Technische Universiteit Delft
Faculteit Elektrotechniek, Wiskunde en Informatica
Delft Institute of Applied Mathematics

Familial searching in DNA databases

(Nederlandse titel: Verwantschapsonderzoek in DNA databases)

Verslag ten behoeve van het
Delft Institute of Applied Mathematics
als onderdeel ter verkrijging

van de graad van

BACHELOR OF SCIENCE
in
TECHNISCHE WISKUNDE

door

MARIEKE DE VRIES

Delft, Nederland
Juli 2016

BSc verslag TECHNISCHE WISKUNDE

"Familial searching in DNA databases"
(Nederlandse titel: "Verwantschapsonderzoek in DNA databases")

MAAIKE MARIEKE DE VRIES

Technische Universiteit Delft

Begeleider

Dr. M.T. Joosten

Overige commissieleden

Drs. E.M. van Elderen

Dr. ir. F.H. van der Meulen

Dr. W.M. Ruszel

Juli, 2016

Delft

Abstract

We start our introduction to DNA matching with the direct match case: we find DNA at a crime scene and find a direct match with a member of our database. To measure the strength of our case against the match, we use likelihoodratios. We differentiate between two different situations: the probable cause case, where we already suspect our match and DNA testing tells us that the profile case and our suspect's DNA matches, and the database match case, where we find a match by comparing our case profile to the DNA profiles in our database. We pose the question: which one of these situations is more incriminating for our match?

The controversy around answering this question is called the database controversy, which is based on the fact that the chosen set of hypotheses, even when logically equivalent, changes the answer of this question. We show that this controversy is not really a controversy by using posterior odds instead of likelihood ratios.

Familial searching is not as straightforward: we discuss two methods for finding relatives of a case profile in a database. One of those is the conditional method, where we look at the minimum subset of members such that the sum of their prior probability times their likelihoodratio is larger than that of the whole database times a factor. The other is called the profile-centred method, where we look at the members in our database that have a likelihood ratio larger than a certain treshold.

We discuss theoretical efficiency (the probability that a family member is included in the subset when it is present in our database) and interpretation, but also model these methods in a simulated database. We look at the probability of detection when a relative is present in our database, and look at the typical size of our subsets for certain efficiencies. Both methods have their advantages and disadvantages, and there is no clear answer on which method is better: this should be decided on a case by case basis.

Preface

Before you lies the thesis *Familial searching in DNA databases*, written in order to obtain the degree of Bachelor of Science from the Delft Institute of Applied Mathematics. This project took place in the department of statistics under the supervision of M.T. Joosten, and was suggested by G. Jongbloed and M.T. Joosten.

During this last quarter I spent a lot of time researching the mathematics of DNA searching and working in R to simulate the described methods. Not everything I researched turned out the way I hoped, and had I had more time I would have loved to spent some more months researching and simulating - but alas, everything comes to an end.

I would like to thank M.T. Joosten for his guidance and help during the project and F.H. van der Meulen, W.M. Ruszel and E.M. van Elderen for taking part of my assessment committee. Lastly, I would like to thank P. Benedysiuk for his feedback on my R code and thesis and P.H. Kanters and M.A. de Vries for supporting me, providing feedback and keeping me sane.

Marieke de Vries
July, 2016

Contents

1 Introduction

Although ideas and practices of the law have been around since the archaic period, developments in science and society still provide new tools to solve crimes and look for justice. One of those 'new' tools is DNA testing, which has led to many convictions ever since its first use in 1986 with the conviction of Colin Pitchfork [1], who had raped and strangled two 15-year-old girls. Initially, the prime suspect was Richard Buckland, a boy from the village. However, after comparing his DNA with that found on the girls there was no match found and thus his innocence could be proven. Coincidentally, Buckland was also the first to have his innocence established by the use of DNA matching. It took an extensive investigation of about 500 DNA samples from the other men from the village before Pitchfork was arrested and convicted.

Likewise, analysing DNA of old cases has led to some people being released from prison (sometimes after spending more than two decades behind bars), as the DNA evidence could establish their innocence. DNA testing and DNA searching in databases are very powerful tools for law enforcement, but there are certain risks when working with this kind of evidence in a case. In this thesis we will be focusing on DNA searching in databases, and mainly on the notion of familial searching.

There has been much debate and many papers written about DNA searching in databases, although much of this research is done in the field of sociology and biology. However, there has been some interesting mathematical research into DNA searching in databases and familial searching. Papers by Meester and Sjerps [2] and Slooten and Meester [3] provide examples of such research, upon which this thesis expands.

We will first take a look at the basic statistics behind DNA searching in databases, where we look into the database controversy and why it is not really a controversy, before delving into familial searching. The main focus of this thesis are the methods used to find relatives of a case profile in a database: the conditional method and the profile-centred method. We first analyse these two methods theoretically and afterwards simulate them in R to compare their performance. And, as always when comparing two methods, we will try to answer the question of which method is better.

2 Some background for statistics

Before delving deeper into the mathematics of DNA searching, we first need to look at some fundamental statistics that we will be using throughout this thesis. First and foremost, the usage of the word *odds*.

Definition 2.1. *For an event A, with $\mathbb{P}(A)$ the probability of A occuring, we define the* ***odds*** *of A occuring as*

$$odds(A) = \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)}$$

Example 2.2. When talking about odds, we mean to express the likelihood of a certain event happening. For example, when rolling a fair six-sided die, one expects to roll a six about one in six times and not six about five in six times; the *odds* of rolling a six are 1 to 5. We will be using the term odds when talking about Bayesian statistics, usually when talking about *prior* and *posterior odds.*

For example, if I were to roll with a fair die or a loaded die (loaded in favour of rolling six), the odds of rolling six is higher for the loaded die than for the fair die. We now take one of the two die and roll it twenty times, and get six about 7 times. Let's say the probability of rolling a six for the loaded dice is $\frac{1}{2}$ and $\frac{1}{10}$ for the other sides. Then the probability of rolling seven sixes is $\binom{20}{7}\frac{1}{2}^{20}$ for the loaded die, whereas for the fair die the probability is $\binom{20}{7}\frac{1}{6}^{7}\frac{5}{6}^{13}$.

The odds of this result being produced by the loaded die is now $\frac{1}{2}^{20}\frac{1}{6}^{-7}\frac{5}{6}^{-13}$ or roughly 3, which means that the loaded die is three times more likely to produce this result. To phrase this more formally, let H_L be the hypothesis that we rolled with the loaded die, H_f that we rolled with a fair die and X our data (i.e. we have rolled 7 sixes in 20 rolls). ■

Definition 2.3. *For any dataset X and any set of two relevant hypotheses (H_0, H_1) relating to the data, we define the corresponding* ***likelihood ratio*** *as*

$$LR = \frac{\mathbb{P}(X|H_0)}{\mathbb{P}(X|H_1)}$$

Example 2.2. (Continued) In our case the likelihood ratio, or the odds, are calculated by $\frac{\mathbb{P}(\text{We throw 7 sixes}|H_L)}{\mathbb{P}(\text{We throw 7 sixes}|H_f)}$, which is roughly 3 (as we have already calculated before). Now that we have the likelihood ratio, we want to know with which die we just rolled: we want the odds of having chosen the loaded die against having chosen the fair die. These also depend on how likely we were to choose one die over the other. We can express these odds by dividing $\mathbb{P}(H_L|X)$ by $\mathbb{P}(H_f|X)$. We will need to use Bayes' theorem ■

Theorem 2.4 (Bayes' Theorem)**.** *For any two events A, B with $\mathbb{P}(B) \neq 0$ we can write*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

Proof. By definition of conditional probabilities, we have $\mathbb{P}(A|B) = \frac{\mathbb{P}(A\bigcap B)}{\mathbb{P}(B)}$ and $\mathbb{P}(B|A) = \frac{\mathbb{P}(A\bigcap B)}{\mathbb{P}(A)}$, so $\mathbb{P}(A\bigcap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$. Thus $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$. □

Definition 2.5. *For any dataset X and any set of two relevant hypotheses (H_0, H_1) relating to the data, we call $\mathbb{P}(H_i|X)$ the* ***posterior probability*** *for H_i. The* ***posterior odds*** *are calculated by*

$$\frac{\mathbb{P}(H_0|X)}{\mathbb{P}(H_1|X)}$$

The ***prior probability*** *for H_i is simply $\mathbb{P}(H_i)$, giving us the* ***prior odds***

$$\frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}$$

Example 2.2. (Continued) Using the Bayes' theorem, we can rewrite the probabilities $\mathbb{P}(H_L|X)$ and $\mathbb{P}(H_f|X)$ to

$$\mathbb{P}(H_L|X) = \frac{\mathbb{P}(X|H_L)\mathbb{P}(H_L)}{\mathbb{P}(X)} \tag{1}$$

And vice versa for $\mathbb{P}(H_f|X)$. In this case, $\mathbb{P}(H_L)$ is the probability that we took the loaded die, without looking at the result of our rolls. If we choose one of the two dice arbitrarily, this probability would be $\frac{1}{2}$ since the prior distribution would be uniform over the dice. We now divide the two probabilities, which gets us the following relation

$$\frac{\mathbb{P}(H_L|X)}{\mathbb{P}(H_f|X)} = \frac{\mathbb{P}(H_L)}{\mathbb{P}(H_f)}\frac{\mathbb{P}(X|H_L)}{\mathbb{P}(X|H_f)} \tag{2}$$

The first fraction, $\frac{\mathbb{P}(H_L|X)}{\mathbb{P}(H_f|X)}$, is called the posterior odds. It expresses the odds in favour of H_L, given our data. The fraction $\frac{\mathbb{P}(H_L)}{\mathbb{P}(H_f)}$ is called the prior odds, as they are the odds in favour of H_L *prior* to looking at the data. This relation holds in general, so for any dataset and two relevant hypotheses we can compare and compute these odds. ■

Theorem 2.6. *For any dataset X and any set of two relevant hypotheses (H_0, H_1) relating to the data for which we can compute the relevant prior odds, posterior odds and likelihoodratio, we have:*

Posterior odds = prior odds · likelihood ratio

Proof. As shown above, we can rewrite the posterior probabilities $\mathbb{P}(X|H_i)$ to

$$\mathbb{P}(H_i|X) = \frac{\mathbb{P}(X|H_i)\mathbb{P}(H_i)}{\mathbb{P}(X)}$$

which gives us, by dividing the two probabilities, the relation we want

$$\frac{\mathbb{P}(H_0|X)}{\mathbb{P}(H_1|X)} = \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} \cdot \frac{\mathbb{P}(X|H_0)}{\mathbb{P}(X|H_1)}$$

□

we will be using this result later on. However, we will look at the application of likelihood ratios when looking at DNA database searching first.

3 Likelihood ratios in DNA matching

Because a DNA match doesn't always mean that the person that matched is the same as the person whom the DNA belongs to, one should be careful when analysing the results of a DNA search in a certain database. This is why we use the concept of likelihood ratios to quantify how likely a person is to actually be the donor of the DNA in question. As shown previously: given two hypotheses H_0 and H_a and certain data, we can calculate the likelihood of the appearance of this data, given the two hypotheses, by

$$LR = \frac{\mathbb{P}(data|H_0)}{\mathbb{P}(data|H_a)} \tag{3}$$

If this likelihood ratio is high, it is an indication that the hypothesis H_0 is more likely than H_a. We can apply this likelihood ratio to the direct database searching case, where we have a DNA sample found at the crime scene (we call the donor C) and found a single match in a database, namely M. We want to know how likely it is that the donor and the match in the database are the same people, and can construct a set of hypotheses to test the relevance of the match. For example, we can use the hypotheses $H_0 : M = C$ and $H_a : M \neq C$, giving the likelihood ratio

$$LR = \frac{\mathbb{P}(M \text{ is the only match}|H_0)}{\mathbb{P}(M \text{ is the only match}|H_a)} \tag{4}$$

which we can calculate depending on the fraction of people that have the DNA profile C, the size of the database and population and some prior probabilities. However, when law enforcement is unable to find a direct match in a database, they sometimes get permission to search for family of the DNA donor in the database. This is called familial searching, which is more complicated than a direct-match case, both on a mathematical and judicial level. We will concern ourselves with the mathematical difficulties and results.

3.1 Familial searching

So let's say we find DNA at the crime scene that we assume belongs to the offender. This gives us a case profile, say C, that then doesn't result into a direct match. Because DNA of family members is more similar to C than that of an arbitrary stranger to C, we can try to find a partial match in the DNA database D which will possibly give us a match with a family member. Similar to the direct database searching case, we can construct likelihood ratios for each person in the database, where we now try compute how likely a person is to be related to the offender.

We do this by constructing two random variables, S and G, both a distribution over DNA profiles, where S is distributed as the profile of a relative of the case profile, and G is distributed as an arbitrary DNA profile. We can now interpret the entries of the database D as realisations of either S or G, and by performing specific search strategies we hope to single out the realisations of S. We say $D = \{e_1, \ldots, e_N\}$ where e_i is a DNA profile of person i in D, and call R the index of the person related to the offender. Then the likelihood ratio

$$LR_i = \frac{\mathbb{P}(e_i|R = i)}{\mathbb{P}(e_i|R \neq i)} = \frac{\mathbb{P}(S = e_i)}{\mathbb{P}(G = e_i)} \tag{5}$$

tells us something about how likely person i is to be related to our case profile. If we're searching for a parent or a child, we often call it the paternity index PI; If we look for siblings, we often refer to it as the sibling index SI. After calculating the respective likelihood ratios, we would like to create a small enough subset of DNA profiles so that with a high degree of certainty it contains a relative of C, but small enough to still be manageable to investigate further. We will discuss this further in chapter 4.

3.2 The choice for likelihood ratios

Before we delve deeper into familial searching, we first need to pay some more attention to the methods that were just described. The use of likelihood ratios can at first seem a bit arbitrary and will certainly need some motivation. The most important part of using a tool to measure the likelihood of the suspect being guilty, is that it is consistent and as little debatable as possible. However, we cannot claim that likelihood ratios are just that, without taking a look at what is called the *database controversy.* We will be following the approach of Meester en Sjerps [2].

3.2.1 Database controversy

To explain the controversy, we look at the case where we find a DNA stain of some sort at the crime scene, which we assume belonged to the culprit. When we then compare the DNA profiles in the database with the profile we found at the crime scene, we find only one match M. We now pose the question: is the case against M stronger now that we have found him through DNA profiling, or is the case where we would find him through other evidence - the so called *probable cause* scenario - stronger?

As before, we use likelihood ratios to quantify how strong our case is. We define our set of hypotheses as follows:

$$\begin{aligned} H_0 &: \quad M \text{ is the donor of the crime stain} \\ H_a &: \quad \text{Someone else is the donor of the crime stain} \end{aligned}$$

And so our likelihood ratio becomes

$$LR = \frac{\mathbb{P}(\text{M is the only match}|H_0)}{\mathbb{P}(\text{M is the only match}|H_a)} \tag{6}$$

We define δ as $\mathbb{P}(M$ left the crime stain), π as $\mathbb{P}$(someone else in the database left the crime stain), n as the population size and N as the size of the database. We assume the DNA stain has profile A that has a certain frequency p_A in our population and also assume that the profiles in our database are independent of each other. We then can calculate the probabilities in the likelihood ratio: $\mathbb{P}(\text{M is the only match}|H_0)$ is the probability that only M has DNA profile A (which is 1 since M is the donor of the crime stain under H_0) and the other people in the database have another profile. Thus, as the database size is N, the probability of none of the other profiles having profile A is $(1 - p_A)^{N-1}$.

For $\mathbb{P}(\text{M is the only match}|H_a)$, we look at the probability that only our match M has profile A but none of the others, (which results in $p_A(1-p_A)^{N-1}$ since M is no longer the donor of the stain and is not guaranteed to have profile A), and that no one in the database (including our match M) has left the crime stain. Using the rules for conditional probabilities, we can rewrite the probability to

$$\begin{aligned}
\mathbb{P}(\text{M is the only match}|H_a) &= \frac{\mathbb{P}(\text{M is the only match, } H_a)}{\mathbb{P}(H_a)} \\
&= \frac{\mathbb{P}(\text{M is the only match})\mathbb{P}(H_a|\text{M is the only match})}{1-\delta} \\
&= p_A(1-p_A)^{N-1}\frac{1-\pi}{1-\delta}
\end{aligned}$$

Since if M is the only match in the database, none of the other people in the database can be the donor of the crime stain: thus, the probability of someone other than M leaving the crime stain is equal to the $1-\pi$, the probability that someone outside of the database left the crime stain. The resulting likelihood is

$$\begin{aligned}
LR &= \frac{\mathbb{P}(\text{M is the only match}|H_0)}{\mathbb{P}(\text{M is the only match}|H_a)} \\
&= \frac{(1-p_A)^{N-1}}{p_A(1-p_A)^{N-1}\frac{1-\pi}{1-\delta}} \\
&= \frac{1}{p_A}\frac{1-\delta}{1-\pi}
\end{aligned}$$

The question remains whether this means the case is stronger than in the probable cause scenario. To see whether it is we will have to compute the relevant likelihood ratios. So let's say M is identified using non-DNA based methods and after a check, his DNA matches the one of the perpetrator. We now have the same two hypotheses as before, meaning that in H_0 we assume the suspect M is the same as the perpetrator, and in H_a we assume the DNA of M and the perpetrator are stochastically independent. Calculating the likelihood ratio gives us:

$$\begin{aligned}
LR &= \frac{\mathbb{P}(M \text{ has profile A, the crime stain has profile A}|H_0)}{\mathbb{P}(M \text{ has profile A, the crime stain has profile A}|H_a)} \\
&= \frac{\mathbb{P}(M \text{ has profile A})}{\mathbb{P}(M \text{ has profile A})\mathbb{P}(\text{the crime stain has profile A})} \\
&= \frac{p_A}{p_A^2} = \frac{1}{p_A}
\end{aligned}$$

Compared to the likelihood ratio in the database searching case, they differ only a factor $\frac{1-\delta}{1-\pi}$. If we assume uniform prior odds, i.e. $\delta = \frac{1}{n}$ and $\pi = \frac{N}{n}$, we have that the likelihood ratios differ a factor of $\frac{n-1}{n-N}$ which is larger than 1 (but only slightly, especially when taking into account that we have a very large population size compared to the database size). Thus it would seem that the likelihood ratio in the database searching case is (slightly) larger than in the probable cause scenario, which would make the case against our match M stronger if he were to be found via DNA profiling.

However, this result has been critized by Stockmarr [4] because the chosen set of hypotheses is data-dependent and therefore, in his opinion, wrong: you can only choose these hypotheses if you already know that M will be the person in the database that matches with our crime stain. In fact, you cannot know that there will be only one match if you haven't received the results of the test yet. So he proposed the following set of hypotheses

$$\begin{aligned} H_0^* &: \text{Someone in the database is the donor of the crime stain} \\ H_a^* &: \text{Someone outside of the database is the donor of the crime stain} \end{aligned}$$

To calculate our new likelihood ratio, we first look at $\mathbb{P}(r$ profiles have profile A, the crime stain has profile $A|H_0^*)$. Under H_0^*, one of the people in the database the donor of the crime stain, so there are $r-1$ people that match by accident. Because every person either has profile A or not, we can consider the amount of people that match to be a binomial random variable distributed like $\text{Bin}(N, p_A)$ with N the size of the database. However, under H_0^* we know one of the people in the database is certainly the donor of the crime stain, meaning that for only $N-1$ people it is uncertain whether they will match or not. This means the number of accidental matches is distributed like a binomial random variable with sample size $N-1$.

Now we calculate: $\mathbb{P}(r$ profiles have profile A, the crime stain has profile $A|H_0^*) = \mathbb{P}(r-1$ profiles are accidental matches$|H_0^*,$ the crime stain has profile $A)\mathbb{P}($The crime stain has profile $A) = \binom{N-1}{r-1}p_A^{r-1}(1-p_A)^{N-r} \cdot p_A = \binom{N-1}{r-1}p_A^r(1-p_A)^{N-r}$.

Under H_a^* our sample size is the same as our database size, since the donor of the crime stain is no longer in our database. This gives us $\mathbb{P}(r$ profiles have profile A, the crime stain has profile $A|H_a^*) = \mathbb{P}(r$ profiles have profile $A|H_a^*,$ the crime stain has profile $A)\mathbb{P}($the crime stain has profile $A) = \binom{N}{r}p_A^r(1-p_A)^{N-r}p_A = \binom{N}{r}p_A^{r+1}(1-p_A)^{N-r}$. The resulting likelihood ratio now becomes

$$\begin{aligned} LR^* &= \frac{\mathbb{P}(r \text{ profiles have profile } A, \text{ the crime stain has profile } A|H_0^*)}{\mathbb{P}(r \text{ profiles have profile } A, \text{ the crime stain has profile } A|H_a^*)} \\ &= \frac{\binom{N-1}{r-1}p_A^r(1-p_A)^{N-r}}{\binom{N}{r}p_A^{r+1}(1-p_A)^{N-r}} = \frac{r}{Np_A} \end{aligned}$$

This results in a different likelihood ratio than before, since we now assume we have only one match and thus r = 1. Thus the likelihood ratio is $\frac{1}{Np_A}$, which is obviously smaller than the previously mentioned likelihood ratio in the probable scenario case. This would mean that the case against our match would be in fact weaker when compared to the other case. The last likelihoodratio is in case of uniform priors, but for the general case we can use the following calculation

$$\begin{aligned} \mathbb{P}(M \text{ is the only match}|H_0^*) &= \frac{\mathbb{P}(M \text{ is the only match}, H_0^*)}{\mathbb{P}(H_0^*)} \\ &= \frac{\delta(1-p_A)^{N-1}}{\pi} \end{aligned}$$

Since $\mathbb{P}(H_0^*) = \pi$ per definition and if we know that the donor of the crime stain is in the database, the chance of only M being matched is the chance of M being the donor and the other people in the database not having profile A. Now for $\mathbb{P}(M$ is the only match$|H_a^*)$ we have that by accident M's profile matches with our perpetrators and the others do not have this profile, so $\mathbb{P}(M$ is the only match$|H_a^*) = p_A(1-p_A)^{N-1}$. This gives us the likelihoodratio

$$\begin{aligned} LR^* &= \frac{\mathbb{P}(M \text{ is the only match}|H_0^*)}{\mathbb{P}(M \text{ is the only match}|H_a^*)} \\ &= \frac{\frac{\delta(1-p_A)^{N-1}}{\pi}}{p_A(1-p_A)^{N-1}} = \frac{\delta}{\pi p_A} \end{aligned}$$

3.2.2 Why it is a false controversy

We have now found two mutually exclusive answers to our question, both of which seem valid, resulting in our afore-mentioned controversy. However, we can show that this controversy is false by using the *posterior odds*. As shown previously, by using Bayesian statistics we have posterior odds = prior odds $\cdot$ likelihood ratio, and so we can look at the prior odds in both instances to find out what the posterior odds are.

For the first set of hypotheses, we have the prior odds $\frac{\mathbb{P}(H_0)}{\mathbb{P}(H_a)}$ with $\mathbb{P}(H_0) = \mathbb{P}(M$ left the crime stain$) = \delta$ versus $\mathbb{P}(H_a) = \mathbb{P}(M$ did not leave the crime stain$) = 1-\delta$, giving us the prior odds $\frac{\delta}{1-\delta}$. Multiplying this with the likelihood ratio associated to this set of hypotheses, we get the posterior odds

$$\frac{\delta}{1-\delta}\frac{1-\delta}{p_A(1-\pi)} = \frac{1}{p_A}\frac{\delta}{1-\pi} \tag{7}$$

As for the second set of hypotheses, H_0^* versus H_a^*, we have $\mathbb{P}(H_0^*) = \mathbb{P}($someone in the database left the crime stain$) = \pi$ and $\mathbb{P}(H_a^*) = \mathbb{P}($someone outside of the database left the crime stain$) = 1-\pi$, resulting in the prior odds of $\frac{\pi}{1-\pi}$. Multiplying with the likelihood ratio yields the following posterior odds

$$\frac{\pi}{1-\pi}\frac{\delta}{p_A\pi} = \frac{1}{p_A}\frac{\delta}{1-\pi} \tag{8}$$

which is equal to the previous posterior odds. Thus, contrary to what we have seen before, the posterior odds actually are consistent across both sets of hypotheses, even though the likelihood ratios are not. It is obvious now that there is no one set of hypotheses that is wrong (nor exclusively right) and we can use both sets of hypotheses when trying to express the strength of the case, as long as we also report the prior (and posterior) odds.

For example, consider the following: a person is murdered and DNA is recovered from under the fingernails of the murderer. The match probability of this DNA profile is around 10^{-5} and we have a database of around 1000 people. The result of our search in the database leaves us with one match: M. The likelihood ratio as suggested by Stockmarr is 10 while the likelihood ratio we suggested first is around 10^5. This is a huge difference, but computing the relevant prior odds will, as discussed before, show that the posterior odds are equal.

One method to compute these prior odds, as suggested by Meester and Sjerps, is by using δ to compute π by assuming that everyone in the population is equally likely to be the donor of the crime stain. Thus

$$\pi = \delta + (1-\delta)\frac{N-1}{n-1}$$

with n the size of the population. In court, this population size can be discussed or the prior and posterior odds can be calculated for multiple values of n. Meester and Sjerps also suggested reporting a table with prior odds and their posterior odds for the case at hand, so the juror can judge the strength of the other evidence and make a choice for the prior odds, which will give him the posterior odds. For high prior odds, the posterior odds will be even higher and thus the case is judged to be pretty strong. However, if the juror believed the other evidence is weak, he can choose a low prior odds, causing the posterior odds to decrease and making the case weaker.

So, though the strength of the DNA evidence remained the same in the two scenarios, the strength of the case was determined by the prior odds. It shows that the strength of the case against our suspect, expressed by the posterior odds, can be high even though our DNA evidence, the strength of which is expressed by the likelihood ratios, does not have to be strong - and our case can be weak if we only have a strong DNA match but no other evidence.

And so, even though at first it seemed that likelihood ratios were not as reliable as we thought, we see that the posterior odds are consistent across all cases and hypotheses. This is what we want, seeing as we looked for a consistent and definite expression for the strength of our case. When looking at the likelihood ratios, it is always wise to look at the relevant prior odds, as they might show that a strong match but low prior odds (usually in a DNA searching case) can be just as incriminating for the match as a low likelihood ratio and high prior odds (often in the probable cause case).

4 Search strategies for familial searching

We already discussed in section 3.1 how we use likelihood ratios when looking for relatives of a specific profile. It is important for law enforcement to narrow down the suspect pool with possible family members in order to make the investigation manageable, but if the conditions are too strict this might possibly lead to an exclusion of an actual family member. We will discuss two strategies for creating a subset of the database, where we try to make it as small as possible but also want to include the relative with probability α. The latter is called the *efficiency* of the method. We follow Slooten and Meester [3] in their approach to calculating the relevant probabilities and defining the two methods we will discuss in section 4.2.

Like before, we introduce some terminology and definitions. In section 3.1 we introduced S and G, both a distribution over DNA profiles, where S is distributed as the profile of a relative of the case profile, and G distributed as an arbitrary DNA profile, both in our database D. We now interpret the entries of the database D as independent realisations of either S or G, and say $D = \{e_1, \ldots, e_N\}$ where e_i is a DNA profile of person i in D, and call R the index of the person related to the offender.

Furthermore, we define $\pi_i = \mathbb{P}(R = i)$, the prior probability that person i is the true family member of the donor of the crime stain, $\pi_D = \sum_{i=1}^{N} \pi_i$ the prior probability that someone in the database D is the true family member and $\pi_0 = 1 - \pi_D$ the probability that no one in the database is the true family member. We also define the likelihood ratios $r_i = LR_i = \frac{\mathbb{P}(S=e_i)}{\mathbb{P}(G=e_i)}, r_0 = 1$ and combine the likelihood ratios into one vector LR_D, with $LR_D = \mathbf{r} = (r_1, \ldots, r_N)$.

4.1 General probabilities for familial searching

In order to compute the efficiency of the methods, we first have to evaluate some probabilities. Most importantly, we want to know the probability $\mathbb{P}(R = i | LR_D = \mathbf{r})$, the probability that the i^{th} person is the family member of the donor of the crime stain, given the likelihood ratios of all the people in the database. We will prove some probabilities and show a couple of interesting results. We start off with the most important theorem:

Theorem 4.1. *For $i \in \{1, \ldots, N\}$, we have*

$$\mathbb{P}(R = i | LR_D = \boldsymbol{r}) = \frac{r_i \pi_i}{\sum_{k=0}^{N} r_k \pi_k}$$
$$\mathbb{P}(R = i | LR_D = \boldsymbol{r}, R \in D) = \frac{r_i \pi_i}{\sum_{k=1}^{N} r_k \pi_k}$$

Before we prove this theorem, we will prove some Lemmas about our likelihood ratios in general.

Lemma 4.2. *Let E be the collection of DNA profiles, $LR(e) = \frac{\mathbb{P}(S=e)}{\mathbb{P}(G=e)}$ and assume for all $e \in E$ we have $\mathbb{P}(S = e) > 0 \Rightarrow \mathbb{P}(G = e) > 0$. Then for all $x \geq 0$, we have*

$$\mathbb{P}(LR(S) = x) = x\mathbb{P}(LR(G) = x) \tag{9}$$

Proof. First we define E_x as the subset of E where $LR(e) = x$ for some $x \geq 0$, i.e. $E_x = \{e \in E : LR(e) = x\}$, so we can rewrite the probability $\mathbb{P}(LR(S) = x)$ to $\sum_{e \in E_x} \mathbb{P}(S = e)$. Keeping in mind that we defined $LR(e) = \frac{\mathbb{P}(S=e)}{\mathbb{P}(G=e)}$, we can replace $\mathbb{P}(S = e)$ by $LR(e)\mathbb{P}(G = e)$ because we assumed $\mathbb{P}(S = e) > 0 \Rightarrow \mathbb{P}(G = e) > 0$. So, we can write

$$\begin{aligned}\mathbb{P}(LR(S) = x) &= \sum_{e \in E_x} \mathbb{P}(S = e) \\ &= \sum_{e \in E_x} LR(e)\mathbb{P}(G = e)\end{aligned}$$

Remember that we defined $E_x = \{e \in E : LR(e) = x\}$ and so we can rewrite $LR(e)\mathbb{P}(G = e) = x\mathbb{P}(G = e)$ for $e \in E_x$. And so since x is constant, we can write

$$\begin{aligned}&= x \sum_{e \in E_x} \mathbb{P}(G = e) \\ &= x\mathbb{P}(LR(G) = x)\end{aligned}$$

And so we have proven $\mathbb{P}(LR(S) = x) = x\mathbb{P}(LR(G) = x)$. □

Lemma 4.3. *If we assume for all* $e \in E$ *we have* $\mathbb{P}(S = e) > 0 \Rightarrow \mathbb{P}(G = e) > 0$, *then we have*

$$\mathbb{P}(LR(S) \geq x) = \mathbb{E}(LR(G)|LR(G) \geq x)\mathbb{P}(LR(G) \geq x) \tag{10}$$

Proof. Using lemma 4.2, we can rewrite the probability $\mathbb{P}(LR(S) \geq x)$ in the following way

$$\begin{aligned}\mathbb{P}(LR(S) \geq x) &= \sum_{y \geq x} \mathbb{P}(LR(S) = y) \\ &= \sum_{y \geq x} y\mathbb{P}(LR(G) = y) \\ &= \sum_{y \geq x} y\mathbb{P}(LR(G) = y|LR(G) \geq x)\mathbb{P}(LR(G) \geq x) \\ &= \mathbb{E}(LR(G)|LR(G) \geq x)\mathbb{P}(LR(G) \geq x)\end{aligned}$$

Where we need the assumption to use Lemma 4.2 in our second step. □

A consequence of this Lemma is that we have

$$\mathbb{P}(LR(G) \geq x) = \frac{\mathbb{P}(LR(S) \geq x)}{\mathbb{E}(LR(G)|LR(G) \geq x)} \leq \frac{\mathbb{P}(LR(S) \geq x)}{x} \leq \frac{1}{x} \tag{11}$$

Lemma 4.4. *If we assume for all* $e \in E$ *we have* $\mathbb{P}(S = e) > 0 \Rightarrow \mathbb{P}(G = e) > 0$, *then we have* $\mathbb{E}(LR(G)) = 1$

Proof. We can write $\mathbb{E}(LR(G)) = \sum_{e \in E} LR(e)\mathbb{P}(G = e) = \sum_{e \in E} \mathbb{P}(S = e) = 1$, where we assume that the family member is a member of the database. □

Now that we have proved these three lemmas, we can start with out proof of Theorem 4.1, which we will repeat once more before we start the proof.

Theorem 4.1. *For $i \in \{1, \ldots, N\}$, we have*

$$\mathbb{P}(R = i | LR_D = \boldsymbol{r}) = \frac{r_i \pi_i}{\sum_{k=0}^{N} r_k \pi_k} \tag{12}$$

$$\mathbb{P}(R = i | LR_D = \boldsymbol{r}, R \in D) = \frac{r_i \pi_i}{\sum_{k=1}^{N} r_k \pi_k} \tag{13}$$

Proof. We start with a proof of equation (12). We first rewrite the left-hand side using the rules for conditional probabilities

$$\begin{aligned}
\mathbb{P}(R = i | LR_D = \mathbf{r}) &= \frac{\mathbb{P}(LR_D = \mathbf{r} | R = i)\mathbb{P}(R = i)}{\mathbb{P}(LR_D = \mathbf{r})} \\
&= \frac{\mathbb{P}(LR_D = \mathbf{r} | R = i)\mathbb{P}(R = i)}{\sum_{k=1}^{N} \mathbb{P}(LR_D = \mathbf{r} | R = k)\mathbb{P}(R = k) + \mathbb{P}(LR_D = \mathbf{r} | R \notin D)\mathbb{P}(R \notin D)}
\end{aligned}$$

The summation in the denominator is hard to work with, so we will take the inverse at both sides. This gets us the equation

$$\frac{1}{\mathbb{P}(R = i | LR_D = \mathbf{r})} = \sum_{k=1}^{N} \frac{\mathbb{P}(LR_D = \mathbf{r} | R = k)}{\mathbb{P}(LR_D = \mathbf{r} | R = i)} \frac{\mathbb{P}(R = k)}{\mathbb{P}(R = i)} + \frac{\mathbb{P}(LR_D = \mathbf{r} | R \notin D)}{\mathbb{P}(LR_D = \mathbf{r} | R = i)} \frac{\mathbb{P}(R \notin D)}{\mathbb{P}(R = i)} \tag{14}$$

We assume all profiles in our database are independent, and thus we can rewrite the probability $\mathbb{P}(LR_D = \mathbf{r} | R = i)$ to $\prod_{j=1}^{N} \mathbb{P}(LR_j = r_j | R = i)$ where $\mathbb{P}(LR_j = r_j | R = i) = \mathbb{P}(LR_j = j)$ for $j \neq i$; likewise for $\mathbb{P}(LR_D = \mathbf{r} | R = k)$. Thus, dividing the two gives us

$$\begin{aligned}
\frac{\mathbb{P}(LR_D = \mathbf{r} | R = k)}{\mathbb{P}(LR_D = \mathbf{r} | R = i)} &= \frac{\prod_{j=1}^{N} \mathbb{P}(LR_j = r_j | R = k)}{\prod_{j=1}^{N} \mathbb{P}(LR_j = r_j | R = i)} \\
&= \frac{\prod_{j \neq i,k} \mathbb{P}(LR_j = r_j)\mathbb{P}(LR_i = r_i | R = k)\mathbb{P}(LR_k = r_k | R = k)}{\prod_{j \neq i,k} \mathbb{P}(LR_j = r_j)\mathbb{P}(LR_i = r_i | R = i)\mathbb{P}(LR_k = r_k | R = i)} \\
&= \frac{\mathbb{P}(LR_i = r_i | R = k)}{\mathbb{P}(LR_i = r_i | R = i)} \frac{\mathbb{P}(LR_k = r_k | R = k)}{\mathbb{P}(LR_k = r_k | R = i)}
\end{aligned}$$

We can rewrite the ratio $\frac{\mathbb{P}(LR_D = \mathbf{r} | R \notin D)}{\mathbb{P}(LR_D = \mathbf{r} | R = i)}$ in the same way, with the only difference being that only the i^{th} term does not cancel out. Together, this yields us the new expression for the right-hand side

$$\sum_{k=1}^{N} \frac{\mathbb{P}(LR_i = r_i | R = k)}{\mathbb{P}(LR_i = r_i | R = i)} \frac{\mathbb{P}(LR_k = r_k | R = k)}{\mathbb{P}(LR_k = r_k | R = i)} \frac{\mathbb{P}(R = k)}{\mathbb{P}(R = i)} + \frac{\mathbb{P}(LR_i = r_i | R \notin D)}{\mathbb{P}(LR_i = r_i | R = i)} \frac{\mathbb{P}(R \notin D)}{\mathbb{P}(R = i)} \tag{15}$$

We now can use Lemma 4.2 to determine $\frac{\mathbb{P}(LR_i = r_i | R = k)}{\mathbb{P}(LR_i = r_i | R = i)}$, $\frac{\mathbb{P}(LR_k = r_k | R = k)}{\mathbb{P}(LR_k = r_k | R = i)}$ and $\frac{\mathbb{P}(LR_i = r_i | R \notin D)}{\mathbb{P}(LR_i = r_i | R = i)}$; if $R = k$ then the profile of i is distributed like G and thus $\mathbb{P}(LR_i = r_i | R = k) = \mathbb{P}(LR(G) = r_i) =$

$\frac{1}{r_i}\mathbb{P}(LR(S) = r_i)$ by Lemma 4.2. If we have $R = i$ then the profile of i is distributed like S and thus $\mathbb{P}(LR_i = r_i|R = i) = \mathbb{P}(LR(S) = r_i)$. So

$$\frac{\mathbb{P}(LR_i = r_i|R = k)}{\mathbb{P}(LR_i = r_i|R = i)} = \frac{\mathbb{P}(LR(G) = r_i)}{\mathbb{P}(LR(S) = r_i)} = \frac{1}{r_i}$$

In the same manner, we can say

$$\frac{\mathbb{P}(LR_k = r_k|R = k)}{\mathbb{P}(LR_k = r_k|R = i)} = \frac{\mathbb{P}(LR(S) = r_k)}{\mathbb{P}(LR(G) = r_k)} = r_k$$

Now for $R \notin D$ versus $R = i$: for $R = i$ the scenario is equal to the one we discussed first, so $\mathbb{P}(LR_i = r_i|R = i) = \mathbb{P}(LR(S) = r_i)$. If $R \notin D$ then every profile in D is distributed like G and so $\mathbb{P}(LR_i = r_i|R \notin D) = \mathbb{P}(LR(G) = r_i)$. So, in a similar way, we have

$$\frac{\mathbb{P}(LR_i = r_i|R \notin D)}{\mathbb{P}(LR_i = r_i|R = i)} = \frac{\mathbb{P}(LR(G) = r_i)}{\mathbb{P}(LR(S) = r_i)} = \frac{1}{r_i}$$

Summarizing these results, we can rewrite expression (15) into

$$\sum_{k=1}^{N} \frac{r_k}{r_i}\frac{\mathbb{P}(R = k)}{\mathbb{P}(R = i)} + \frac{1}{r_i}\frac{\mathbb{P}(R \notin D)}{\mathbb{P}(R = i)}$$

Remember that this expression was equal to $\frac{1}{\mathbb{P}(R=i|LR_D=\mathbf{r})}$ and we defined at the start of this chapter notation for the remaining expressions, and so rewriting expression (14) gives us

$$\begin{aligned}\mathbb{P}(R = i|LR_D = \mathbf{r}) &= \frac{1}{\sum_{k=1}^{N} \frac{r_k}{r_i}\frac{\pi_k}{\pi_r i} + \frac{1}{r_i}\frac{\pi_0}{\pi_i}} \\ &= \frac{1}{\sum_{k=0}^{N} \frac{r_k \pi_k}{r_i \pi_i}} = \frac{r_i \pi_i}{\sum_{k=0}^{N} r_k \pi_k}\end{aligned}$$

As required. Now with equation 13 the computation is nearly the same; we can rewrite the probability $\mathbb{P}(R = i|LR_D = \mathbf{r}, R \in D)$ into

$$\begin{aligned}\mathbb{P}(R = i|LR_D = \mathbf{r}, R \in D) &= \frac{\mathbb{P}(LR_D = \mathbf{r}|R = i)\mathbb{P}(R = i)}{\mathbb{P}(LR_D = \mathbf{r}, R \in D)} \\ &= \frac{\mathbb{P}(LR_D = \mathbf{r}|R = i)\mathbb{P}(R = i)}{\sum_{k=1}^{N}\mathbb{P}(LR_D = \mathbf{r}|R = k)\mathbb{P}(R = k)}\end{aligned}$$

(where we don't have the extra $\mathbb{P}(LR_D = \mathbf{r}|R \notin D)\mathbb{P}(R \notin D)$ term because we already assume that we have a family member in D)

In a similar fashion, we can derive the expression

$$\frac{1}{\mathbb{P}(R=i|LR_D=\mathbf{r},R\in D)}=\sum_{k=1}^{N}\frac{r_k}{r_i}\frac{\mathbb{P}(R=k)}{\mathbb{P}(R=i)}=\sum_{k=1}^{N}\frac{r_k\pi_k}{r_i\pi_i}$$

And thus we have proved equation 13:

$$\mathbb{P}(R=i|LR_D=\mathbf{r},R\in D)=\frac{r_i\pi_i}{\sum_{k=1}^{N}r_k\pi_k}$$

□

Note that we only needed Lemma 4.2 to prove this theorem.

4.2 Different methods

As mentioned before, we will discuss two methods: the conditional method and the profile-centered method. We will first explain how their approach to finding a subset works, and will discuss some important differences in section 4.3 afterwards.

4.2.1 Conditional method

For the first subset method, we look at the subset D^k corresponding to the k profiles that have the highest products $r_i\pi_i$. We then look at the minimal k, say k_α, for which

$$\sum_{j\in D^k}r_j\pi_j\geq\alpha(r_1\pi_1+\cdots+r_N\pi_N)$$

holds for $0\leq\alpha\leq1$, and write D^c_α for D^{k_α}. This D^c_α will be our group of possible family members and will contain an actual family member with a certain probability. We now want to determine the efficiency of this method, the probability $\mathbb{P}(R\in D^c_\alpha|R\in D)$. We can calculate $\mathbb{P}(R=i|R\in D)$ first by using the law of total probability and theorem 4.1

$$\begin{aligned}\mathbb{P}(R=i|R\in D)&=\sum_{\mathbf{r}}\mathbb{P}(LR_D=\mathbf{r}|R\in D)\mathbb{P}(R=i|R\in D,LR_D=\mathbf{r})\\&=\sum_{\mathbf{r}}\mathbb{P}(LR_D=\mathbf{r}|R\in D)\frac{r_i\pi_i}{\sum_{k=1}^{N}r_k\pi_k}\end{aligned}$$

So for $\mathbb{P}(R\in D^c_\alpha|R\in D)$ we will simply need to sum over all the $i\in D^c_\alpha$:

$$\begin{aligned}\mathbb{P}(R\in D^c_\alpha|R\in D)&=\sum_{\mathbf{r}}\mathbb{P}(LR_D=\mathbf{r}|R\in D)\mathbb{P}(R\in D^c_\alpha|R\in D,LR_D=\mathbf{r})\\&=\sum_{\mathbf{r}}\mathbb{P}(LR_D=\mathbf{r}|R\in D)\sum_{i\in D^c_\alpha}\frac{r_i\pi_i}{\sum_{k=1}^{N}r_k\pi_k}\end{aligned}$$

By definition of D^c_α, we have $\sum_{i\in D^c_\alpha} r_i\pi_i \geq \alpha \sum_{k=1}^N r_k\pi_k$ and $\sum_{\mathbf{r}} \mathbb{P}(LR_D = \mathbf{r}|R \in D) = 1$, so we have

$$\mathbb{P}(R \in D^c_\alpha|R \in D) = \sum_{\mathbf{r}} \mathbb{P}(LR_D = \mathbf{r}|R \in D) \sum_{i\in D^c_\alpha} \frac{r_i\pi_i}{\sum_{k=1}^N r_k\pi_k} \geq \alpha$$

And so we have shown that the efficiency of this method is at least α.

4.2.2 Profile-centred method

The construction of the profile-centred method requires less information than the conditional method: we don't look at the prior probabilities, but solely at the likelihood ratios. Recall that we interpret the profiles in D as realisations of either S (a family member of the crime stain donor) or G (a random member of the population). For this method, we look at $\mathbb{P}(LR(S) \geq t)$ for some t, which is the probability that the likelihoodratio for a family member is at least t. Let t_α be the largest t for which

$$\mathbb{P}(LR(S) \geq t) \geq \alpha$$

for $0 \leq \alpha \leq 1$, where α is our efficiency. This t_α is our treshold for the new subset of possible family members; we define

$$D^P_\alpha = \{i \in D : LR_i \geq t_\alpha\} \tag{16}$$

We can check the efficiency $\mathbb{P}(R \in D^P_\alpha|R \in D)$ by a very quick calculation: if $R \in D^P_\alpha$ then $LR(R) \geq t_\alpha$, and thus

$$\mathbb{P}(R \in D^P_\alpha|R \in D) = \mathbb{P}(LR(R) \geq t_\alpha) = \mathbb{P}(LR(S) \geq t_\alpha) \geq \alpha$$

And so the efficiency indeed is at least α. We can even calculate the probability of D containing family member when we do not find a relative in our subset D^P_α by using the law of total probability and Bayes' theorem:

$$\begin{aligned}
\mathbb{P}(R \in D|R \notin D^P_\alpha) &= \frac{\mathbb{P}(R \notin D^P_\alpha|R \in D)\mathbb{P}(R \in D)}{\mathbb{P}(R \notin D^P_\alpha)} \\
&= \frac{(1-\alpha)\pi_D}{\mathbb{P}(R \notin D^P_\alpha|R \in D)\mathbb{P}(R \in D) + \mathbb{P}(R \notin D^P_\alpha|R \notin D)\mathbb{P}(R \notin D)} \\
&= \frac{(1-\alpha)\pi_D}{(1-\alpha)\pi_D + (1-\pi_D)} = (1-\alpha)\frac{\pi_D}{1-\alpha\pi_D}
\end{aligned}$$

Since $\mathbb{P}(R \in D) = \pi_D$ as defined at the start of this chapter and $\mathbb{P}(R \notin D^P_\alpha|R \notin D) = 1$, because D^P_α cannot contain a relative if it is not present in the database in the first place.

Remark 4.5. *Another way to construct our subset is to look at members whose profiles have an exceptionally high likelihood ratio if they were unrelated, as proposed by Sjerps and Kloosterman [5]. As with D_α^P we define a treshold s_β such that*

$$\beta = \mathbb{P}(LR(G) \geq s_\beta)$$

for some $\beta \in (0,1)$, giving us a new subset $\{i \in D : LR_i \geq s_\beta\}$. We define this β and s_β such that we hope to include a fraction β of our database in our subset. We can relate this to our subdatabase D_α^P in the following way

$$\alpha \leq \mathbb{P}(LR(S) \geq t_\alpha) = \mathbb{E}(LR(G)|LR(G) \geq t_\alpha)\mathbb{P}(LR(G) \geq t_\alpha)$$

By applying Lemma 4.3. If we choose β such that $t_\alpha = s_\beta$, then we can rewrite

$$\alpha \leq \beta \cdot \mathbb{E}(LR(G)|LR(G) \geq s_\beta)$$

Meaning that we cannot express α in β alone but it also depends on the case profile in question. This means that we also cannot give a good lower bound for the efficiency, in contrast to D_α^P which has a uniform lower bound α.

4.3 Comparison of the two methods

We have shown that both subsets have an efficiency of at least α, but there are some important differences that we will need to discuss. First of all, we require a lot more information to construct D_α^c, since we need to know the prior probabilities π_i for all i. We do not need this information for D_α^P, so it would seem that since D_α^c requires more information it is the better method to use.

However, interpretation wise, it is maybe more appealing to use D_α^P. A common interpretation in statistics and probability theory is the *frequentist interpretation*, where we interpret the probability of an event as its relative frequency when conducting a large amount of trails. If we look at the (random) treshold for the two subsets, the distribution of D_α^c's treshold depends on both S and G. A frequentist interpretation requires resampling of our database: if we would create a large amount of sets of one copy of S versus $N-1$ copies of G, then about $\mathbb{P}(R \in D_\alpha^P)$ times we will have included the copy of S.

If we compare this with D_α^P, its treshold only depends on S. We also don't have to resample the database for our interpretation of D_α^P: we can interpret it as one realisation of $N-1$ copies of G and can repeatedly add one copy of a realisation of S. In this case again a fraction of $\mathbb{P}(R \in D_\alpha^P)$ this realisation of S is included in our subset. This interpretation is especially important because we are not dealing with a theoretical application: the results of our analysis will be used by inspectors and possibly in court, and will have to be used to make decisions about the remainder of the investigation. From this viewpoint, the usage of D_α^P is more helpful for law enforcement.

Besides theoretical anlysis, we can apply these models in a simulated database to compare their performance. We will do so in chapter 5.

5 Modeling in R

Now that we have examined two methods to look for relatives of a case profile in a database, we can apply these techniques in a simulated database with simulated case profiles. We use the package *DNAprofiles*, written by Maarten Kruijver [6] for R. This package gives us a way to sample a database according to allelic frequencies in the Netherlands, and has a pre-built function to determine the likelihood ratios we discussed in previous chapters.

Since it is (quite a bit) harder to find siblings in a database than to find parents or siblings, we will be focussing on the results when looking for siblings rather than other types of relatives. Furthermore, we cannot compute the prior probabilities π_i in our simulation, since these probabilities are case dependent. Therefore, we assume a uniform prior on the whole database, meaning $\pi_i = \frac{1}{n}$ for all $i \in D$, where n is the size of the population.

5.1 Probability of detection

First and foremost, we will be looking at the *probability of detection* when using the two methods. We sampled a large database D ($N = 10^5$, about the size of the actual DNA database that is managed by the Dutch Forensic Institute [7]) and then sampled about 100 profile cases C_i for which we determined the probability that we would detect a relative in the database. This method differed slightly for the two methods, so we will discuss them seperately. These simulation strategies are based on the simulations by Slooten and Meester [3]. See Appendix 8.1 for the R code.

5.1.1 Conditional method

For the conditional method, we sampled 1000 siblings $S_{i,j}$ ($i \in \{1, \dots, 100\}, j \in \{1, \dots, 1000\}$) for every case profile we sampled and determined for every sibling if they would be included in $\bar{D}^c_\alpha$ for a certain efficiency α, where $\bar{D}$ is the database D with the sampled sibling added. For all those profiles, we calculated the likelihood ratio r_i for being a sibling of the profile for all members of the database and all simulated siblings. We will refer to this likelihoodratio as their *Sibling Index* (SI) from now on. We first define the *rank* of every sibling by the number of database members that have a higher SI, where a sibling has rank k if there are $k - 1$ members of the database with a higher SI.

If the rank of the sibling is 1, it will always be included in $\bar{D}^c_\alpha$ since the priors are uniform and thus we only look at the k_α largest SI. If the rank of the sibling is higher than 1, this means that for some α's the sibling might not be included in $\bar{D}^c_\alpha$: the smaller the efficiency, the smaller the size of our selection. In section 5.2.1 we will discuss this in greater detail. We define $t_{i,j}$ as the largest $t \geq 0$ such that $S_{i,j} \notin \bar{D}^c_t$ (and 0 if its rank is 1). This means that $S_{i,j}$ is only included in our subset if $\alpha > t_{i,j}$, and thus we can estimate the probability of detection for a given α.

For $\alpha \in \{0.01, \dots, 1\}$ we calculated $\beta_{i,\alpha}$, the fraction of $t_{i,j}$ that are smaller than α, for every case sample. We then took the average over all case samples, which gave us the estimate β_α for our probability of detection, the probability that if we add a relative of a random case profile to our simulated database D, then the relative would be included in $\bar{D}^c_\alpha$.

One thing to note is that this estimate is database dependent, meaning that the probability of detection can only be defined in relation to our simulated database. For every case profile the probability of detection is not only affected by the exact profiles in the database but also by the database size, which we will discuss in section 5.1.3.

Our simulation in R gave us the result shown in the following Figure

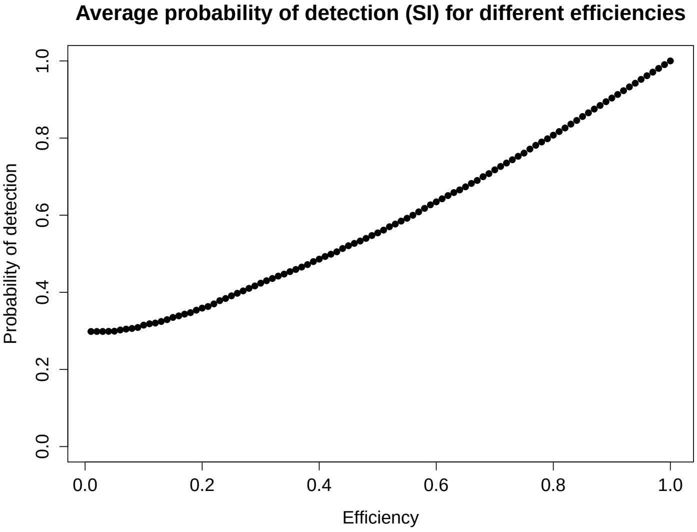

Figure 1

We see that the probability of detection is bigger than the efficiency for every α, although the difference between the two gets smaller as the efficiency gets higher. This is due to the way we defined the subset $\bar{D}_{\alpha}^{c}$ and the fact that we assumed a uniform prior: we only look at the k_α largest SI's, which will contain on average a higher percentage of actual siblings when k_α is small. The higher α, the more we add members with a lower SI score and thus on average more non-related profiles.

5.1.2 Profile-centred method

As discussed in the last chapter, the interpretation of the efficiency is a bit easier for the profile-centred method than for the conditional method. The same holds for the computation of the probability of detection: because the distribution of the treshold depends only on the case profile, simulating the probability of detection requires only finding the treshold for the case profile in question and calculating the percentage of siblings that have a SI higher than the calculated t_α.

In the simulation, we computed the treshold by using one of the functions in the package that determined the cumulative distribution function for the distribution of the SI for a relative of the case profile. We computed this treshold for every α and sampled a large amount siblings for every one of the 100 case profiles. The estimate β_α of the probability of detection is found by taking the average of the percentage of siblings that have a higher SI. As mentioned before, this method does not depend on the database size nor on the exact profiles in the database.

Our simulation, with in this case 50.000 siblings per case profile, gave us the following result:

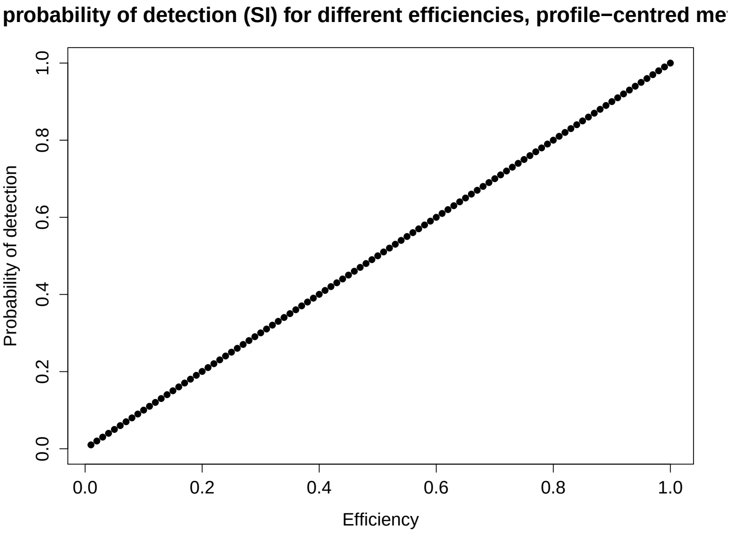

Figure 2

This figure is less exciting than that of the conditional method, but also quite easy to explain: we defined t_α as the maximum t for which $\mathbb{P}(LR(S) \geq t) \geq \alpha$ held. This means that, because we have a large amount of siblings, on average about α of the siblings are included in the set $\bar{D}^P_\alpha$. Thus, our estimate β_α is close or equal to our efficiency for all α's.

5.1.3 Influence of database size

We already remarked that the probability of detection for the conditional method was database dependent, contrary to the profile-centred method. For different sizes of databases, we can compute the average probability of detection to see how big the influence really is.

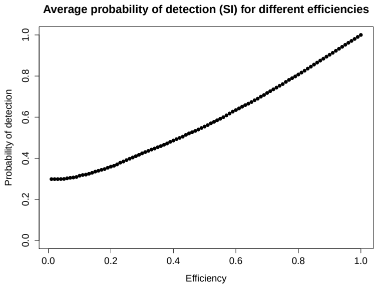

Figure 3: N = 100.000

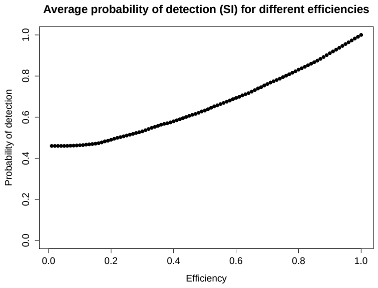

Figure 4: N=10.000

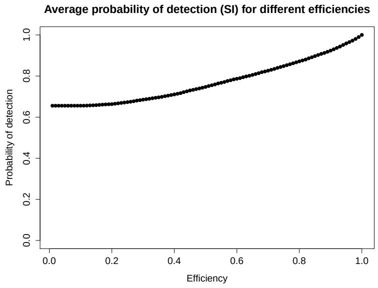

Figure 5: N=1.000

We can easilly see that if the database is smaller, the probability of detection is higher. This is of course no surprise: remember that the probability of detection is determined by the fraction of relatives that would be included in our set $\bar{D}^c_\alpha$, the selected subset of the database D with a relative added. Therefore if the database is smaller, there are less members in the database that can accidentally favourably match with our profile case and, in general, less members that can compete for highest SI with our relative. As a result, high SI values stand out more and get included in $\bar{D}^c_\alpha$ moreoften, and therefore relatives of a case profile are easier to detect.

5.2 Size of selected subset

In reality, we often will not have a relative of our case profile in our database. According to a spokesperson at the NFI, out of the 33 cases where the Dutch Forensic Institute used familial searching to look for suspects, 5 instances led to a suspect for the police to investigate [8]. Therefore, it is useful to know what the typical size of D^c_α or D^P_α is when there are no relatives present in the database D. This size depends both on the chosen efficiency and the frequency of the case profile, and we will discuss these two factors seperately.

We again sample a database D with $N = 10^5$ and sample 100 case profiles. However, this time we do not sample any family members, because we are no longer working under the assumption that a relative is present. This will not only affect the size of the selection but is also a better representation of the reality: we have no way of knowing whether there is a relative present in our database before we run some tests. See Appendix 8.2 for the relevant R codes.

5.2.1 Efficiency

We first look at the influence of the efficiency on the size of the database. We ran both methods for our database D and 100 case profiles and averaged the size of the selection for $\alpha \in \{0.01, \dots, 1\}$. Taking the natural log of this size gives us the following Figures

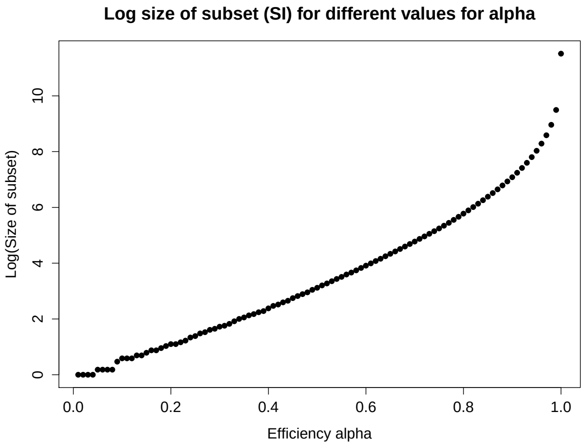

Figure 6: Size of D^c_α

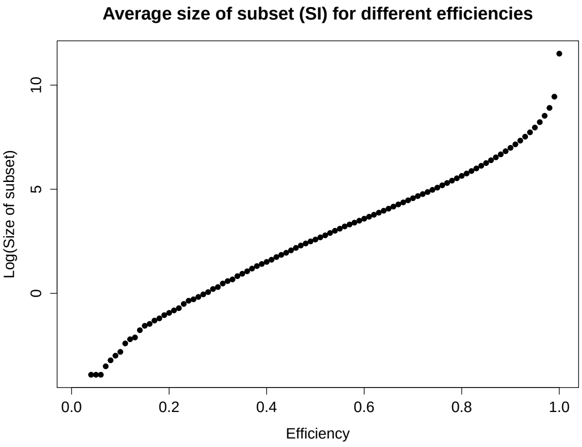

Figure 7: Size of D^P_α

One thing to note is that the conditional method will, by construction, always include at least one member of the database in D^c_α. The profile-centred method will not: this method finds a treshold t_α which will be very high for low values of α, meaning that in many cases there are no members with a high enough SI. We can see that the average size of D^P_α starts to exceed 1 for α bigger than 0.2, whereas the size of D^c_α is already around 6 for the same efficiency.

For both methods there seems to be an almost exponential relation between the size of the subset and the efficiency. We also see that the function is monotone increasing for both methods, which makes sense: for the profile-centred method, the treshold t_α is strictly decreasing for α, meaning that more members will be included in D^P_α as our efficiency gets higher. As for the conditional method, remember that we chose a uniform prior on our database. This means that D^c_α consists only of those with the highest SI, with no more members included than strictly necessary to have $\sum_{i \in D^c_\alpha} \pi_i \geq \alpha \pi_D$. Naturally, the higher our α, the more members we will need to include in D^c_α.

5.2.2 Profile frequencies

Secondly, we wrote a function in R to find the frequency of a given profile and used this to plot the size of the subset against the profile frequency. For the function in R we found the frequency by computing the frequency per locus, which is $p_a p_b (2 - \delta_{a,b})$ for a locus with alleles (a, b) with $\delta_{a,b} = \mathbb{1}_{\{a=b\}}$. We used the allelic frequencies from the R package at 10 SGMplus loci, which are the loci documented for the SGM Plus DNA profiling system. The profile frequencies p we found were between 10^{-12} and 10^{-20}, so we plotted $-\log_{10}(p)$ against the size of D^P_α or D^c_α for $\alpha \in \{0.7, 0.8, 0.9\}$, with p the frequency of the profile. The results were as follows

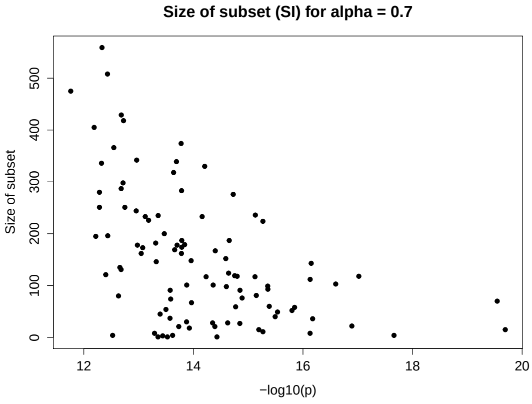

Figure 8: Conditional method, $\alpha = 0.7$

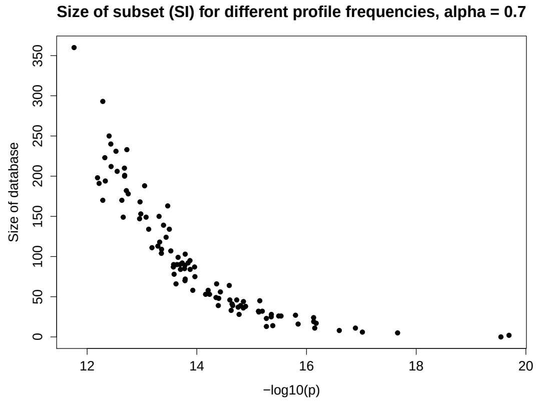

Figure 9: Profile-centred method, $\alpha = 0.7$

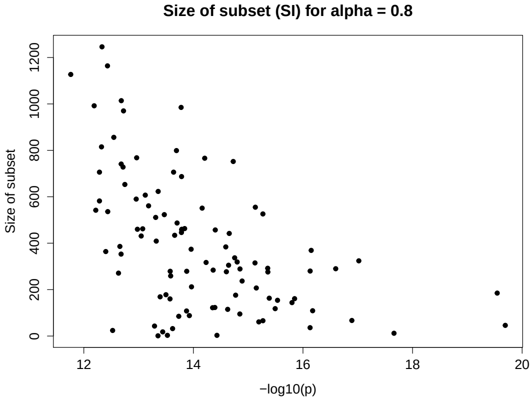

Figure 10: Conditional method, $\alpha = 0.8$

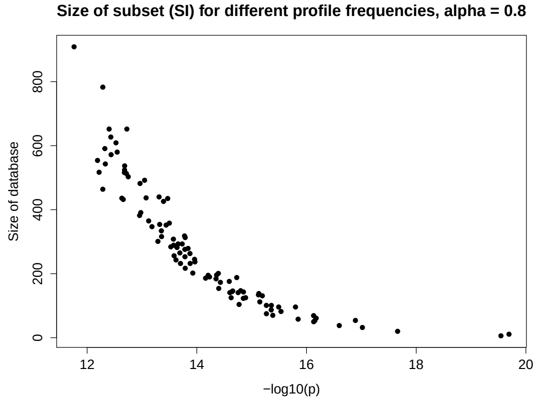

Figure 11: Profile-centred method, $\alpha = 0.8$

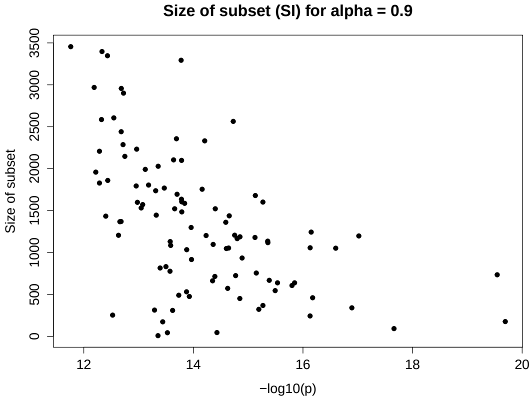

Figure 12: Conditional method, $\alpha = 0.9$

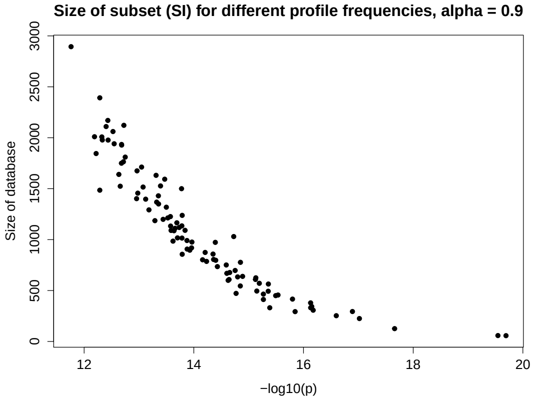

Figure 13: Profile-centred method, $\alpha = 0.9$

We see again that as α increases, the selection size increases. However, we also see that the rarer the profile, the lower the selection size - this is visible in the conditional method, but even more so in the profile-centred method. This is also a result of the way we defined our subsets: if the frequency of our profile is low, the SI are lower on average, meaning that for the profile-centred method there are less members whose SI will exceed our t_α. For the conditional method our π_D is low, which means we will need fewer members with a high SI in our subset D^c_α.

5.3 Comparison

Now that we simulated both methods and looked at some interesting properties, we can analyse the methods. We start with the probability of detection: we've seen that for every α, the conditional method is on average better in including the relative than the profile-centred method. However, when looking at the standard deviation of the probability of detection, we find the following

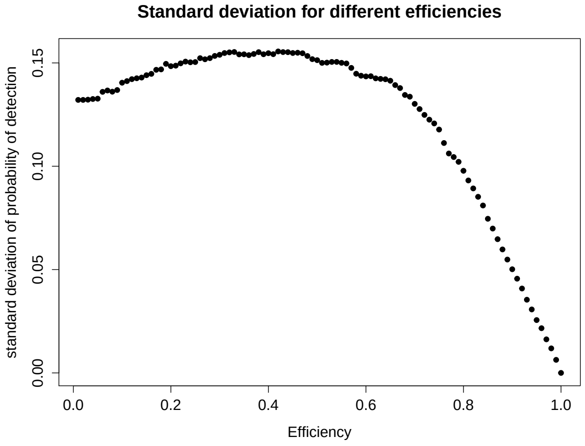

Figure 14: σ for β_α, conditional method

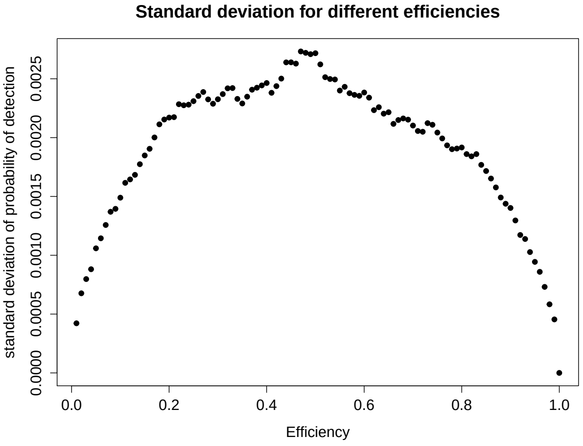

Figure 15: σ for β_α, profile-centred method

We see that the standard deviation for the conditional method is between 0.10 and 0.15 for most values for α, which is a large standard deviation for a probability. However for the profile-centred method the standard deviation is at most 0.0025, which is quite low. If we look at three different profiles with different frequencies, we can see how the standard deviation translates into a difference in probability of detection:

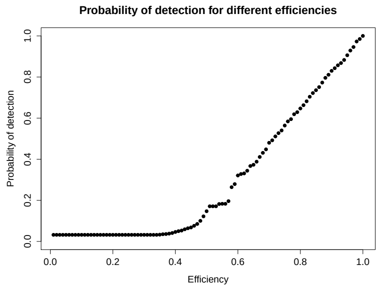

Figure 16: $p = 1.05 \cdot 10^{-12}$

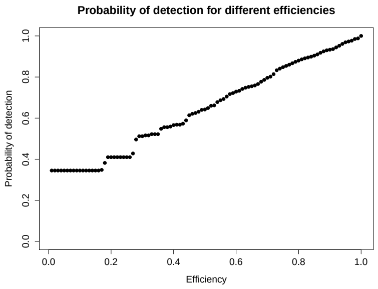

Figure 17: $p = 1.06 \cdot 10^{-15}$

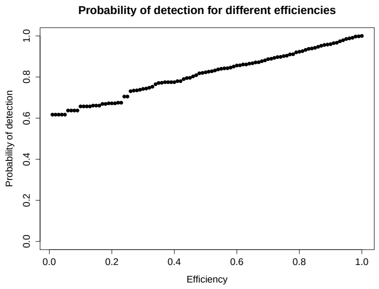

Figure 18: $p = 5.47 \cdot 10^{-17}$

If the profile has a low frequency, then the conditional method has some difficulty finding the relative among the other relatives in the database. In contrast to what we have seen before, our estimate β_α is even lower than α for almost all α. However for lower frequencies the conditional method seems to perform even better than average. Compare this to the profile-centred method, where we see the following:

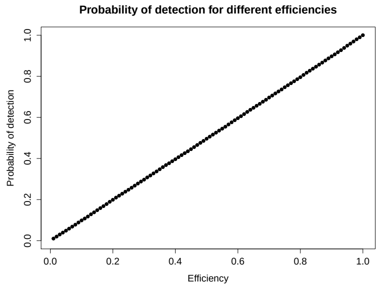

Figure 19: $p = 4.898 \cdot 10^{-12}$

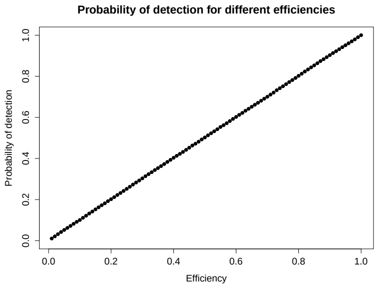

Figure 20: $p = 5.86 \cdot 10^{-15}$

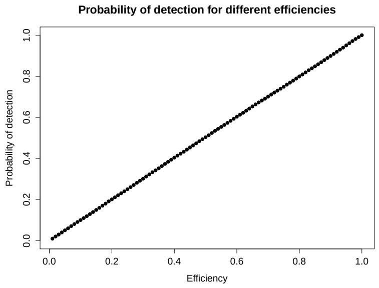

Figure 21: $p = 7.67 \cdot 10^{-17}$

As expected, the profile-centred method seems to be consistent over all profiles. It seems that for profiles with a high frequency, the profile-centred method is a better choice since it is consistent for lower α and performs better than the conditional method in this case. For a low-frequency profile, the conditional method seems to be a better choice. One thing to note is that in practice, we will often go for a higher efficiency. When testing in a lab, we will want to know that we include our relative with a relatively high probability; we also see that the for higher efficiencies, our two methods do not seem to differ much.

Aside from the probability of detection, the size of the selection is important for the future of the investigation too. Police and other investigators are dealing with limited resources and when looking for a relative, having to investigate over 3000 database members takes a long time. We see in section 5.2.2 that on average the size of the selection is a fair amount smaller for the profile-centred method than for the conditional method, especially for profiles with a lower frequency. Therefore, the profile-centred method has a clear advantage over using the conditional method, as it reduces the average amount of people we will need to investigate.

Both methods have their advantages and disadvantages. For the profile-centred method, the predictability of its performance is one of its advantages. When we know the frequency of the case profile, we can use this to predict D_α^P like previous simulations in this chapter. In addition, we can use equation (11) to look at the amount of unrelated members included in D_α^P. This tells us that the probability that an unrelated members has a KI higher than t_α is at most $\frac{\alpha}{t_\alpha}$, so we expect to include at most $N\frac{\alpha}{t_\alpha}$ unrelated members in our subset D_α^P. If the subset turns out not to contain any relative of the case profile, we can say something about the probability that D contains a relative. If π_D was small, we can approximate the probability $\mathbb{P}(R \in D | R \notin D_\alpha^P)$ discussed in section 4.2.2 by $(1-\alpha)\pi_D$. So, if our method did not detect a relative, the probability that D still contains a relative has decreased by a factor $1-\alpha$. This is useful to know for the investigators that need to decide whether to test additional database members after no new suspect has been found.

The conditional method, however, is more powerful because it takes into account more information related to the case. This is only helpful when we can use information from the case to determine priors - when we can only use uniform priors, this method does not seem to perform much better than the profile-centred method, which is more attractive to use because of its interpretation and consistency. However, often we can use properties like age, ethnicity or geography (if there are eye-witnesses, for example) to determine some kind of prior distribution.

When working with the conditional method, it is useful to look at the KI of the members in D_α^c. Lemma 4.4 tells us that $\mathbb{E}(LR(G)) = 1$, so if the sum of all likelihood ratios is smaller than N this might be an indication that D does not contain a relative, whereas if the sum is larger than N this might indicate the presence of a relative in our database. So, if all likelihood ratios are small, this could lead to a fairly big subset D_α^c which is not likely to contain a relative and not always worthy of investigation.

All in all, there is no clear answer to which method is better: making a choice between the two methods should be done on a case by case basis. If there is no way to define better prior probabilities than a uniform distribution, it might be better to use the profile-centred method. However, when a suitable prior distribution can be calculated, it seems to be worthwhile to use the conditional method. Nevertheless, since the profile has a big impact on the probability of detection and the size of the selection, it is advisable to run some simulations before choosing either one of the methods. This way, the probability of finding a relative and the workload can be estimated before making a choice between the methods.

6 Conclusions

In the last chapters, we have examined the statistics behind DNA searching in databases and familial searching, and looked at two methods for finding relatives in a database.

We have explained why we use likelihood ratios when working with DNA in criminal cases: even though they might seem to be inconsistent because of the database controversy, using posterior odds shows that they are not. However, when stating the strength of the evidence, which is calculated by likelihood ratios, the prior and posterior probabilities should be stated as well: without them, we cannot say anything about the strength of the case itself. The case against our match can have high prior odds but a weak match, which often happens in the probable cause case, or can have a strong match but low prior odds, as in the DNA searching case, and be equally incriminating.

We examined two methods for searching for relatives: the conditional method and the profile-centred method. During the theoretical analysis we noted that the first requires more information and thus is more powerful, but the latter is easier to interpret. Especially when looking at the frequentist interpretation, the conditional method requires resampling of the database while the profile-centred method does not.

When simulating the two methods, we found that the profile-centred method is more consistent and is more selective, giving us a smaller subset to investigate while not being that much worse at detecting relatives. It also is able to tell us that the probability of a relative being present in the database, if it's not found in D_α^P, will decrease by a factor $1-\alpha$ if π_D is low. The conditional method is indeed more powerful when we can choose a useful prior distribution. We can use properties like age, ethnicity and geography to influence the selection made by this method, which often positively influences the probability of detection. However when there is no way to determine a prior distribution, it does not seem worthwhile to choose this method over the profile-centred method.

When looking at our simulations, we also see that for different profile frequencies, the two methods seem to perform differently. For profiles with a high frequency, the conditional method seems to perform quite poorly in comparison with the profile-centred method, where profiles with a low frequency lead to a high overall probability of detection. In general, the profile-centred method seems to include less members on average for every profile, making future investigation easier.

All things considered, both methods have their up- and downsides and therefore choosing one of the two methods should be done on a case by case basis. If it is possible to determine a prior distribution aside from a uniform distribution, it seems that the conditional method is often a better pick. Otherwise, the profile-centred method is advised. However, aspects such as the profile frequency should be considered as well, and before choosing either one of the methods it is advised to run some simulations with the case profile to try to predict the workload and probability of finding a relative.

For future research it might be interesting to look for new methods that try to combine the strengths of the two methods and balance out the flaws. Ideally, this new method would be powerful, predictable in its outcome and have both an easy interpretation as a high average probability of detection. Familial searching is a tool that has already proven its right to exist, but methods can always be altered or invented in hopes of improving the workload or probability of detection.

7 Discussion

Familial searching, like most tools used in the judicial system, is not perfect and its implementation will cause some problems. One of those problems is that it can - and will - increase the racial bias that already exists in our system. The Dutch Bureau for Statistics showed [9] that in 2002, in total 0.9% of the native Dutch population were ever registered as a suspect in a case, in contrast to 2.2% of the immigrant population. To put that into perspective: this means 61.000 of the total 163.000 suspects were first or second generation immigrants, which made their representation in the suspect pool twice as large as in the general Dutch population.

If we use familial searching more often, this will cause the gap between representation of the immigrant and autochthonous population in the suspect pool to widen even more. During research for this thesis, I found this aspect of familial searching to be very interesting and worthy of some attention, but when trying to model this gap and the increase that familial searching would cause, I repeatedly found myself having trouble with the lack of data needed to model this trend. Much of the research in this field is of a sociological nature, but not much research is done in the biological and mathematical field, making it hard to model this in any way close to the real world. However I would still like to research this one day, as I think it is an interesting subject worthy of more research. It is not only mathematically interesting, but could provide a sound mathematical foundation for the political debate around familial searching.

Last but not least I would like to point out that the codes I wrote for the simulations are in no way fully optimized. Right now, the conditional method is fairly slow compared to the profile-centred method. Although I am not new to R, I have no doubt that these methods can be written better to speed up the simulations.

8 Appendix

8.1 Section 5.1

§5.1.1

```
data(freqsNLsgmplus)

# set db sample size
N = 1e5
# set relative sample size
M = 1000

set.seed(100)

# sample a reference db
db <- sample.profiles(N,freqsNLsgmplus)
b <-matrix(,nrow=100,ncol=100)

# run for 100 profile cases
for(i in 1:100){
  x <- sample.profiles(1, freqsNLsgmplus)
  rs <- sample.relatives(x,M,type="FS")

  # compute SI for all database members
  SI <- ki.db(x,db,"FS")
  SIvector <- as.vector(SI)
  SI_r <- ki.db(x,rs,"FS")

  #compute rank for siblings
  ranks <- c()
  for(j in 1:M){
    ranks[j]<-sum(SI>SI_r[j,1]) + 1
  }

  # CONDITIONAL METHOD, compute t_i,j
  # we take uniform priors
  tp<-c()
  for(n in 1:M){
    high<-c()
    if(ranks[n]>1)
    {
      high<-sort(SIvector,decreasing=TRUE)[1:ranks[n]-1]
      for (q in seq(from=0.01,to=1,by=0.01)){
        if (sum(high) < q*(sum(SI)+SI_r[n])) break
        tp[n]<-q
      }
    }
    else{
      tp[n]=0
    }
  }

  k=1
  # compute b_alpha
  for(a in seq(from=0.01, to=1, by=0.01)){
    b[i,k]<-sum(tp<a)/M
    k=k+1
  }
}

ba<-colMeans(b)
plot(seq(from=0.01, to=1, by=0.01),ba,ylim=c(0,1),main="Average probability of detection
for different efficiencies",xlab="Efficiency",ylab="Probability of detection",pch=19)
```

§5.1.2

```
data(freqsNLsgmplus)

# set db sample size
N = 1e5
M = 50000
set.seed(100)

# sample a small reference db
db <- sample.profiles(N,freqsNLsgmplus)
prob<-matrix(,nrow=100,ncol=100)

# compute for all case profiles
for(i in 1:100){
  x <- sample.profiles(1, freqsNLsgmplus)
  rs <- sample.relatives(x,M,type="FS")

  # compute SI for all database members
  SI <- ki.db(x,db,"FS")
  SIr <- ki.db(x,rs,"FS")

  # compute cdf for SI of true sibling
  cdf.fs <- ki.cdf(x,hyp.1="FS",hyp.true="FS")
  cdfinv <- inverse(cdf.fs)

  m<-1
  # set efficiency and find t_alpha
  for(a in seq(from=0.01,to=1,by=0.01)){
    t <- cdfinv(1-a)
    prob[i,m]<-sum(SIr>=t)
    m<-m+1
  }
}

p<-colMeans(prob)/M

plot(seq(from=0.01,to=1,by=0.01),p,main="Average probability of detection (SI)
for different efficiencies, profile-centred method",xlab="Efficiency",
ylab="Probability of detection",pch=19)
```

```
#function for finding the inverse of the cdf function
inverse = function (f, lower = 0, upper = 1e12) {
  function (y) uniroot((function (x) f(x) - y), lower = lower, upper = upper)[1]
}
```

8.2 Section 5.2

§5.2.1

```
#function for finding the frequency of a certain profile
#compared to freqsNLsgmplus

profile.freq<-function(x){
  data(freqsNLsgmplus)

  lijst1 <-unlist(freqsNLsgmplus[1],use.names=FALSE)
  lijst2 <-unlist(freqsNLsgmplus[2],use.names=FALSE)
  lijst3 <-unlist(freqsNLsgmplus[3],use.names=FALSE)
  lijst4 <-unlist(freqsNLsgmplus[4],use.names=FALSE)
  lijst5 <-unlist(freqsNLsgmplus[5],use.names=FALSE)
  lijst6 <-unlist(freqsNLsgmplus[6],use.names=FALSE)
  lijst7 <-unlist(freqsNLsgmplus[7],use.names=FALSE)
  lijst8 <-unlist(freqsNLsgmplus[8],use.names=FALSE)
  lijst9 <-unlist(freqsNLsgmplus[9],use.names=FALSE)
  lijst10 <-unlist(freqsNLsgmplus[10],use.names=FALSE)

 p <- lijst1[x[1]]*lijst1[x[2]]*(2-diracD(x[1],x[2]))*lijst2[x[3]]*lijst2[x[4]]*
    (2-diracD(x[3],x[4]))*lijst3[x[5]]*lijst3[x[6]]*(2-diracD(x[5],x[6]))*
    lijst4[x[7]]*lijst4[x[8]]*(2-diracD(x[7],x[8]))*lijst5[x[9]]*lijst5[x[10]]*
    (2-diracD(x[9],x[10]))*lijst6[x[11]]*lijst6[x[12]]*(2-diracD(x[11],x[12]))*
    lijst7[x[13]]*lijst7[x[14]]*(2-diracD(x[13],x[14]))*lijst8[x[15]]*lijst8[x[16]]*
    (2-diracD(x[15],x[16]))*lijst9[x[17]]*lijst9[x[18]]*(2-diracD(x[17],x[18]))*
    lijst10[x[19]]*lijst10[x[20]]*(2-diracD(x[19],x[20]))

  return(p)
}
```

```
#dirac delta function

diracD<-function(a,b){
  if(a==b){
    return(1)
  }
  else{
    return(0)
  }
}
```

Conditional method

```
data(freqsNLsgmplus)

# set db sample size
N = 1e5

```

```
set.seed(100)

# sample a small reference db
db <- sample.profiles(N,freqsNLsgmplus)

ksize<-matrix(,nrow=100,ncol=2)
for(j in 1:100)
{
  x <- sample.profiles(1, freqsNLsgmplus)

  # compute SI for all database members
  SI <- ki.db(x,db,"FS")

  m<-1
  # set minimum efficiency
  a = 0.9

    # CONDITIONAL METHOD
    # we take uniform priors
    i <- 0
    high <- c()

    # find size of subset
    for (k in 1:N)
    {
      i <- i + 1
      l <- find.kth.element(SI,k)
      high <- c(high,l)
      if (sum(high)/N >= a*sum(SI)/N) break
    }

    ksize[j,1]<-i
    ksize[j,2]<-profile.freq(x)
}

plot(-log10(ksize[,2]),ksize[,1],main="Size of subset (SI) for alpha = 0.9",
xlab="-log10(p)",ylab="Size of subset")
```

Profile-centred method

```
data(freqsNLsgmplus)

# set db sample size
N = 1e5
set.seed(100)

# sample a small reference db
db <- sample.profiles(N,freqsNLsgmplus)
ksize<-matrix(,nrow=100,ncol=2)
```

```

for(i in 1:100){
  x <- sample.profiles(1, freqsNLsgmplus)

  # compute SI for all database members
  SI <- ki.db(x,db,"FS")

  # compute cdf for SI of true sibling
  cdf.fs <- ki.cdf(x,hyp.1="FS",hyp.true="FS")
  cdfinv <- inverse(cdf.fs)

  # set efficiency and find t_alpha
  a <- 0.9
  t <- cdfinv(1-a)

  ksize[i,1]<-sum(SI>=t)
  ksize[i,2]<-profile.freq(x)
}

plot(-log10(ksize[,2]),ksize[,1],main="Size of subset (SI) for different profile
frequencies, alpha = 0.9", xlab="-log10(p)",ylab="Size of database")
```

§5.2.2

Conditional method

```
data(freqsNLsgmplus)

# set db sample size
N = 1e5
set.seed(100)

# sample a small reference db
db <- sample.profiles(N,freqsNLsgmplus)

ksize<-matrix(,nrow=100,ncol=100)
for(j in 1:100)
{
  x <- sample.profiles(1, freqsNLsgmplus)

  # compute SI for all database members
  SI <- ki.db(x,db,"FS")

  m<-1
  # set minimum efficiency
  for(a in seq(from=0.01,to=1,by=0.01)){
    # CONDITIONAL METHOD
    # we take uniform priors
    i <- 0
    high <- c()

    for (k in 1:N)
    {
      i = i + 1
      l <- find.kth.element(SI,k)
      high <- c(high,l)
      if (sum(high)/N >= a*sum(SI)/N) break
    }

    ksize[j,m]<-i
    m<-m+1
  }
}

k<-colMeans(ksize)
plot(seq(from=0.01,to=1,by=0.01),log(k),main="Log size of subset (SI) for different
values for alpha", xlab="Efficiency alpha",ylab="Log(Size of subset)")
```

Profile-centred method

```
data(freqsNLsgmplus)

# set db sample size
N = 1e5
set.seed(100)

# sample a small reference db
db <- sample.profiles(N,freqsNLsgmplus)
ksize<-matrix(,nrow=100,ncol=100)

for(i in 1:100){
  x <- sample.profiles(1, freqsNLsgmplus)

  # compute SI for all database members
  SI <- ki.db(x,db,"FS")

  # compute cdf for SI of true sibling
  cdf.fs <- ki.cdf(x,hyp.1="FS",hyp.true="FS")
  cdfinv <- inverse(cdf.fs)

  m<-1
  # set efficiency and find t_alpha
  for(a in seq(from=0.01,to=1,by=0.01)){
    t <- cdfinv(1-a)
    ksize[i,m]<-sum(SI>=t)
    m<-m+1
  }
}
```

```

k<-colMeans(ksize)
plot(seq(from=0.01,to=1,by=0.01),log(k),main="Average size of subset (SI) for different
efficiencies",xlab="Efficiency",ylab="Log(Size of subset)")
```

9 References

[1] Butler, J. M. (2010). *Fundamentals of Forensic DNA Typing.* London: Academic Press.

[2] Meester, R. and Sjerps, M. (2003). *The Evidential Value in the DNA Database Search Controversy and the Two-Stain Problem.* Biometrics 59, 727-732

[3] Slooten, K. and Meester, R. (2014). *Probabilistic strategies for familial DNA searching.* Journal of the Royal Statistical Society 63, 361-384

[4] Stockmarr, A. (1999). *Likelihood Ratios for Evaluating DNA Evidence When the Suspect is Found Through a Database Search.* Biometrics 55, 671-677

[5] Sjerps, M. and Kloosterman, A.D. (1999). *On the consequence of DNA profile mismatches for close relatives of the suspect.* International Journal of Legal Medicine 112, 176-180

[6] https://cran.r-project.org/web/packages/DNAprofiles/index.html

[7] Van der Beek, C.P. (2015). *Jaarverslag 2015: Nederlandse DNA-databank voor strafzaken.* Retrieved from Netherlands Forensic Institute: https://dnadatabank.forensischinstituut.nl/dna_dossier/jaarverslagen_dna_databank/

[8] Winterman, P. (2016). *Dankzij je bloedeigen broer de cel in.* Algemeen Dagblad, 26 May, 10-11

[9] Blom, M., Oudhof, J., Bijl, R.V. and Bakker, B.F.M. (2005). *Verdacht van criminaliteit: allochtenen en autochtonen nader bekeken.* Retrieved from Dutch Central Bureau of Statistics: https://www.cbs.nl/NR/rdonlyres/FA26E32C-250B-4D1D-B3B9-35085B6C48BC/0/2005verdachtvancriminaliteit.pdf