m = d$ or $c$ there are $k_i - 1$ free parameters for each of these. So when we consider the whole parameter space for $p_{di}$ and $p_{ci}$ there are $2k_i - 2$
free parameters. Under the null hypothesis there is only one parameter vector
\( p_{bi} = (p_{bi1}, \ldots, p_{bik_i}) \) so there are \( k_i - 1 \) free parameters. The difference is \( k_i - 1 \)
free parameters. Hence, \( -2 \ln \lambda(X_{di}, X_{ci}) \) is approximately \( \chi^2 \) distributed with
\( k_i - 1 \) degrees of freedom for large sample sizes.

One problem that we may encounter is that some alleles that are present in one
database may be very rare or many times not even present in another database.
This violates the assumption of a large sample size so in the following tests rare
alleles are pooled together into a new "allele", so that all allele counts reach at
least five.

Table 6.4 shows the P-values resulting from the likelihood ratio test between
the reference database and database A. There are two loci that give quite low
P-values (\( \leq 0.05 \)); D19S433 and D8S1179. These loci also showed the largest
\( L_1 \) and \( L_2 \) distances. However, when applying the Bonferroni correction that
we introduced in chapter 3 we get the limit for significance at 0.005 which all
P-values exceed. Using the \( \chi^2 \) statistic (3.18) gives an overall P-value of 0.1171
which is greater than the rule-of-thumb limit of 0.05. Hence, it seems that the
deviations between the reference database and database A may be explained by
random effects only.

Next, Table 6.5 shows the P-values resulting from the likelihood ratio test be-
tween the reference database and database B. Again D19S433 and D8S1179
show quite low P-values. Using the \( \chi^2 \) statistic (3.18) gives an overall P-value
of 0.0137 which hints that the difference may not be explained by random effects
only. No P-value is below the conservative limit of 0.005 given by the Bonferroni
correction.

<table>
<thead>
<tr>
<th>Locus (i)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3S1358</td>
<td>0.1365</td>
</tr>
<tr>
<td>vWA</td>
<td>0.2106</td>
</tr>
<tr>
<td>D16S539</td>
<td>0.6628</td>
</tr>
<tr>
<td>D2S1338</td>
<td>0.9175</td>
</tr>
<tr>
<td>D8S1179</td>
<td>0.0367</td>
</tr>
<tr>
<td>D21S11</td>
<td>0.9526</td>
</tr>
<tr>
<td>D18S51</td>
<td>0.2978</td>
</tr>
<tr>
<td>D19S433</td>
<td>0.0356</td>
</tr>
<tr>
<td>TH01</td>
<td>0.4181</td>
</tr>
<tr>
<td>FGA</td>
<td>0.3598</td>
</tr>
</tbody>
</table>

Table 6.4: P-values for the likelihood ratio test between the reference database
and database A.

An important factor in this kind of hypothesis testing is the sample size. Not
only for the sake of asymptotic theory but also for how sensitive the test is
in terms of detecting a deviation from the null hypothesis. As mentioned in
chapter 3, if the null hypothesis is incorrect, then the power of a test is its
ability (probability) to correctly detect this departure from the null hypothesis.
Clearly we want this quantity to be large (close to one) when there is a deviation
from the null hypothesis.
6.1. Allele discrepancies

<table>
<thead>
<tr>
<th>Locus (i)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3S1358</td>
<td>0.3037</td>
</tr>
<tr>
<td>vWA</td>
<td>0.1820</td>
</tr>
<tr>
<td>D16S539</td>
<td>0.2603</td>
</tr>
<tr>
<td>D2S1338</td>
<td>0.9518</td>
</tr>
<tr>
<td>D8S1179</td>
<td>0.0191</td>
</tr>
<tr>
<td>D21S11</td>
<td>0.6352</td>
</tr>
<tr>
<td>D18S51</td>
<td>0.2718</td>
</tr>
<tr>
<td>D19S433</td>
<td>0.0169</td>
</tr>
<tr>
<td>TH01</td>
<td>0.1045</td>
</tr>
<tr>
<td>FGA</td>
<td>0.1552</td>
</tr>
</tbody>
</table>

Table 6.5: P-values for the likelihood ratio test between the reference database and database B.

A simulation of the power of the test for different distances between the reference database population and the database A population is shown in Figure 6.4. To make the simulation "realistic" we first generate two vectors of allele proportions $p_1$ and $p_2$ that are similar to a known vector of allele proportions (here locus vWA is used but other loci give similar results) and measure the $L_1$ distance between them (similar results is reached when the $L_2$ distance is used). Next we generate observations from a multinomial distribution 10000 times with parameter $p_1$ and with sample size equal to the size of the reference database, and also generate observations 10000 times from a multinomial distribution with parameter $p_2$ and with sample size equal to the size of database A. From these observations we perform 10000 tests and check how many times the null hypothesis is rejected at the 0.005 level. It "should" be rejected since $p_1$ and $p_2$ are unequal. This result serves as the estimated power of the test for this distance of allele proportions. We repeat the process all over 250 times and get new estimated powers of the test at different distances of allele proportions. From Figure 6.4 we see that when the distance is small (< 0.15 say) the power is quite low (< 0.40), so we do not expect the test to detect small departures from the null hypothesis. When the distance is somewhat larger (> 0.20 say) the power is quite high (> 0.80) so there is a good chance that the test will detect this level of departure from the null hypothesis. The conclusions are the same when comparing the reference database and database B, see Figure 6.5.

If the sample sizes are large enough then the test will reject the point null hypothesis no matter how small the difference is. The samples contained in database A and B are large so when we test $H_0: p_{da} = p_{ba}$ for database A and B, we will most likely reject the null hypothesis even if the difference is practically irrelevant. This is one of the weaknesses with testing a point null hypothesis. As shown in Figure 6.6 the estimated power is equal to one even for $L_1$ distances as small as 0.03. If we perform likelihood ratio tests between database A and database B, indeed almost all resulting P-values are very low; all below $10^{-6}$ except the P-value for locus vWA which is 0.1554. vWA is also the locus that has the smallest $L_1$ and $L_2$ distances between the estimated allele proportions from database A and database B.
Figure 6.4: Estimated power for the likelihood ratio test between the reference database and database A.

Figure 6.5: Estimated power for the likelihood ratio test between the reference database and database B.
6.1. Allele discrepancies

Figure 6.6: Estimated power for the likelihood ratio test between database A and database B.

6.1.2 Interval estimation

Another way of telling if there are any major differences between the allele proportions for the different databases is by constructing a confidence region for the difference of allele proportions for a specific locus and see whether the zero vector is contained within this region. Fitzpatrick and Scott [32] give a method for finding the total coverage probability (confidence level) for the combination of all confidence intervals of differences of proportions; Assuming that we have two multinomial samples of size \( N_1 \) and \( N_2 \) from two different populations of DNA profiles we estimate the allele proportions \((p_{1i1}, \ldots, p_{1i_k})\) and \((p_{2i1}, \ldots, p_{2i_k})\) respectively using (3.5). A conservative confidence interval for \( p_{1ij} - p_{2ij} \) with coverage probability \( \alpha \) is

\[
I_{p_{1ij} - p_{2ij}} = \hat{p}_{1ij} - \hat{p}_{2ij} \pm \frac{\lambda(\alpha/2)}{2} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}
\]

(6.4)

where \( \lambda(\alpha/2) \) is the upper 100\((1 - \alpha/2)\)% point of the standard normal distribution. The total coverage probability of the univariate confidence intervals \( I_{p_{1ij} - p_{2ij}} \) is then at least equal to

\[
T(\alpha) = \begin{cases} 
1 - 2\alpha & \alpha \leq 0.016 \\
6\Phi\left(\frac{3\lambda(\alpha/2)}{\sqrt{8}}\right) - 5 & 0.016 < \alpha \leq 0.15
\end{cases}
\]

(6.5)

where \( \Phi(\cdot) \) denotes the standard normal distribution function. \( \alpha \) is set to 0.024 which gives a total coverage probability of at least 0.95. In the comparison between the reference database and database A, the only confidence regions that do not include zero are the ones for loci D19S433 and D8S1179, the same
loci that showed low P-values in the likelihood ratio test. The same result is reached when comparison is made between the reference database and database B.

6.1.3 Bayes factor test

Another way of investigating the differences or similarities between the database populations is by estimating the Bayes factor that we encountered in section 4.1.1. As in section 6.1.1 we begin with independent, random variables \(X_{mi} = \text{allele counts for DNA database } m \text{ at locus } i\). \(X_{mi}\) is multinomially distributed with probability mass function \(f(x_{mi}|p_{mi})\). But now we consider \(p_{mi}\) to be a random variable with probability density function \(g(p_{mi}|\tilde{\alpha}_{mi})\) that describes our prior belief about the allele proportions. As earlier we will take these as a random variable with probability density function \(g\)

Also assume that our prior knowledge tells us that \(\tilde{\alpha}_{bi}\), \(\tilde{\alpha}_{ci}\) and \(\tilde{\alpha}_{di}\) that are given by our prior knowledge. Also assume that our prior knowledge tells us that \(\tilde{\alpha}_{ci} \neq \tilde{\alpha}_{di}\) so that these hypotheses do not coincide.

The Bayes factor is the ratio between the posterior odds and the prior odds of \(H_0\). If the Bayes factor is greater than one, then the observation of the data whatever that was.

In our case, the Bayes factor is

\[
\frac{P(X_{di} = x_{di}, X_{ci} = x_{ci}|H_0)}{P(X_{di} = x_{di}, X_{ci} = x_{ci}|H_1)} = \frac{P(X_{di} = x_{di}|H_0)}{P(X_{di} = x_{di}|H_1)} \frac{P(X_{ci} = x_{ci}|H_0)}{P(X_{ci} = x_{ci}|H_1)} \frac{\prod_j \frac{\Gamma(\sum_j x_{dij} + \alpha_{bij})}{\Gamma(\alpha_{bij})}}{\prod_j \frac{\Gamma(\sum_j x_{cij} + \alpha_{cij})}{\Gamma(\alpha_{cij})}},
\]

\[
\frac{\prod_j \frac{\Gamma(\sum_j x_{dij} + \alpha_{dij})}{\Gamma(\alpha_{dij})}}{\prod_j \frac{\Gamma(\sum_j x_{cij} + \alpha_{cij})}{\Gamma(\alpha_{cij})}},
\]

\[
\frac{\prod_j \frac{\Gamma(\sum_j x_{dij} + \alpha_{bij})}{\Gamma(\alpha_{bij})}}{\prod_j \frac{\Gamma(\sum_j x_{cij} + \alpha_{cij})}{\Gamma(\alpha_{cij})}}.
\]

(6.6)

In the last step we use that each integral is known, namely as the Dirichlet-multinomial distribution [6]. What we now have to consider is the choice of the parameter vectors.
One approach is to take a subset of the data to construct prior distributions of \( p_{ci} \) and \( p_{di} \), see e.g. Garthwaite et al. [33]. In our case we may begin with a Dirichlet\((1, 1, \ldots, 1)\) distribution as the prior distributions of \( p_{ci} \) and \( p_{di} \) for all \( i \). In this case all values for the random vectors \( p_{ci} \) and \( p_{di} \) are equally likely. Next we form a posterior distribution of \( p_{di} \) by updating the prior Dirichlet\((1, 1, \ldots, 1)\) distribution with part of the data from database \( d \). Since the data is assumed multinomially distributed, the posterior distribution will be Dirichlet distributed in the same way as in section 3.1. This posterior distribution is then used as the prior distribution \( g(p_{di} | \tilde{\alpha}_{di}) \) in equation (6.6). In the same way we find a prior distribution \( g(p_{ci} | \tilde{\alpha}_{ci}) \) for \( p_{ci} \). To find the prior distributions \( g(p_{di} | \tilde{\alpha}_{bi}) \) and \( g(p_{ci} | \tilde{\alpha}_{bi}) \) we temporarily regard the data from the subsets of database \( d \) and \( c \) as one large sample \( x_{bi} \) taken from a multinomial distribution. Updating the prior Dirichlet\((1, 1, \ldots, 1)\) distribution with this sample yields a posterior distribution which is used as the prior distributions \( g(p_{di} | \tilde{\alpha}_{bi}) \) and \( g(p_{ci} | \tilde{\alpha}_{bi}) \) in (6.6).

We try two different ways of selecting subsets of the databases. In the first investigation we take 20% of the DNA profiles from each database to construct the prior distributions \( g(p_{di} | \tilde{\alpha}_{di}) \), \( g(p_{ci} | \tilde{\alpha}_{ci}) \), \( g(p_{di} | \tilde{\alpha}_{bi}) \) and \( g(p_{ci} | \tilde{\alpha}_{bi}) \). The allele counts of the rest of the DNA profiles are then used as the observed data \( x_{di} \) and \( x_{ci} \) in (6.6). At this point we can calculate the Bayes factor and see whether the result points towards \( H_0 \) (if the Bayes factor > 1) or towards \( H_1 \) (Bayes factor < 1). To capture the sampling error associated with the selection of DNA profiles in the subset we perform the test 20 times for each locus.

The result when comparing the reference database and database A is given in Figure 6.7 using a \( \log_{10} \) scale. The ordering of loci are the same as in previous tables. See for example Table 6.1. Basically all logged Bayes factors are above zero, i.e. the Bayes factors are above one, supporting \( H_0 \) for all loci. The Bayes factors for locus number 5, D8S1179, and number 8, D19S433, seem to be the ones closest to zero on the logged scale with somewhat smaller support for \( H_0 \). D8S1179 and D19S433 also has quite large \( L_1 \) and \( L_2 \) distances when comparing the reference database with database A. A short inspection indicates that the agreement between the Bayes factors and the distance measures, regarding the ordering of discrepancy between loci, seems to be less than when using the likelihood ratio test. However, the way we measure the \( L_1 \) and \( L_2 \) distances are very different from the computing of Bayes factors and the P-values in the likelihood ratio tests.

Since the sizes of the databases are unequal, the number of profiles that are used to find prior distributions will be different between databases when taking a 20% subset as above. So in the second investigation we take 50 profiles from each database to construct the prior distributions. The result when comparing the reference database and database A is given in Figure 6.8 using a \( \log_{10} \) scale. All Bayes factors are above zero on the \( \log_{10} \) scale pointing towards the null hypothesis for all loci.

Similar conclusions are reached when comparing the reference database with database B. The result, when using subsets of 50 profiles for the priors, is shown in Figure 6.9.
Figure 6.7: Distribution of the Bayes factors on a $\log_{10}$ scale for all 10 loci between the reference database and database A. Subsets of 20% of the profiles are used to construct the priors.

Figure 6.8: Distribution of the Bayes factors on a $\log_{10}$ scale for all 10 loci between the reference database and database A. Subsets of 50 profiles are used to construct the priors.
6.2 Likelihood ratio discrepancies

If the underlying populations of the DNA databases are different this may be reflected in the likelihood ratios; we simulate a number of DNA profiles as we did in section 3.3 and regard each of these as a profile from a "suspect" that matches a hypothetical stain from a crime scene. We then calculate the LRs using the estimated allele proportions from each of the databases and observe what discrepancies there are. A visual comparison can be made if we plot the LRs obtained when using the estimated allele proportions from one database versus the LRs obtained when using the estimated allele proportions from another database. If we assume that the allele proportions are the same, then the points in the figure should lie along a straight line through the origin and with slope one. Due to the large span of LRs we use a log$_{10}$ scale.

In the simulations and in the calculations of the LRs we may use full or partial profiles (5 loci) and try some different values on $\theta$ (e.g. 0, 0.01, 0.02). Similar conclusions are reached for any of these settings so we focus the presentation of the results to a single case; partial profiles over 5 loci and a value of $\theta$ at 0.01.

In Figure 6.10 the LRs for 200 simulated DNA profiles over five loci are compared using estimated allele proportions from the reference database (LR1) and from database A (LR2). If there are no differences between the two databases then the points should lie along the diagonal line $LR1 = LR2$ that is shown in the figure. We then fit a line $\log_{10}(LR1) = \beta_0 + \beta_1 \log_{10}(LR2)$. This is done using a
method known as the orthogonal least square method\footnote{This is different from the ordinary least square method that minimizes the sum of the squared vertical differences between each observed value and the line, assuming that one of the variables is measured without error.} \cite{5}; we select the line that minimize the sum of the squared orthogonal differences between each observed value and the line. We see that both lines agree quite well on this interval but that the points show a certain degree of variation along the fitted line. The situation is basically the same when we compare the reference database to database B, see Figure 6.11. In Figure 6.12 the LRs are calculated using the estimated allele proportions from database A and database B respectively. As expected, both lines agree very well on this interval and the deviation of the points from the fitted line seems to be small.

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{figure6.10.png}
\caption{Comparison of LRs when using estimated allele proportions from the reference database (LR1) and from database A (LR2) respectively. Partial profiles over 5 loci are used. \(\theta\) is set to 0.01.}
\end{figure}
6.2. Likelihood ratio discrepancies

Figure 6.11: Comparison of LRs when using estimated allele proportions from the reference database (LR1) and from database B (LR2) respectively. Partial profiles over 5 loci are used. $\theta$ is set to 0.01.

Figure 6.12: Comparison of LRs when using estimated allele proportions from database A (LR1) and from database B (LR2) respectively. Partial profiles over 5 loci are used. $\theta$ is set to 0.01.
6.3 Most probable matching profile discrepancies

As a final comparison of the databases we examine if the most probable matching profile appear in expected numbers in the different databases, and if there are discrepancies between the databases.

Given the most probable matching profile $g_{mpmp}$ from the reference database between two unrelated individuals we assess the empirical number of matches against this profile for the profiles in the different databases. For each profile we examine how many genotypes that agree with $g_{mpmp}$ and then calculate the fraction of the individuals that match the profile $g_{mpmp}$ at exactly 0, 1, ..., 10 genotypes. See Figure 6.13. The distribution of matches seems to agree reasonably well for the different databases.

![Figure 6.13: Distributions of the relative frequency of matches against $g_{mpmp}$.](image)

In chapter 4 we saw how to calculate the match probability between two unrelated individuals given the profile of one of them. If we take this profile as $g_{mpmp}$ we may estimate how likely it is that an hypothetical unrelated individual, with the untyped profile $G_s$, will match $g_{mpmp}$ at exactly $l$ loci. The calculations of these probabilities, shown below, require that we examine a very large number of combinations of loci. We therefore restrict $l$ to a maximum of 6 loci in order to reduce the computing effort, and we choose a "scaled down" version of $g_{mpmp}$ consisting of the 6 loci with the highest match probabilities.

For $l = 0, \ldots, 6$, define the events

$$A_l = \text{The profile } G_s \text{ matches } g_{mpmp} \text{ at exactly } l \text{ loci}$$
6.3. Most probable matching profile discrepancies

\[ P(A_0 \cup A_1 \cup \ldots \cup A_6) = \sum_{l=0}^{6} P(A_l) = 1 \]  

(6.7)

Next we note that a match at exactly \( l \) loci can be done in \( n = \binom{6}{l} \) ways, denoted \( A_1^l, A_2^l, \ldots, A_6^l \), for \( l = 1, \ldots, 6 \). So

\[ P(A_l) = P(\text{exactly one of } A_1^l, A_2^l, \ldots, A_6^l \text{ occur}) = \]

\[ \sum_{p=1}^{n} P(A_p^l) - 2 \sum_{p=1}^{n-1} \sum_{q=p+1}^{n} P(A_p^l \cap A_q^l) + 3 \sum_{p=1}^{n-2} \sum_{q=p+1}^{n-1} \sum_{r=q+1}^{n} P(A_p^l \cap A_q^l \cap A_r^l) - \]

\[ \ldots (-1)^{n+1} \sum_{p=1}^{n} \cdots \sum_{z=n}^{n} P(A_p^l \cap \ldots \cap A_z^l) \]

(6.8)

by Råde and Westergren [34]. And \( P(A_0) = 1 - \sum_{l=1}^{6} P(A_l) \).

The probabilities \( P(A_p^l \cap \ldots \cap A_z^l) \) are then calculated as the product of the match probabilities for each locus involved. For example, if \( A_1^1 \) is a match at locus 1 and 2 and \( A_2^2 \) is a match at locus 1 and 3 then \( P(A_1^1 \cap A_2^2) \) is the match probability over locus 1, 2, and 3. Using the same framework as in section 4.2;

\[ P(A_1^1 \cap A_2^2) = \prod_{i=1}^{3} P(G_i^a = (a_{ij}, a_{ik}) | G_i^a = (a_{ij}, a_{ik}), H_d) \]

(6.9)

where \((a_{ij}, a_{ik})\) is the genotype from \( g_{\text{mpmp}} \) at locus \( i \). Formula (4.10) and (4.11) then applies with the estimated allele proportions from the reference database.

The resulting probabilities \( P(A_l) \) can then be interpreted as how large fraction of a number of hypothetical unrelated individuals that will match the most probable matching profile at exactly \( l \) loci. Let us refer to this group of individuals as the hypothetical population. The comparison to this population may indicate if the combination of the most common alleles and genotypes (those given by \( g_{\text{mpmp}} \)) appear in expected number in the different databases. In Figure 6.14 the relative frequencies of matches against \( g_{\text{mpmp}} \) at exactly \( l \) loci for the hypothetical population, the reference database, and database A and B are given. The results for the hypothetical population differ the most at 0 and 2 matches to the other databases, but there are no extreme discrepancies of the results in comparison to the different databases. Hence, by visual inspection we do not find any strong evidence against an assumption that combinations of common alleles and genotypes appear in expected numbers. The restriction to 6 loci instead of using all 10 loci has, in one aspect, little effect on the results because, as we see in Figure 6.13, there are no matches at 6 loci or above. However the results may alter if we take some other set of 6 loci. But some experimentation with other sets of loci than the one we use yield similar results in terms of similarities between the different distributions. In comparison to Figure 6.13 there are some differences in the relative frequencies of matches against \( g_{\text{mpmp}} \). This is because we have restricted the calculations to 6 loci while in the previous comparison we used all 10 loci. The similarities between the different distributions remain though.
Chapter 6. Database discrepancies

6.4 Summary

We investigate discrepancies and similarities between the reference database, which we have used earlier in this thesis, and two casework databases of DNA profiles at SKL that we refer to as database A and database B. The reference database consists of 205 full profiles, database A consists of 14983 full profiles and database B consists of 77055 full profiles. The three databases are regarded as samples from three populations and our task is to find out if these three populations are similar in aspect of their DNA profiles.

Using two distance measures, the Euclidean and the Manhattan, we obtain estimates of the distances between the allele proportions for each locus between the different databases. The two largest distances between the reference database and database A and B seem to be for loci D19S433 and D8S1179. The smallest distance seems to be for locus D2S1338. The comparison of database A and B reveal only minor distances with the largest being for locus TH01.

Next we perform statistical tests to find out if there are differences in the allele proportions that are larger than we expect if the differences are due to random effects only. For the first test method, the likelihood ratio test, we do not find any significant differences between the reference database and database A. There are no significant differences between the reference database and database B in the pairwise comparisons of loci. One method for combining the tests results indicate that the overall differences may not be explained by random effects only. Significant differences are found for all loci except for locus vWA when comparing database A to database B. However, the test is sensitive to the large sample sizes of database A and B and the significant differences are probably
Another way of telling if there are any major differences between the allele proportions for the different databases is by constructing a set of probable values of the differences in allele proportions and see whether the zero vector is contained within this set of values. In the comparison between the reference database and database A, the only set that does not include zero is for loci D19S433 and D8S1179. The same result is reached when comparison is made between the reference database and database B.

A third testing strategy using so called Bayes factors is applied. We cannot find any convincing evidence against the hypothesis that the sampled allele data have arisen from similar populations when comparing the reference database to either database A or database B.

Next we generate a number of DNA profiles and compute LRs using estimated allele proportions from one of the databases with the LRs using estimated allele proportions from one of the other databases. If there are no differences between the databases then the two sets of LRs should have an equal distribution. A visual inspection of the distribution of LRs between the reference database and either database A or B shows that the values are reasonably similar. A comparison between database A and database B shows a very good agreement between the distributions of LRs.

In the last comparison we investigate if the relative frequency of full or partial matches against the most probable matching profile between unrelated individuals are similar in the different databases. A visual inspection indicates that the agreement between the different databases is quite good. Further we examine if the combination of the most common alleles and genotypes appear in expected numbers in the different databases. No clear evidence pointing in the other direction is found.
Let us return to the issues that were addressed in the introduction of the thesis and make a summary of the main results along with some further conclusions;

- How are population genetic factors taken into account into the process of estimating the strength of DNA evidence?

In chapter 3 and 4 we investigated some concepts in population genetics that are important in the process of estimating the strength of DNA evidence; population allele proportions, inbreeding coefficients and tests of Hardy-Weinberg’s law and linkage equilibrium. The tests of Hardy-Weinberg’s law and linkage equilibrium are valuable mainly for finding indications of genotyping or data entry errors. The tests presented are more appropriate for the data at hand than earlier tests applied at SKL. A different way of testing for deviations of Hardy-Weinberg’s law is mentioned by Buckleton et al. [11]; that is, by estimating $\theta$. However, this requires data sampled from different subpopulations which may be a topic for future investigation. In the calculation of the strength of DNA evidence against a suspect it is assumed that there is random mating within subpopulations so that the $f$ coefficient is equal to zero. Further, there is an uncertainty that the allele proportions from the database population are the same as in the suspect’s subpopulation since we generally do not have genotype data from the required subpopulation. This uncertainty is equal to $\theta$. It was stated in chapter 4 that a large value of $\theta$ is in favor of the suspect and a reasonable value of $\theta$, in the light of Norwegian and English population studies, might be 0.01.

- How is the strength of DNA evidence affected when the suspect put the blame on some of his or her close relatives?

In chapter 4 we examined two methods for combining hypotheses given by the defense. Both of these are applications of Bayes’ theorem. The second of these two is only briefly mentioned in the references that have been used but here a more thorough examination of it was given. The result from this method is a LR, which is the common choice for assessing the strength of DNA evidence. Further, the method does not need a prior opinion of the suspect’s probability of being the donor of the stain which might be more acceptable to a court. The drawback of these two methods is that you (read: the court) must give assessment of prior probability ratios for being the donor of the stain for some
groups of individuals but this may be a difficult task. However, the consequence of the choice of prior ratios for both of these methods, may be illustrated with graphs such as Figure 4.3 that may be helpful. Also the prior probability ratios for which the LR reaches a new level on the scale of conclusion can be useful. In the thesis only simple situations were demonstrated. Situations where there are a large number of groups of possible donors of the stain may be harder to illustrate with graphs etc. More work on how to de facto apply these methods in a trial may be an issue for further examination.

- What DNA profiles minimizes the strength of DNA evidence against a suspect under different scenarios? How low is the strength and how is it affected by errors in the estimation procedure?

In chapter 5 we introduced the most probable matching profile, \( g_{mpmp} \), as the profile that minimizes the LR under a specific hypothesis given by the defense. A tool in Excel was developed that found these profiles and their corresponding LR. Some of the results were presented in Tables 5.1-5.5. The LRs of \( g_{mpmp} \) may be used as general lower bounds for each relationship. Further, we investigated how these results are affected by sampling error, i.e. the error due to the fact that we only took a sample of DNA profiles instead of typing the whole population. Of course sampling error is unavoidable in all investigations where a random sample is taken but it might be a difficult task to measure this error. Different methods are suggested in the literature and we have seen two of them applied to the LR of \( g_{mpmp} \): the bootstrap method and a Bayesian method. Here, the effect of the sampling error is most interesting for LRs just above or below some level on the scale of conclusions. In some cases the range of plausible values of the LR may imply that two different levels on the scale of conclusions may be justified by the DNA evidence. The range of plausible values for the LR will differ from case to case but as we have seen the effect of the sampling error is not insignificant for cases where the LR is close to another level on the scale of conclusions. However, conservativeness in the calculations of the LRs at SKL are adopted by applying a reasonably large value of \( \theta \) and by using minimum allele proportions (allele proportion below 2% are set to 2%). This reduces some of the impact of the sampling error in many cases. Further, as discussed below, the sampling error is expected to be reduced if a larger sample size is used in the construction of a new reference database.

- What similarities and differences are there between the different databases of DNA profiles at SKL?

In chapter 6 we examined the differences in allele proportions between the different databases using both descriptive and inferential statistics. We found little evidence of major differences of both practical and statistical significance between the databases. The discrepancies between database A and B seemed to be very small, while the discrepancies between the reference database and database A or B was a little bit larger. There were some indications of discrepancies in allele proportions between the reference database and database A and B at locus D19S433 and D8S1179. More research to validate this and to find reasons for it can probably be done. One limitation here is that we tested if the allele proportions are exactly the same, which they of course are not. By the Bayesian method we found no convincing evidence against the hypothesis that
the sampled allele data have arisen from similar populations when comparing the reference database to either database A or database B. We must note here that this method is quite sensitive to the choice of prior values as we saw in section 6.1.3.

If we would use either database A or B as reference database we saw that the calculated LRs would be reasonably similar to the LRs given when using the true reference database. Also, the agreement of partial matches against the most probable matching profile was quite high. Further discussions of if the data in database A and B really can be considered as a legitimate random sample of profiles can also be done but this requires more insight of these databases not brought up here.

7.1 More loci

In a near future the number of loci in the DNA typing at SKL will increase from 11 to 16, including the gender indicating locus. In the construction of the new reference database the recommended statistical tests of Hardy-Weinberg’s law and linkage equilibrium in section 3.2.1 may be used in the validation process. Naturally, with more loci a match between a stain and a suspect will yield even larger LRs than before. This affects some of the results in this study but the principles remain the same. Also, since we sometimes end up with partial profiles the LRs may not be that large after all in those cases.

In chapter 4 we considered the situation where e.g. full siblings to the suspect were candidates of being the true donor of the crime stain. According to Buckleton and Triggs [35] the relative balance of the contribution to the strength of DNA evidence has changed with more loci introduced in the DNA typing process. The few related individuals considered have a relatively larger impact on the strength of DNA evidence than the many unrelated individuals than before. In the article this is exemplified by using formula (4.15) with the assumption that the suspect has a full sibling of the same gender. Further it is assumed that these two have the same prior probability as the average of 125 million unrelated individuals. Then a simple calculation tells how large fraction of the remaining posterior probability that is due to the sibling’s presence when we for a moment ignore the suspect’s impact on the posterior probability. When 16 loci are used basically 100% of the remaining probability is due to the full sibling. Hence, the question of how to combine several hypotheses presented by the defense is becoming more important. Returning to section 4.4, we can expect that typing 16 loci instead of 11 loci will make it much more unlikely that someone else, e.g. a full sibling or an unrelated individual will have the same profile as the suspect. So in the examples given even larger minimum prior ratios will be required between a sibling and an unrelated person to drop down from one level on the scale of conclusion to another one.

The LRs of the \( g_{mpmp} \) in section 5.1 will become larger. For full profiles the minimum LR will always be at +4 except perhaps for full siblings as indicated in 5.1.1. Also the situation for partial profiles has to be investigated, which can easily be done using the Excel tool. If a larger sample size is used in
the construction of the new reference database than before then the sampling error is expected to be reduced, and since we introduce more loci there will probably be less cases where the LR is just above or below some level on the scale of conclusions because we often end up with LRs that is vastly greater than required for a +4 conclusion.

As for database discrepancies it is harder to tell how the results may be affected. Other sample sizes and new profiles will probably alter the results to some degree, although the size and direction is difficult to guess upon.
Bibliography


Copyright

The publishers will keep this document online on the Internet - or its possible replacement - for a period of 25 years from the date of publication barring exceptional circumstances. The online availability of the document implies a permanent permission for anyone to read, to download, to print out single copies for your own use and to use it unchanged for any non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional on the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility. According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement. For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its WWW home page: http://www.ep.liu.se/

Upphovsrätt

Detta dokument hålls tillgängligt på Internet - eller dess framtida ersättare - under 25 år från publiceringsdatum under förutsättning att inga extraordinäre omständigheter uppstår. Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns det lösningar av teknisk och administrativ art. Upphovsmannens ideella rätt innefattar rätt att bli nämnt som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannens litterära eller konstnärliga anseende eller egenart. För ytterligare information om Linköping University Electronic Press se förlagets hemsida http://www.ep.liu.se/

© 2010, Ronny Hedell