Allard Veldman

Evidential strength of Y-STR haplotype matches in forensic DNA casework

Master thesis, defended on September 28, 2007

Thesis advisors:
Erik van Zwet (Universiteit Leiden)
Marjan Sjerps (Nederlands Forensisch Instituut)

Mathematisch Instituut, Universiteit Leiden

Nederlands Forensisch Instituut

This thesis was written at the Netherlands Forensic Institute (NFI) between January 1 and August 31, 2007.

Contents

Summary

If a match is found between two Y chromosome STR profiles in a forensic case, a number of methods are available to evaluate this match, based on a large database of known profiles: the simple counting method, the more elaborate haplotype surveying and mismatch distribution methods, and a suggestion for dealing with rare haplotypes made by Charles Brenner.

The details of each method are discussed, as well as its advantages and disadvantages. A new approach called the *high-profile distribution* is introduced, which, like Brenner's suggestion, tries to make use of the fact that the database contains many unique profiles. Finally, the solutions provided by all of the methods are analyzed and compared.

The differences are small, except when the matching profile is very rare in (or even absent from) the database. Since for such profiles, none of the methods provides a very reliable figure, one could use any of the methods, provided that one explains to the court that there is a high level of uncertainty in one's answer. I advise to use the counting method, for two reasons. First, it is by far the easiest one to explain in court, allowing all parties to understand the strength of the evidence. Second, the other conventional methods (haplotype surveying, the mismatch distribution, and Brenner's correction for rare profiles) all contain (minor) errors, the implications of which are not so clear, while the high-profile distribution is still in an experimental phase.

Thus, I suggest to report to the court a frequency table, in which the numbers of copies of the crime scene profile in a few relevant databases (for instance, the total database available at www.yhrd.org, and the European, Western European or Dutch sub-databases, each database expanded with one copy of the crime scene profile) are listed, accompanied by the corresponding frequency estimates. In addition, one should stress the fact that paternal relatives of the suspect in general have the exact same Y-STR-profile.

In the final chapter, some special cases are treated, and suggestions for further research are provided.

Chapter 1

Introduction

DNA analysis plays a major part in modern forensic research. Due to mutation and recombination, no two persons have the same DNA (except for identical twins), and this enables us to determine if a cell sample, for example from a blood or semen stain, was left by a suspect or victim, provided that a cell sample from that person is also available. Unfortunately, a complete human DNA sequence consists of 6 billion symbols (A, C, G, or T, which stand for the *nucleotides* adenine, cytosine, guanine, and thymine) and decoding it is very expensive and time-consuming. Instead, the standard procedure is to type and compare only a few small parts of the DNA sequence, preferably those parts that have the highest discriminative power, i.e. that are most likely to differ between two persons. The positions at the genome where these parts are located are called *loci*, and the different variants of the DNA sequence at a locus are called *alleles*. For a general introduction to DNA, see Butler [2005] or Jobling & Tyler-Smith [2003].

The most widely used parts of the genome are the so-called short tandem repeats (STRs), strands of DNA that consist of a sequence of $2-6$ nucleotides, repeated a number of times. In contrast with single nucleotide polymorphisms (SNPs, where a single nucleotide has at one time mutated into another one), these STRs have high mutation rates, typically about once in every 1000 generations, so that the number of repeats varies across a population. Another advantage of STRs is that typing them is relatively easy using a polymerase chain reaction (PCR), which requires only a few DNA molecules to start with and produces billions of copies of a fragment of choice within a few hours. The number of repeats can then be determined by using electrophoresis to measure the length of the amplified fragment.

Because all autosomes (chromosomes other than the sex chromosomes) come in pairs, PCR analysis produces two numbers for each locus, denoting the numbers of repeats within the two copies of the same fragment on the two chromosomes; these numbers can of course be the same. A person's *DNA profile* is the collection of his repeat numbers on all typed loci. Since the frequency of each number of repeats within the population can be inferred from a database, one can calculate the probability that a person drawn at random from the population has the same pair of repeat numbers (see Buckleton [2005], Evett & Weir [1998]).

Due to the recombining nature of DNA and because target STRs are usually chosen on different chromosome pairs, repeat numbers corresponding to different loci can be considered more or less statistically independent (although one needs to adjust for the effects of structured populations, in which mating occurs primarily between members the same subpopulation; see Balding [2005), and as a result the probabilities can, with the proper corrections for population structure, be multiplied to obtain the probability that a random person has the exact same DNA profile. The standard set of loci used in the Netherlands guarantees that this probability will always be smaller than 10^{-9}, which in most cases is small enough to remove all reasonable doubt that identical profiles originated from the same source.

In some cases, however, it is impossible to obtain a complete profile from a sample. If the sample is a mixture of DNA from two persons, for example, and the amounts of material left by one of them is more than 10 times greater than the amount left by the other one, then the main contributor's profile can completely obscure the other one. But if the main contributor is female and the minor one is male, as is common in samples from rape cases, there is one piece of DNA that *can* be attributed to the male donor: the Y chromosome.

A good way to obtain an STR-profile for this person is thus to choose all target loci on the Y chromosome. The resulting profile is referred to as a Y-STR profile. However, the evidential strength of this profile is not to be confused with that of an autosomal profile.

First of all, every man has only one Y chromosome, so each locus provides just one number of repeats; for this reason, a Y-STR profile is also called a Y *haplotype*. But more importantly, all STR loci must lie on the non-recombining part of the Y chromosome (otherwise, the female contributor would have a copy as well), so they do not experience independent inheritance like autosomal STRs. A son always inherits his father's exact Y-STR profile, a few possible mutations excepted.

This has some serious implications: if a suspect matches a Y-STR profile obtained from a stain, this means that all his brothers, paternal cousins, and so on, probably match this profile too, along with an unknown number of men who share a more distant paternal ancestry with the suspect. Even for totally unrelated persons (if such a concept makes sense, since everybody is related if traced back in time far enough), the match probability for a Y-STR profile is much higher than for an autosomal profile, because the match probabilities for the various loci cannot be multiplied.

The evidential strength of a Y haplotype match is the subject of this thesis. The reasons for this investigation at this moment are the fact that the Netherlands Forensic Institute (NFI) recently started performing Y-STR typing in casework, and the lack of a scientifically accepted method to evaluate the evidential strength.

Besides giving an overview of the current opinions on this subject and introducing a new method, I will therefore present a recommendation on how to report a match between two Y haplotypes.

1.1 Data

The main source of information on frequencies of Y haplotypes is the huge Y Chromosome Haplotype Reference Database available at www.yhrd.org [Willuweit & Roewer 2007], maintained and regularly updated by Sascha Willuweit and Lutz Roewer of the Institute of Legal Medicine and Forensic Sciences, Humboldt University, Berlin, Germany. This database consists of Y haplotypes submitted by laboratories from all over the world, each of which participated in a quality control exercise before being allowed to file detected haplotypes.

At the moment of publication of this thesis, release 22 of this database has been published, containing $52,655$ from 464 populations, all typed for the seven STR loci DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393. $50,867$ of those have also been typed for the twin locus DYS385a/b [1], and $23,981$ of those have additionally been typed for the loci DYS438 and DYS439. $26,395$ of the haplotypes in the database are from Europe, and all but 90 of these have been typed for the 9-locus *minimal haplotype*, consisting of the aforementioned loci excluding DYS438 and DYS439.

Six samples from the Netherlands are included in the YHRD database: one from each of the provinces of Zeeland, Limburg, Groningen, and Friesland, one from the city of Leiden, and one from different parts of the country, containing 371 haplotypes in total; all of these samples were assembled by Peter de Knijff from the Forensic Laboratory for DNA Research (FLDO) in Leiden.

The database allows us to search for a particular haplotype in all constituent population samples, and also provides frequency estimates for the Western European, Eastern European, and South-Eastern European populations, using the haplotype surveying method (see section 2.3).

A version of the European part of the database from 2004 is available for download, listing 12727 individual haplotypes and the population samples they were found in. I will use this version for my comparison of the various frequency estimation methods, since the current database (version 22, August 10, 2007)

[1] The twin locus DYS385a/b produces two repeat numbers, corresponding to two sequences at different locations on the Y chromosome, which cannot be told apart by conventional PCR analysis. The two alleles should therefore be separated by a hyphen, e.g. DYS385*11-14, following the guidelines of the ISFG [Gusmão et al. 2006].

cannot be downloaded.
In this 2004 database, the most common profile, consisting of the repeat numbers $14-13-29-24-11-13-13$ at the seven standard loci, occurs 661 times and therefore has a database frequency of $\frac{661}{12727} = 5.19 \cdot 10^{-2}$. This profile is the center of a cluster of frequent profiles that are genetically close to it. Another group of frequent profiles is clustered around profile $14-12-28-22-10-11-13$, which has frequency $\frac{271}{12727} = 2.13 \cdot 10^{-2}$.
Locally, frequencies can be much higher. An extreme example is the Finnish population, which forms part of our database; in this population of size 399, the most frequent profile $(14-14-30-11-14-14)$ occurs 100 times, while there are only 25 copies of this profile in the rest of the European database.

Data on mutation rates for all of these loci are available from a number of studies [Willuweit & Roewer 2007]. The total numbers of observed mutations per locus, along with 95%-confidence intervals for the mutation rates, are shown in Table 1.1.

Locus	Father-son pairs	Mutations	Mutation rate (x 10^{-3})	95% C.I. (x 10^{-3})
DYS19	8944	22	2.46	1.54 – 3.72
DYS389I	7148	13	1.82	0.97 – 3.11
DYS389II	7135	19	2.66	1.60 – 4.16
DYS390	8426	20	2.37	1.45 – 3.66
DYS391	8375	25	2.98	1.93 – 4.40
DYS392	8339	4	0.48	0.13 – 1.23
DYS393	7128	6	0.84	0.31 – 1.83
DYS385a/b	13468	30	2.23	1.50 – 3.18
DYS438	3887	2	0.51	0.062 – 1.86
DYS439	3864	22	5.69	3.57 – 8.61

Table 1.1: Mutation rates, obtained from www.yhrd.org, visited on August 13, 2007

The estimated mutation rates range from $0.48 \cdot 10^{-3}$ to $5.69 \cdot 10^{-3}$, with mean $2.20 \cdot 10^{-3}$ and median $2.30 \cdot 10^{-3}$.

Chapter 2

Existing methods for evaluating Y haplotype matches

In this chapter I will give an overview of all methods that are currently being used or developed for estimating match probabilities of Y chromosome profiles. First, I will recapitulate the situation of interest and introduce some notations.
In the analysis of a DNA sample recovered from a crime scene, a Y chromosome trace can be detected (this is done with a so-called amelogenin test, incorporated in standard autosomal DNA profiling). However, either no full autosomal profile of this man – whom I will call C, since in most cases he is the culprit – can reliably be obtained, or the information provided by the autosomal profile is not yet conclusive. A Y-STR profile $\pi = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7)$ is derived as well, where the π_i are the respective numbers of repeats at the core loci DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393; the twin locus DYS385a/b is omitted for the rest of this analysis, because the haplotype surveying method uses genetic distances between haplotypes, which are ambiguous for this locus.
Suspect s is also typed for these loci, and is discovered to have profile π as well. This match could indicate that the crime scene sample originated from s, but it could also have been left by someone else sharing the same profile, either coincidentally or by common ancestry. The important question is: what is the strength of the DNA evidence?
Before discussing the answers that the various methods provide to this question, I will present the weight-of-evidence theory common to all methods, which can, for instance, be found in more detail in the book by Balding [2005].

2.1 Weight-of-evidence theory: likelihood ratios

Write $A \equiv \alpha$ if individual A has DNA profile α, $A \equiv B$ if individuals A and B have the same DNA profile, and denote by $\mathcal{S}$ the population of all possible suspects. Then a finder of fact needs to evaluate the probability that s is the culprit given the evidence, i.e. $P(C = s|C \equiv s \equiv \pi)$. Equivalently, because $P(C = s|C \equiv s \equiv \pi) + P(C \neq s|C \equiv s \equiv \pi) = 1$, he could also evaluate the odds against s:

$$\frac{P(C \neq s|C \equiv s \equiv \pi)}{P(C = s|C \equiv s \equiv \pi)} = \sum_{\substack{i \in \mathcal{S} \\ i \neq s}} \frac{P(C = i|C \equiv s \equiv \pi)}{P(C = s|C \equiv s \equiv \pi)}.$$

Now, applying Bayes theorem, we see that for an individual i,

$$\frac{P(C = i|C \equiv s \equiv \pi)}{P(C = s|C \equiv s \equiv \pi)} = \frac{P(C \equiv s \equiv \pi|C = i)}{P(C \equiv s \equiv \pi|C = s)} \frac{P(C = i)}{P(C = s)}$$

$$= \frac{P(i \equiv s \equiv \pi|C = i)}{P(s \equiv \pi|C = s)} \frac{P(C = i)}{P(C = s)}$$

We assume that *a priori*, the identity of the culprit doesn't affect anyone's probability of having profile π, so we can leave out the conditioning on $C = i$ and $C = s$:

$$\frac{P(i \equiv s \equiv \pi|C = i)}{P(s \equiv \pi|C = s)} \frac{P(C = i)}{P(C = s)} = \frac{P(i \equiv s \equiv \pi)}{P(s \equiv \pi)} \frac{P(C = i)}{P(C = s)}$$

$$= P(i \equiv \pi|s \equiv \pi) \frac{P(C = i)}{P(C = s)}.$$

We conclude that for every possible suspect i, two quantities need to be assessed: first, the prior probability $P(C = i)$ that i is the culprit, relative to $P(C = s)$, and second, the conditional probability that i has profile π, given that s has it; both probabilities should be conditioned on all other evidence. Since the first assessment is up to the judge, a forensic DNA expert need only consider the second one, called the *match probability.* Of course, this probability depends on the genetic relationship of i and s, since related individuals are more likely to have matching profiles than unrelated ones. This effect is even stronger for Y-STR profiles than for autosomal profiles.

The match probability for paternal relatives is easy to derive from the mutation rates μ_i in Table 1.1: if the relatives are k steps apart in a male family tree (e.g. $k = 2$ for a paternal grandfather and grandchild, or two brothers), the probability that no mutation has taken place is $\prod_{i=1}^{7}(1-\mu_i)^k \approx 0.9865^k$ (assuming that mutations are independent events), and since the probability of multiple mutations canceling each other out is negligible, 0.9865^k is also the probability that the two profiles are equal.

For most members of the suspect population, there will be no known paternal relationship to the suspect, so we cannot use mutation probabilities. Hence, a reasonable thing to do is to estimate the frequency of haplotype π in the suspect population, or in a bigger population from which the suspect population was sampled. We will assume that our European database from 2004 (containing

12727 haplotypes) can serve as this bigger population, postponing the discussion regarding this assumption to chapter 6.
Under this assumption, we are regarding individuals i and s as independent samples from the European population, so the information we obtain about i from the observation of the suspect's profile π is exactly the same as from the observation of any of the profiles in the database, because all are independent samples from the same population. This means that we can add the suspect's copy of π to the database, and use this extended database of size $n = 12728$ as the basis for our estimates.
We will denote by $\hat{p}_j$ the estimated population frequency of a profile j, and by f_j the absolute frequency of j in this extended database.

2.2 Counting method

The first method consists of simply estimating the population frequency as equal to the observed frequency in the extended database. This amounts to the formula

$$\hat{p}_\pi = \frac{f_\pi}{n}.$$

An obvious advantage of this calculation is its simplicity, which is an advantage for explaining the method in court. Also, since it is just the classical maximum likelihood estimator (MLE) for multinomial samples, it doesn't make any assumptions about population genetics, and thus represents only factual information contained in the database.
However, there is a wealth of extra information provided by this very database, and the failure of the counting method to make use of this information can be seen as a disadvantage. Moreover, this estimator only produces reliable results if both the sample size and the number of observed copies of π in this database are big enough. In the worst case, the suspect's copy of π can be the only one in the database. Since it is considered better to give *conservative* estimates (frequency estimates that are higher than the actual values, favouring the suspect and minimizing the probability that innocent people are convicted) than non-conservative ones, we could resolve this by reporting the upper 95% confidence limit. If f_π is indeed equal to 1, this 95% confidence limit is $3.73 \cdot 10^{-4}$ (calculated numerically using the binomial distribution, solving $(1-p)^{12728} + 12728p(1-p)^{12727} = 0.05$), much higher than the MLE of $\frac{1}{12728} \approx 7.86 \cdot 10^{-5}$.

2.3 Haplotype surveying

L. Roewer et al. [2000] feel that this conservative 95% confidence limit dramatically reduces the power of Y-STR haplotyping, and therefore propose another method, called *haplotype surveying*. They use a Bayesian approach, deriving a prior distribution of the frequency of a haplotype from its genetic distance to the other haplotypes in the database, and then transforming it into a posterior distribution via the likelihoods of observing the actual number of copies in the database under the assumed prior probabilities.

According to Roewer et al., classical population genetics theory tells us that the prior distribution of P_i, the population frequency of profile i, should be a $\beta(u_i, v_i)$-distribution with density function

$$\phi_i(p) = \frac{\Gamma(u_i + v_i)}{\Gamma(u_i) \cdot \Gamma(v_i)} p^{u_i - 1}(1-p)^{v_i - 1},$$

where u_i and v_i are profile-specific parameters.

Instead of u_i and v_i, the mean μ_i and standard deviation σ_i of the prior distribution are estimated. Since

$$\mu_i = \frac{u_i}{u_i + v_i} \text{ and } \sigma_i^2 = \frac{u_i v_i}{(u_i + v_i)^2 (u_i + v_i + 1)},$$

u_i and v_i can be derived from μ_i and σ_i:

$$u_i = \frac{\mu_i^2 (1 - \mu_i)}{\sigma_i^2} - \mu_i \text{ and } v_i = u_i \frac{1 - \mu_i}{\mu_i}.$$

In order to estimate μ_i and σ_i, the authors suggest that both parameters depend on the genetic distance of profile i to all the other profiles, more precisely on the weighted inverse molecular distance

$$W_i = \frac{1}{n} \sum_{j \neq i} \frac{f_j}{d_{ij}},$$

where d_{ij} is the minimum number of mutation steps separating profiles i and j, in this case adopting the single-step model. So $\mu_i = \mu(W_i)$ and $\sigma_i = \sigma(W_i)$. Each value of W_i thus determines μ_i and σ_i of the prior distribution.

To estimate these, all European haplotypes in the database are divided into 15 equally sized groups, according to W-value. Each set of observed frequencies corresponding to one of these groups is then treated as a sample from a single β-distribution, the parameters μ and σ of which are estimated by the sample mean and standard deviation. These values are taken as estimates for $\mu(\bar{W})$ and $\sigma(\bar{W})$, where $\bar{W}$ is the average value of W for this group.

To the set of fifteen triples $(\bar{W}, \mu, \sigma)$ that arises from this process, 'exponential' regression is applied to obtain the two functions $\mu(W)$ and $\sigma(W)$:[2]

$$\mu(W) = 1.11 \cdot 10^{-4} + e^{41.20W - 11.30}$$

[2] These formulas are based on a version of the YHRD database from September 1999, comprising 2439 haplotypes; the latest formulas for the Western European population are $\mu(W) = e^{36.3543W - 13.7255}$ and $\sigma(W) = e^{35.1207W - 14.0974}$.

$$\sigma(W) = 2.37 \cdot 10^{-4} + e^{30.86W-9.22}$$

The likelihood of observing f_i copies of a haplotype i in a database of n individuals, as a function of the prior frequency P_i, is a binomial probability given by

$$P(f_i|P_i) = \binom{n}{f_i} P_i^{f_i}(1-P_i)^{n-f_i},$$

so the posterior distribution of f_i has density function

$$\phi_i'(p) = \frac{P(f_i|p)\phi_i(p)}{\int_0^1 P(f_i|q)\phi_i(q)dq},$$

which is a $\beta(u_i + f_i, v_i + n - f_i)$-distribution.

In the original article [Roewer et al. 2000], the formulas were slightly different than presented here, due to the fact that at first, the authors only considered haplotypes already present in the database, and thus removed one copy of each haplotype because the first observation only 'indicated its existence'. M. Krawczak corrects this mistake in a comment [Krawczak 2001] and also explains the need for the addition of one copy of the suspect's profile π to form the extended database in forensic casework, as discussed before.
The haplotype surveying method produces a posterior frequency distribution, instead of a point estimate. If one would like to have such an estimate, the mean of this distribution, equal to $\frac{u_i+f_i}{u_i+v_i+n}$, is the most logical choice. There are, however, other possible choices, like the mode (Krawczak suggests this, but his argument seems to miss its point[3]) or a 95% credibility interval; one could even provide a graph of the posterior distribution.

2.4 The mismatch distribution

Luisa Pereira et al. [2000] have suggested that calculating a haplotype's frequency is not enough to estimate a match probability, because according to them, this procedure is based on the 'dichotomy equal/not-equal', whereas both categories are heterogeneous: 'equal' could mean that two haplotypes are either identical by descent (both having descended from the same ancestor without experiencing

[3] Krawczak draws an analogy with a coin that is taken at random from a set of three coins, two of which always show heads and one of which always shows tails. The coin in our hands can be any of these three, so the probability of one toss showing heads is $\frac{2}{3}$. However, when asked for the outcome of ten tosses, Krawczak reasons that the most sensible guess is '10 times heads'. If one wants to maximize the probability of giving the right answer, this is indeed the most sensible guess, but in terms of minimum squared error, '$6\frac{2}{3}$ times heads' is better. The number of tosses is irrelevant for this argument; it simply comes down to a choice between the mean and the mode, and the coin example does not help us to make that choice for the posterior frequency distribution described above.

any mutation) or just identical by state, which means that they have experienced mutations, but ended up equal after all. A frequency estimate only produces the sum of these two probabilities and is thus misleading.

However, it is this suggestion that is misleading, since the distinction between identity by descent or by state is not at all relevant to our problem. According to the formula derived in the weight-of-evidence section 2.1, we should try to estimate the probability of one individual having haplotype π, conditional on another individual having it. For this equality, it does not matter if their haplotypes are identical by descent or by state, since from the evidence we can never tell which is the case.

The authors conclude their article by warning us not to calculate frequency estimates without taking their 'mismatch distribution' into account, but they fail to specify how this distribution should affect our estimate. We therefore will not investigate this paper any further.

2.5 Charles Brenner's suggestion for dealing with rare haplotypes

Charles Brenner has suggested another method of calculating match probabilities for *singletons*, haplotypes that occur only once in a database. The suspect's haplotype π can be such a singleton, if it was absent from the original database. Brenner considers the fraction κ of singletons in the extended database, including π. Referring to an article by H. Robbins [1968], he reasons that the probability that the next individual sampled will have a new profile (that is, a profile that did not yet occur in the database), can be estimated by κ. He uses this information to calculate the match probability for an individual i unrelated to the suspect s in the following way.

Denote by A the event that the profile of i is already in the extended database, and by B the event that it is equal to π. Then according to Robbins, $P(A) = 1-\kappa$; and as there are n profiles in the database, $P(B|A) = \frac{1}{n}$. Therefore

$$P(B) = P(B|A) \cdot P(A) = \frac{1}{n} \cdot (1-\kappa) = \frac{1-\kappa}{n}$$

Brenner uses the American database presented by Budowle et al. [2005] and finds that for this database, $\kappa = 0.9$, so the match probability for an unobserved profile would be $\frac{1}{10n}$. In our much bigger European database of size $n = 12728$, there are 1396 – 1398 singletons (depending on the number of copies of π in the original database), so $\kappa \approx 0.11$ and if π is a singleton, then the match probability is $\frac{1-0.11}{12728} \approx 6.99 \cdot 10^{-5}$.

However, as Brenner himself points out, if we extend this argument to all haplotypes present in the database, there seems to be something wrong; I will discuss this in more detail in the next chapter.

Chapter 3

Comments

In this chapter I will discuss the validity of the two most promising methods, the haplotype surveying method and Brenner's modification of the counting method, as well as their advantages and disadvantages. The idea behind the latter method will serve as a basis for the high-profile method, a new approach presented in chapter 4.

3.1 Haplotype surveying

The haplotype surveying method is a Bayesian one, i.e, it starts with a prior distribution of a profile's frequency, and then updates it with information obtained from a sample. Both stages need to be executed correctly in order for the posterior distribution to be correct. First, let's discuss the prior distribution.

The effect of the prior distribution

A question one could ask first is whether the exact form of the prior is really that important: often the data is so overwhelming that the prior distribution is "swamped" by it. When the likelihood of the data under the assumption of one frequency is a thousand times bigger than under the other, the former will have a much larger probability in the posterior distribution, unless the prior substantially favours the latter. Such informative priors effectively rule out certain values of the population frequency, so we should have great confidence in our prior beliefs if we wish to use them.
The prior distribution Roewer et al. [2000] use is $\beta(u, v)$, where u and v depend on W, the mean inverted distance to the other profiles. For the two extreme W-values discernible in Figure 3 of their article, $W = 0.02$ and $W = 0.14$, (u, v) equals $(0.109, 785.7)$ and $(0.275, 67.3)$, corresponding to $(\mu, \sigma) = (1.39 \cdot 10^{-4}, 4.21 \cdot 10^{-4})$ and $(4.07 \cdot 10^{-3}, 7.69 \cdot 10^{-3})$ respectively; in comparison, a uniform distribution corresponds to $(u, v) = (1, 1)$. Transformation

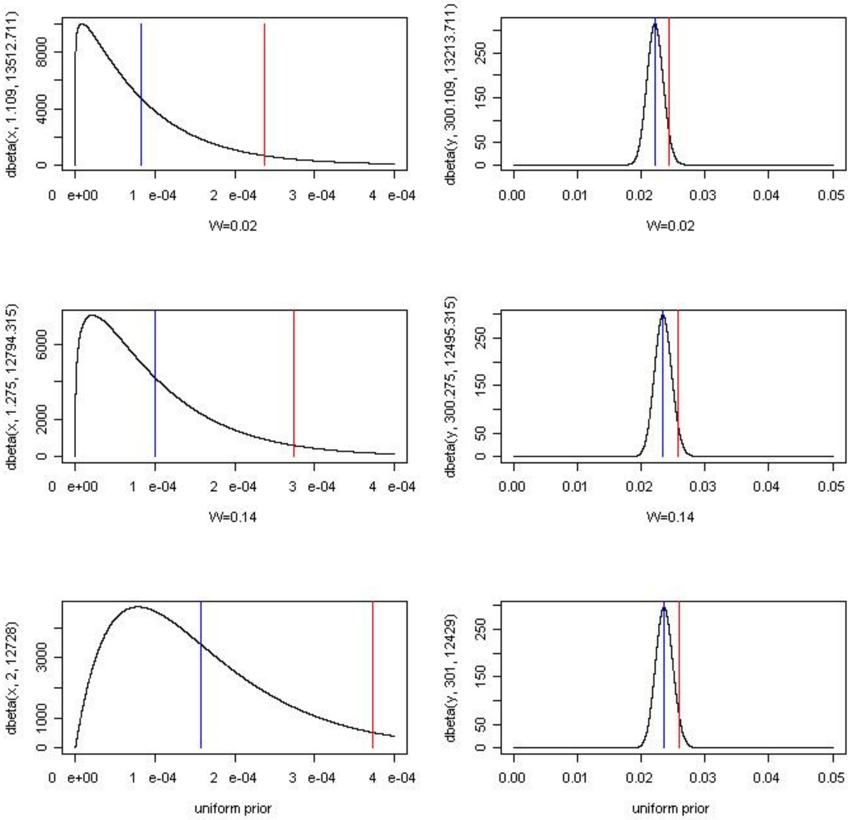

Figure 3.1: Illustration of the possible influence of the prior distribution on the posterior distribution. The three posterior functions on the left correspond to a haplotype with one copy in the database, the three distributions on the right to one with 300 copies.

from the prior to the posterior distribution simply consists of adding to u and v the observed counts of the profile of interest and of all other profiles combined. To illustrate the effect the choice of the prior may have, let's draw graphs of the posterior distributions corresponding to these three priors, both for a profile occurring only once in a database of 12728, and for a profile occurring 300 times.

The blue vertical line indicates the position of the mean of the posterior distribution, the red line the position of the 95%-quantile. As is illustrated by these plots, the choice of the prior has little effect on the common profile (the posterior mean ranging from 0.0222 for $W = 0.02$ to 0.0236 for a uniform prior), but this is not the case for the rare profile: here, the posterior mean almost doubles, from $8.21 \cdot 10^{-5}$ (or 1 in $12,200$) for $W = 0.02$ to $1.57 \cdot 10^{-4}$ (1 in $6,400$) for the uniform prior. We conclude that, certainly for rare profiles, it is worthwhile to have a close look at the prior distribution.

The choice of the prior distribution

First of all, the choice of the beta distribution is not so obvious. The real stationary distribution of any finite population model, after all, should be discrete, since allele frequencies must be multiples of $\frac{1}{N}$, where N is the male population size. However, this discrete distribution is already hard to compute for relatively small populations, let alone for large populations like the European or Dutch ones. S.G. Wright has developed a formula for a continuous approximation [Wright 1937], which comes very close when certain conditions are satisfied. This formula, based on the diffusion method, is

$$\Phi(x) = \frac{C}{V_{\delta x}} e^{2 \int \frac{M_{\delta x}}{V_{\delta x}} dx}$$

where $\Phi(x)$ is the stationary distribution of the allele frequency, and $M_{\delta x}$ and $V_{\delta x}$ are the mean and variance, respectively, of the change in allele frequency when going to the next generation, depending on the current allele frequency x. $M_{\delta x}$ depends on the mutation model, whereas the variance is due to drift only, and is given by $(x + M_{\delta x})(1 - x - M_{\delta x})/N_e$ [Wright 1937]. The latter expression, since $M_{\delta x}$ is in general very small with respect to x, is approximately equal to $x(1 - x)/N_e$, where N_e denotes the *effective* population size, i.e. the size of a theoretic random-mating population with the same sampling variance as the population of interest.

The hard part of Wright's formula is the calculation of $M_{\delta x}$. If $\mathcal{S}$ denotes the set of all alleles, A the allele we want to know the stationary frequency distribution of, F_i the frequency of allele i, and p_{ij} the probability of an i allele becoming, through mutation, a j allele in one generation (define $p_{ii} = 1 - \sum_{j \in \mathcal{S}} p_{ij}$), we can write

$$M_{\delta x} = \sum_{i \in \mathcal{S}} E[F_i | F_A = x] \cdot p_{iA} - x$$

To calculate $E[F_i | F_A = x]$, one would typically need the (unknown) stationary distribution of the state of the entire population, but for certain simple models we can do without it. I will consider two models.

1. One of the simplest mutation models one could think of consists of two alleles A and B, and mutation probabilities $p_{AB} = u$ and $p_{BA} = v$. If the frequency of allele A is x, then the frequency of the other allele must be $1 - x$; in other words, $E[F_B | F_A = x] = 1 - x$, so $M_{\delta x} = x \cdot (1 - u) + (1 - x) \cdot v - x = -u \cdot x + v \cdot (1 - x)$.

Wright's equation then yields

$$\begin{aligned}\Phi(x) &= \frac{C}{V_{\delta x}} e^{2\int \frac{M_{\delta x}}{V_{\delta x}} dx} \\ &= \frac{N_e C}{x(1-x)} e^{2\int \frac{N_e(-u\cdot x + v\cdot(1-x))}{x(1-x)} dx} \\ &= \frac{N_e C}{x(1-x)} e^{2\int (\frac{-N_e u}{1-x} + \frac{N_e v}{x}) dx} \\ &= \frac{N_e C}{x(1-x)} e^{2(N_e u \log(1-x) + N_e v \log x)} \\ &= \frac{N_e C}{x(1-x)} (1-x)^{2N_e u} x^{2N_e v} \\ &= N_e C (1-x)^{2N_e u - 1} x^{2N_e v - 1}.\end{aligned}$$

Since $\Phi(x)$ is a probability density function, we should have $\int_{x=0}^{x=1} \Phi(x) dx = 1$, so $C = \frac{\Gamma(2N_e(u+v))}{N_e(\Gamma(2N_e u)(\Gamma(2N_e v)}$.
$\Phi(x)$ then becomes a $\beta(2N_e v, 2N_e u)$-function:

$$\Phi(x) = \frac{\Gamma(2N_e(u+v))}{\Gamma(2N_e u)\Gamma(2N_e v)} (1-x)^{2N_e u - 1} x^{2N_e v - 1}.$$

Hence, for two alleles, a beta distribution seems plausible, provided that the population has reached equilibrium.

2. For a one-dimensional infinite alleles stepwise mutation model, Kimura and Ohta [1978] have found an approximation to $M_{\delta x}$:

$$M_{\delta x} \approx -vx + \frac{v}{2}(1-x)(b_0 + b_1 x),$$

where $\frac{v}{2}$ is the mutation rate from each allele to one of its neighbouring alleles, and b_0, b_1 are constants that depend on $N_e v$. Inserting this approximation into Wright's formula yields

$$\Phi(x) = C e^{N_e v b_1 x} (1-x)^{2N_e v - 1} x^{N_e v b_0 - 1},$$

which for small b_1 is close to a $\beta(N_e v b_0, 2N_e v)$-distribution. Kimura and Ohta show that for $N_e v = 0.05$, $(b_0, b_1) = (0.9314, 0.0472)$, while for $N_e v = 0.5$, $(b_0, b_1) = (0.6040, 0.3177)$. In general, the smaller $N_e v$ is, the fewer different alleles one would expect in a population (because either there are less people, or drift dominates mutation), and the more the stationary distribution will look like the beta distribution from the two alleles model. As we saw in the introduction, STR mutation rates are typically high, so even for relatively small populations, one should not use the beta distribution without checking whether it is a good approximation.

Now we should try to derive $M_{\delta x}$ for the more complex multi-dimensional stepwise mutation model; it would be nice if the resulting stationary distribution would still look like a beta function, because that would validate the haplotype surveyors' assumptions. I have tried to derive this generalization, but unfortunately, it turned out to become too complex for an analytical treatment. Hence, the assertion of Roewer et al. [2000] that the choice of a beta prior is "standard population genetic theory" is overoptimistic. In fact, it is an unproven assumption.

For calculation purposes, it is a convenient assumption, since the class of beta distributions is conjugate to the class of binomial likelihood functions, i.e., if the prior function is beta, and the likelihood function is binomial, the posterior function will be beta too.

The formula for the weighted inverse distance

Apart from the choice of a beta prior distribution, there are some other strange aspects of the haplotype surveying method. For instance, according to the original article, the formula for the weighted inverse distance is

$$W_i = \frac{1}{N} \sum_{j \neq i} \frac{N_j}{d_{ij}}.$$

In a weighted average formula, the total weight should always be 1, but this is not the case here, since $\sum_{j \neq i} \frac{N_j}{N} = \frac{N-N_i}{N}$. Of course, we saw that the prior distribution only has a strong influence on the estimates of rare profiles, for which the difference between $N - N_i$ and N is small. Still, the correct formula, $W_i = \frac{1}{N-N_i} \sum_{j \neq i} \frac{N_j}{d_{ij}}$ should be used. According to private correspondence with M. Krawczak, this mistake has been corrected now in the calculations on the YHRD website [Willuweit & Roewer 2007].

A more fundamental issue is that the inverse genetic distance of two profiles is used as a measure of the correlation between the frequencies of the profiles; according to this measure, our prior belief in the existence of a profile A is as much influenced by finding one copy of one of A's direct neighbours as it is by finding four copies of a profile at distance 4 from A, or by finding 16 copies of a profile at distance 16. In any stepwise mutation model where mutations are equally likely in all directions, correlation between profile frequencies in the stationary distribution of a population decreases rapidly as a function of the distance between the profiles. The 16 profiles at distance 16 from A tell us almost nothing about the frequency of profile A, and should therefore have negligible effect on the calculation of W. It may be enough to consider only the direct neighbours of A, having a genetic distance of 1.

Estimating the parameters of the beta prior distribution

The final step in the development of the prior frequency distribution is the calculation of the parameters u and v, through the related parameters μ and σ of the beta distribution. They are both found by regression on W, grouping data points together to obtain average values of μ and σ for the respective W-intervals. If, however, one chooses to use group averages of data instead of single data points, one should be cautious about the outcome. If standard regression is performed, the resulting estimates for μ and σ will be biased.

To illustrate this, consider one of the groups, consisting of m profiles, numbered 1 through m. Each profile i gives us a value w_i for W and a value f_i for the observed frequency in the database. Now according to the regression model, f_i is a realization of a random variable F_{w_i} with expected value $EF_{w_i} = \mu_{w_i}$ and variance $Var(F_{w_i}) = \sigma^2_{w_i}$. Roewer et al. use the pairs (w_i, f_i) to generate two pairs $(\bar{w}, \hat{\mu}_{\bar{w}})$ and $(\bar{w}, \hat{\sigma}^2_{\bar{w}})$, where $\bar{w} = \frac{1}{m}\sum_{i=1}^{m} w_i$, and $\hat{\mu}_{\bar{w}} = \bar{f} = \frac{1}{m}\sum_{i=1}^{m} f_i$ and $\hat{\sigma}^2_{\bar{w}} = \frac{1}{m-1}\sum_{i=1}^{m}(f_i - \bar{f})^2$ are considered to be close to the real values $\mu_{\bar{w}}$ and $\sigma^2_{\bar{w}}$.

Let's compute the expected values of $\hat{\mu}_{\bar{w}}$ and $\hat{\sigma}^2_{\bar{w}}$:

$$\begin{aligned}
E\hat{\mu}_{\bar{w}} &= E[\frac{1}{m}\sum_{i=1}^{m} F_{w_i}] \\
&= \frac{1}{m}\sum_{i=1}^{m} EF_{w_i} \\
&= \frac{1}{m}\sum_{i=1}^{m} \mu_{w_i};
\end{aligned}$$

$$\begin{aligned}
E\hat{\sigma}^2_{\bar{w}} &= E[\frac{1}{m-1}\sum_{i=1}^{m}(F_{w_i} - \bar{F})^2] \\
&= E[\frac{1}{m-1}\sum_{i=1}^{m}(F^2_{w_i} - \bar{F}^2)] \\
&= \frac{1}{m-1}\sum_{i=1}^{m}(E[F^2_{w_i}] - E[\bar{F}^2]) \\
&= \frac{1}{m-1}\sum_{i=1}^{m}(E[F^2_{w_i}] - (EF_{w_i})^2 - (E[\bar{F}^2] - (E\bar{F})^2)) \\
&+ \frac{1}{m-1}\sum_{i=1}^{m}((EF_{w_i})^2 - (E\bar{F})^2) \\
&= \frac{1}{m-1}\sum_{i=1}^{m}(VarF_{w_i} - Var\bar{F}) + \frac{1}{m-1}\sum_{i=1}^{m}(\mu^2_{w_i} - (\frac{1}{m}\sum_{i=1}^{m}\mu_{w_i})^2) \\
&= \frac{1}{m}\sum_{i=1}^{m}\sigma^2_{w_i} + \frac{1}{m-1}\sum_{i=1}^{m}(\mu^2_{w_i} - (\frac{1}{m}\sum_{i=1}^{m}\mu_{w_i})^2).
\end{aligned}$$

The second term in the last equation is the dispersion of the values of $\mu(W)$ for the haplotypes in the group.

We see that if $\mu(W)$ and $\sigma^2(W)$ are convex functions of W, which is suggested by the data, $\hat{\mu}_{\bar{w}}$ and $\hat{\sigma}^2_{\bar{w}}$ are positively biased because of Jensen's inequality; the effect is even stronger in $\hat{\sigma}^2_{\bar{w}}$ due to the extra term involving the dispersion of $\mu(W)$. The biases can be avoided by simply using the single data points in the regression, first performing the regression of μ on W, then calculating the squared deviations of all data points from the resulting function $\mu(W)$ and regressing those values on W to find $\sigma^2(W)$.

The validation of the haplotype surveying assumptions

In their original article, the haplotype surveyors provide us with a validation for their method: they find that the expected number of profiles with a certain frequency, calculated from all prior distributions, is always close to the actual number of profiles with that frequency in the database. However, this resemblance is not as striking as it looks at first sight, since the histogram in which these numbers are represented uses a logarithmic scale. Thus we see that the number of profiles which occur twice (once, after subtraction of one copy of each profile) deviates some 40% from its expected value. Still, the figures seem remarkable. There is, however, good reason to be skeptic about this method of validation.

The expected numbers are calculated from the prior frequency distributions of all profiles. The means of these prior distributions, in turn, have been obtained from a regression function $\mu(W)$ which uses the database frequencies as y-coordinates for its data points. Regardless of the W-values associated with these data points, the set of μ-values of the prior distributions will therefore resemble the set of database frequencies of all profiles. Automatically, the expected number of singletons, for instance, according to these prior functions will also be close to the number of singletons in the database. Large deviations from the expected frequency could occur for individual profiles, but since the histogram only represents gross figures, this possible evidence against the haplotype surveying method is left out. While the number of singletons is what it was predicted to be, there is no way of telling whether these rare profiles also had small prior functions. The claim that W is a predictor for the frequency of a profile can therefore not be tested convincingly with a histogram of this kind.

The objections to the haplotype surveying method from the last few pages should make us cautious to use it in the present form. However, even after applying the advised corrections, we still have to consider one important issue. When using a beta distribution, or any other prior distribution resulting from a population model, we implicitly assume that our population has been around long enough to have reached equilibrium as a Markov chain. A population that is not yet in equilibrium will exhibit founder effects: clusters of similar profiles will show up, originating from a small number of individuals. In such a population, there will be strong correlations between frequencies of neighbouring profiles, since the founder's profiles have only had the time for one or two mutations, resulting in a dense cloud of profiles, clustered around the original profile.

This explanation for the dependence of μ on W is much more plausible than the equilibrium one in our situation, since the history of Europe shows continuous migration at a large scale, which introduces new profiles to the population all the time and may outweigh the effects of mutation and drift. Also, because there is evidence that the size of the population varied to a large extent in the past, the people who lived when it was small for the last time (the last *population bottleneck*) are likely to have caused a founder effect.

3.2 Brenner's method

Unlike the haplotype surveying method, the counting method does not make use of genetic information, like distances between profiles, or of any mutation model. Instead, it only looks at the frequencies of the profiles in the database to estimate the corresponding population frequencies. Thus, the problem reduces to that of estimating probabilities from a sample of n coloured marbles drawn from an urn, with replacement (actually, without replacement, but the population size is so big that this doesn't matter).

The maximum likelihood estimator for the probability of a specific colour is the one used in the counting method, equaling the sample frequency. However, this estimator performs poorly when the sample size is small compared to the number of colours, which is true in our case. All estimated frequencies are multiples of $\frac{1}{n}$, so it is impossible to obtain an accurate estimate for a profile with a true population frequency significantly less than this number.

Also, the maximum likelihood estimator distributes all probability mass over the observed profiles. Since there are bound to be some profiles that do exist in the population, but do not occur in the database, the MLE effectively denies their existence and consequently overestimates, on average, the frequencies of the observed profiles.

To adjust for this effect, we need an estimate for the total probability of the unseen profiles. Brenner's κ, based on Robbins's article [Robbins 1968], is in fact an unbiased estimator for this probability. The main problem, however, is how to distribute the remaining probability mass of $1 - \kappa$ over the observed profiles. One option is to do this proportionately to the database counts, like Brenner does. This leads to estimates $\frac{(1-\kappa)f_i}{n}$ for profiles with database frequencies f_i. Assuming that these adjusted estimates are close to the real population frequencies, we draw a remarkable conclusion: all profiles that have been observed, including the most common ones, have been observed $\frac{1}{1-\kappa}$ times as often as one would expect in a representative sample! This feature is inherent in the method, and it leads to nonconservative estimates for common profiles.

One could wonder if the same objection applies to rare profiles, since it is also very unlikely that a large number of profiles with true frequency $\frac{1-\kappa}{n}$ are all observed once, still $\frac{1}{1-\kappa}$ times as often as expected. However, there is an essential difference here, since the observed singletons are probably part of a larger group of low-frequency profiles, the others just having been missed in the sample. If the population consisted of six billion unique profiles, most multinomial samples of length 12728 would contain only singletons, and it would be correct to estimate the population frequencies a lot smaller than the sample frequency of $\frac{1}{12728}$.

A second way of distributing the $1-\kappa$ probability over the database profiles is the one proposed by Good [1953]. In the same way that the number of singletons is used to estimate the total probability of the unseen profiles, he uses the total frequency of the doubletons in the database to estimate the population frequencies of the singletons. If we denote by n_t the number of profiles observed t times, Good's method would estimate the frequency of each singleton at $\frac{2n_2}{n_1 n}$, which for the European database, depending on the database frequency of the suspect's profile π, approximately comes down down to $\frac{2\cdot 379}{1397\cdot 12728} \approx 0.543\cdot \frac{1}{12727} \approx 4.26\cdot 10^{-5}$. In the same way, the frequency of a profile occurring t times is estimated at $\frac{(t+1)n_{i+1}}{n_t n}$, so that each profile's estimate is derived from the profiles occurring exactly once more.

This procedure ensures that all estimates nicely add up to one, but realistic results are only obtained for low-frequency profiles, where there are a lot of profiles sharing the same frequency in the database. For common profiles, we could encounter division by zero, and moreover, it seems absurd to use one profile's count to estimate another profile's frequency. Good himself signals this flaw and suggests solving it by *smoothing* the database frequencies in some way, so that irregularities in the data are flattened. He does not specify how this smoothing should be conducted, and no one has found a satisfactory solution for this problem yet.

These problems are big enough for us to discard Good's method for the moment, and to try to improve on Brenner's method. As said before, we would like to distribute the $1-\kappa$ probability in such a way that the common profiles roughly receive their maximum likelihood probability (i.e. the database frequency), while the rare profiles get less, all estimates still adding up to 1.

There may be a canonical way to do this, suggested by Alan Orlitsky et al. [2004]. Consider a sample consisting of three balls, of which two have colour A and one has colour B. Instead of the probability of drawing two As and one B, we could calculate, for every probability distribution over the colours, the probability of drawing two identical colours and one different from these, regardless which colour was drawn twice. If we only allow two colours with positive probability, it turns out that observing exactly two identical balls in a sample of three is more likely under a $(\frac{1}{2}, \frac{1}{2})$-probability distribution than under a $(\frac{2}{3}, \frac{1}{3})$-distribution, although the latter is the maximum likelihood distribution for the specific observation of two As and one B.

This approach, in a more rigorous form, is the subject of the next chapter.

Chapter 4

The high-profile distribution

At the end of the previous chapter, we hinted at a new method of estimating haplotype frequencies based on a small sample relative to the total number of distinct haplotypes, taking into consideration that there probably are many haplotypes that have been missed in the sample. In this chapter, we will follow Orlitsky et al. [2004a] in developing this method in more detail and in providing an algorithm to calculate these improved estimates. First, we need some definitions.

Consider a sequence $\bar{x}$ of haplotypes, coded by letters from an alphabet $\mathcal{A}$, e.g. $\bar{x} = \bar{a} = cadcegcacfgbe$ (this should not be confused with a DNA sequence, where every letter stands for a nucleotide; in this sequence $\bar{a}$, each letter represents an entire Y-STR profile, for instance $c = 14 - 13 - 29 - 24 - 11 - 13 - 13$. Think of a sequence $\bar{x}$ as a database of Y-STR profiles). The *pattern* $\Psi(\bar{x})$ of $\bar{x}$ is the sequence obtained by replacing every symbol of $\bar{x}$ by the order of its first appearance (for example, the sequence $\bar{a}$ contains 7 distinct symbols. b is the last one of these to appear, so we replace b by 7.). In this way, $\Psi(\bar{a}) = \bar{\psi} = 1231451216574$. The length of the pattern $\bar{\psi}$ is denoted by $l(\bar{\psi})$ $(= 13)$.

Obviously, there are several sequences inducing this pattern $\bar{\psi}$, and it makes sense to define the overlying set $X_{\bar{\psi}} = \Psi^{-1}(\bar{\psi}) = \{\bar{x} \in \mathcal{A}^{l(\psi)} : \Psi(\bar{x}) = \bar{\psi}\}$. If $m = m_{\bar{\psi}}$ denotes the highest number appearing in $\bar{\psi}$, the cardinality of $X_{\bar{\psi}}$ is $\frac{(\#\mathcal{A})!}{(\#\mathcal{A}-m)!}$, since for each number in $\bar{\psi}$ we can choose any letter from $\mathcal{A}$, but we cannot choose the same letter twice (for the first number, there are $\#\mathcal{A}$ choices, for the second one $\#\mathcal{A} - 1$, and so on).

In the DNA context, the sequence $\bar{x}$ of haplotypes is a realization of a random variable $\bar{X}$, which is a concatenation of random variables $X_1, X_2, \ldots, X_m$, which are independently and identically distributed according to a certain distribution P. Each sequence thus has an associated probability $P(\bar{X} = \bar{x}) = \prod_{i=1}^{m} P(X_i = x_i)$. The probability $P(\bar{\psi})$ of a particular *pattern* is the sum of the probabilities of all sequences with this pattern. If before drawing the marbles we do not know the possible colours or their number, Orlitsky et al. [2004a] suggest that we should find the distribution $\hat{P}$ which maximizes this pattern probability, called the *high-*

profile distribution.

Since changing the order of the symbols constituting a sequence or permuting the symbols does not alter the pattern probability, we will assume, from now on, that each pattern arises in *canonical form*, i.e. that it looks like $1^{f_1}2^{f_2}\dots m^{f_m}$, satisfying $f_1 \geq f_2 \geq \cdots \geq f_m$. We can abbreviate such a pattern to $(f_1, f_2, \dots, f_m)$, where f_i is understood as indicating the number of occurrences of symbol i. Our sequence $\bar{a}$, for instance, is coded by $(4, 2, 2, 2, 1, 1, 1)$.

Following the same line of thought, we will also denote all discrete probability distributions with finite support as vectors; (p_1, p_2, p_3), for example, will denote a probability distribution over three colours with probabilities p_1, p_2, and p_3. The specific colours associated with these three probabilities are irrelevant, since we are only looking at patterns.

Let's consider the pattern $(2, 1)$. The classical maximum likelihood estimator for the underlying distribution is $(p_1, p_2) = (\frac{2}{3}, \frac{1}{3})$, because this maximizes the probability $p_1^2 p_2$ of the sequence *aab*. The pattern probability of $(2, 1)$, however, equals $P(aab) + P(bba) = \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9}$. If we assume that there are only two colours, with probabilities p and $1 - p$, to maximize this pattern probability we have to maximize $p^2(1-p) + (1-p)^2 p$. The solution is $p = \frac{1}{2}$ (yielding $\hat{P}(2, 1) = \frac{1}{4}$), leading to the conclusion that the observed pattern is most likely under the assumption that the two colours have equal probability of being drawn. This corresponds to our intuition that all probabilities should be presumed equal if there is no evidence to the contrary. Samples of size three can never produce equal numbers of observations for two colours, so in some sense this is unfair towards the hypothesis of equal probabilities. Our method of calculation deals with this unfairness elegantly.

In order to prove that the $(\frac{1}{2}, \frac{1}{2})$-distribution is optimal amongst all probability distributions, including those over more than two colours, suppose that there is a distribution $P = (p_1, p_2, \dots, p_{n-1}, p_n)$ such that $P(2, 1) > \frac{1}{4}$; without loss of generality, assume that $p_1 \geq p_2 \geq \cdots \geq p_n$. Let Q be the distribution over $n - 1$

symbols satisfying $q_i = p_i$ for $i < n-1$ and $q_{n-1} = p_{n-1} + p_n$. Then

$$\begin{aligned}
Q(2,1) &= \sum_i \sum_{j,j\neq i} q_i^2 q_j \\
&= \sum_{i<n-1} \sum_{j<n-1,j\neq i} p_i^2 p_j + (\sum_{i<n-1} p_i^2)(p_{n-1}+p_n) + (p_{n-1}+p_n)^2 \sum_{j<n-1} p_j \\
&= \sum_{i<n-1} \sum_{j<n-1,j\neq i} p_i^2 p_j + (\sum_{i<n-1} p_i^2)(p_{n-1}+p_n) + (p_{n-1}^2+p_n^2) \sum_{j<n-1} p_j \\
&+ 2p_{n-1}p_n \sum_{j<n-1} p_j + p_{n-1}^2 p_n + p_n^2 p_{n-1} - p_{n-1}^2 p_n - p_n^2 p_{n-1} \\
&= P(2,1) + p_{n-1}p_n(2 \sum_{j<n-1} p_j - p_{n-1} - p_n) \\
&= P(2,1) + p_{n-1}p_n(2 - 3(p_{n-1}+p_n)).
\end{aligned}$$

Since p_{n-1} and p_n were the two lowest probabilities of distribution P, we must have $p_{n-1} + p_n \leq \frac{2}{3}$, so $Q(2,1) \geq P(2,1)$. We can continue grouping the two lowest-probability colours together in this way until there are only two colours left, in each step increasing the pattern probability of $(2,1)$, so the optimal distribution must contain two colours. Since we saw that $(\frac{1}{2}, \frac{1}{2})$ is the optimal two-coloured distribution, we conclude that it is optimal overall.

In this example, the number of colours in the high-profile distribution was the same as in the sample. However, this does not hold for every pattern; in general, for fixed pattern size, the greater the number of colours in a pattern is, the more unobserved colours will be assumed by the high-profile distribution. For instance, patterns consisting of all ones, for which every colour is unique in the sample, are most likely for distributions over a large number of colours. The optimal probability distribution would be a uniform distribution over an infinite number of colours, or, if the number of colours is bounded (in our context by the population size), over the maximum number of colours. There is also a category of patterns for which the number of colours in the high-profile distribution is finite, yet strictly greater than the number of colours in the pattern. The smallest example of the latter category is the pattern $(2,1,1)$ of three colours, with high-profile distribution $(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$ [Orlitsky et al. 2004a].

Orlitsky et al. also allow distributions P with a continuous part: if a number from this region is drawn (the probability of this event is called the *continuous size* and is denoted by q_P, while the number of colours k_P with a discrete probability is called the *discrete size*), it is always unique in the sequence. For patterns containing a lot of ones, the high-profile distribution can be one with a continuous part. However, since we are estimating population frequencies of haplotypes, we will limit ourselves to discrete distributions (the algorithm designed by Orlitsky et al. [2004b] does the same). Properties 2 and 3 from the next section support this choice.

4.1 Properties

Before discussing how to compute the high-profile distribution for the European Y-STR data, I would like to list a few properties of this distribution, borrowed from Orlitsky's article [Orlitsky et al. 2004a]. Outlines of proofs of these properties can be found in the same article.
Trivial patterns are () (the empty pattern) and (1), the pattern consisting of just one symbol.

Property 1. For all patterns $\bar{\psi}$, there exists a distribution P (possibly with a continuous part) achieving $\hat{P}(\bar{\psi})$.

This property ensures that the high-profile distribution exists for every pattern. It is not known if this distribution is unique, although for all patterns considered in the article, it is.

Property 2. For all non-trivial patterns, the continuous size $\hat{q}$ satisfies $\hat{q} \leq \frac{n_1}{n}$, where n_1, as in chapter 2, is the number of singletons in the pattern, and n is the length of the pattern.

Property 3. The discrete size $\hat{k}$ of any high-profile distribution is finite.

The upper bound on the continuous size is convenient, since this tells us that if we restrict the parameter space to discrete distributions with bounded size, the result will probably be close to the actual maximum. Any deviation will result in conservative estimates for observed profiles, since the optimal distribution within the space of finite, discrete distributions will always have smaller support than the optimal distribution overall.

For the last property, let $P^{\bar{\alpha}} = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ be a distribution over $\{1, 2, \ldots, m\}$ such that $\sum_{i=1}^{m} \alpha_i = 1$. Consider, for each integer n, the sequence in which symbol i appears $\alpha_i n$ times (assume that all $\alpha_i n$ are integers for simplicity), and let $\phi_{\bar{\alpha}}^n$ denote the corresponding profile. Then the following property holds.

Property 4. As n tends to infinity, $\hat{P}_{\phi_{\bar{\alpha}}^n} \rightarrow P^{\bar{\alpha}}$, in terms of both the K-L divergence $D(P^{\bar{\alpha}}||\hat{P}_{\phi_{\bar{\alpha}}^n})$ and the L_1 distance $||P^{\bar{\alpha}} - \hat{P}_{\phi_{\bar{\alpha}}^n}||_1$.

Since the classical maximum likelihood estimate is equal to $P^{\bar{\alpha}}$, this shows that these two estimators converge as n tends to infinity. And since it is clear from the outline of the proof of this property that, for fixed m, the convergence is uniform for all choices of the α_i [Orlitsky et al. 2004a], it can be shown that the high-profile estimator converges in probability to the underlying probability distribution.

4.2 Computation

Finding the high-profile distribution can be very difficult, especially for large patterns. The space of all probability distributions over k colours is $(k-1)$-dimensional, and on top of that, for each of those distributions, the pattern probability is the sum of $\frac{k!}{(k-m)!}$ sequence probabilities; since calculation time of a sequence probability is roughly linear in the sequence length, it is clear that for larger patterns, it soon becomes impossible to search the entire space of possible probability distributions. Instead, Orlitsky et al. [2004b] propose an *expectation maximization* algorithm to approximate the high-profile distribution.

Expectation maximization, in general, is a useful technique in missing data problems where a certain parameter $\theta \in \Theta$ needs to be estimated. It is an iterative algorithm and thus takes a starting value θ_0 and produces a sequence of approximations $(\theta_0, \theta_1, \theta_2, \dots)$, in each step increasing the likelihood of the observed data. If we denote by $\mathbf{y}$ and $\mathbf{z}$ the observed and missing data, respectively, then an iteration of the algorithm, calculating θ_{t+1} from θ_t, consists of computing, for each $\theta \in \Theta$, the likelihood of the complete data (observed and missing data combined) as a function $f_\theta(\mathbf{z})$ of the missing data $\mathbf{z}$. After that, it calculates the expected value of this function $f_\theta(\mathbf{z})$ under the old distribution for $\mathbf{z}$, with parameter θ_t. The value of θ for which this expected value is maximized is then taken as the next approximation θ_{t+1}. This procedure is summarized by the following formula:

$$\theta_{t+1} = \arg\max_{\theta \in \Theta} E_{\theta_t}[P_\theta(\mathbf{y}, \mathbf{z})]$$

It can be shown that this step does indeed not decrease the likelihood of the observed data $\mathbf{y}$.

To use this algorithm to approximate the high-profile distribution, we view the pattern $\bar{\psi}$ as the observed data and the exact sequence $\bar{X}$ as the missing data. The parameter θ, or rather P in this case, is the underlying probability distribution of the data. Instead of the likelihood of the complete data, we use the expected value of the loglikelihood, since this leads to easier formulas and takes its maximum in the same point. The above formula can now be rewritten as

$$P_{t+1} = \arg\max_{Q \in \mathcal{P}} E_{P_t}[\log Q(\bar{X})],$$

since $Q(\bar{X}, \bar{\psi}) = Q(\bar{X})$.

If $\mathcal{P}$ is the space of probability distributions over k colours, and $f_i(\bar{x})$ is the number of times symbol i appears in sequence $\bar{x}$, the expectation can be expanded to

$$E_{P_t}[\log Q(\bar{X})] = \sum_{i=1}^{k} c_i \log Q(i),$$

where $c_i = \sum_{\bar{x} \in X_{\bar{\psi}}} f_i(\bar{x}) P_t(\bar{X} = \bar{x})$. This equation takes the same form as the loglikelihood of a multinomial sample, so the expectation is maximized if the $Q(i)$ are equal to the coefficients c_i, scaled so that they sum to 1. Each iteration has now been reduced to calculating the values of the c_i, but since this involves a summation over all sequences with the prescribed pattern, it can still be exponential in the length of the pattern. We therefore follow the advice of Orlitsky et al. and use the Metropolis algorithm to calculate these values.

This algorithm, originally designed for calculating properties of large systems of interacting molecules, can also be used to compute expected values for discrete random variables with large supports. For such a random variable X, it views the elements of its support as states in a Markov chain, and defines the transition probabilities in such a way that the stationary distribution of the chain equals the original distribution of X. The expected value of any function f of X can then be approximated by simulating a random walk in the system and averaging the value of $f(a)$ over all states a visited in the walk. The transition probabilities from state A are defined in the following way:

First, a candidate B for the next state is selected according to a certain distribution P_A. Next, a number u is drawn from a $U(0,1)$-distribution, and the transition from A to B is accepted if $u < g(A,B) := \frac{P(X=B) \cdot P_B(A)}{P(X=A) \cdot P_A(B)}$; if not, the transition is denied and the next state is again A. If the distributions P_A are chosen in such a way that the Markov chain is irreducible and aperiodic, it converges to a unique stationary distribution. This distribution must be equal to P, the distribution of X, since P satisfies the conditions for stationarity for each state A:

$$
\begin{aligned}
&\sum_i P(X=i)\cdot p_{iA} \\
&= \sum_i P(X=i)\cdot P_i(A)\cdot \min\{\frac{P(X=A)\cdot P_A(i)}{P(X=i)\cdot P_i(A)},1\} \\
&+ P(X=A)\sum_i P_A(i)\cdot (1-\min\{\frac{P(X=i)\cdot P_i(A)}{P(X=A)\cdot P_A(i)},1\}) \\
&= \sum_i \Big(P(X=i)\cdot P_i(A)\cdot \min\{\frac{P(X=A)\cdot P_A(i)}{P(X=i)\cdot P_i(A)},1\} \\
&+ P(X=A)\cdot P_A(i)\cdot (1-\min\{\frac{P(X=i)\cdot P_i(A)}{P(X=A)\cdot P_A(i)},1\})\Big) \\
&= \sum_{i,g(i,A)<1} P(X=A)\cdot P_A(i) \\
&+ \sum_{i,g(i,A)\geq 1} \Big(P(X=i)\cdot P_i(A) + P(X=A)\cdot P_A(i)\cdot (1-\frac{P(X=i)\cdot P_i(A)}{P(X=A)\cdot P_A(i)}\Big) \\
&= \sum_i P(X=A)\cdot P_A(i) \\
&= P(X=A)\sum_i P_A(i) \\
&= P(X=A)
\end{aligned}
$$

As we saw on the previous page, calculation of the coefficients c_i, which correspond to the solution of an iteration in our expectation-maximization algorithm, amounts to averaging the frequency of symbol i over all sequences in $X_{\bar{\psi}}$, weighted by the probabilities of drawing those sequences in a random sample of length n from a distribution P_n.

As the states of the Metropolis algorithm, we choose the elements of $X_{\bar{\psi}}$, represented by vectors $(a_1, a_2, \ldots, a_m)$, where a_i is the symbol assigned to the ith component of the canonical representation of $\bar{\psi}$, thus occurring n_i times in that particular sequence.

$f : X_{\bar{\psi}} \rightarrow \{1, \ldots, n\}^{\#\mathcal{A}}$ is the function that maps each sequence to its vector of symbol counts, and for each $\bar{x} \in X_{\bar{\psi}}$, a candidate for transition is selected by drawing two numbers j, k from $\{1, 2, \ldots, n\}$ without replacement. If both j and k occur in $(a_1, a_2, \ldots, a_m)$, the candidate sequence is obtained by swapping these two symbols; if only one of them does, it is replaced by the other one; and finally, if neither of them occurs in the current sequence, the transition candidate is equal to the current sequence and automatically accepted.

Now that the states and transition probabilities of the Markov chain have been defined, we have to decide how long the random walk should be allowed to

continue. On the one hand, we want to be sure that there are enough steps for the walk to reach the stationary distribution and for the variance in the average value of f to be small, but on the other hand, we would like to keep computation time low. Another issue, related to this one, concerns the number of steps the expectation-maximization algorithm needs to converge to a stable distribution.
We ran some tests to determine satisfactory values for these parameters and incorporated them into the R-program highprofile.r (see Appendix), which takes a sequence in canonical form as its input and calculates the corresponding high-profile distribution, using the expectation-maximization algorithm along with the Metropolis approximation of each iteration. This program is freely available from the author.

One peculiar aspect of the high-profile estimator should be mentioned: by allowing all possible pairings of haplotypes to components of the probability distribution, it does not estimate population frequencies belonging to specific haplotypes; the only thing it estimates is the entire vector of population frequencies. If we do want such an estimate, as in forensic cases, it is not clear how we should do this.
One option is to assign the highest estimated frequency to the most common haplotype, the second highest estimated frequency to the second most common haplotype, and so on. Alternatively, we could use the conditional posterior distribution over all pairings induced by the high-profile distribution, and compute the expected value of each haplotype's frequency under this conditional distribution.
Further research is needed to resolve this issue.

Chapter 5

Comparison

To compare our new approach to the other methods, let's see what the estimated frequencies are for the rarest Y-STR profiles (with database frequencies f of at most 5), for a more common profile with frequency 30, and for a profile occurring 300 times. These numbers are the frequencies of the profiles in the extended database of $n = 12728$ (see chapter 2).

Since the haplotype surveying method requires both a frequency and a weighted inverse distance for its calculation, we used prior distributions corresponding to two W-values, $W = 0.02$ and $W = 0.14$, that appear to be extreme from the haplotype surveyors' graph [Roewer et al. 2000]. We incorporated the corrections Michael Krawczak suggested in his comment on this method [Krawczak 2001] and chose to represent the posterior distributions by their mean values.

The estimate pertaining to the high-profile method is calculated as follows: first, the high-profile probability vector belonging to the extended database is calculated using the expectation maximization algorithm (three runs of this algorithm with different starting values showed nearly equal results). Then this vector is ordered decreasingly, and linked to the decreasing vector of database frequencies. If the frequency of the profile of interest is unique in the database, the corresponding component of the ordered high-profile vector is used as a frequency estimate. If not, we use the mean of those components of the high-profile vector that correspond to the profiles having the same frequency as the profile of interest.

Table 5.1 shows the resulting estimates.

The table shows that for common profiles, the estimates of the haplotype surveying and high-profile methods are close to the database frequency (this does not hold for the Brenner estimate, since the ratio between this estimate and the database frequency is a constant factor $1 - \frac{1397}{12728} \approx 0.890$). This similarity between the three estimates is a good thing, since the database frequency is the maximum likelihood estimator for the population frequency, and for this sample size it is a very reliable one (a database frequency of 300 gives rise to

Database frequency	Counting method	Brenner	Haplotype surveying		high-profile method
			$W = 0.02$	$W = 0.14$	
1	$7.86 \cdot 10^{-5}$	$6.99 \cdot 10^{-5}$	$8.21 \cdot 10^{-5}$	$9.96 \cdot 10^{-5}$	$5.46 \cdot 10^{-5}$
2	$1.57 \cdot 10^{-4}$	$1.40 \cdot 10^{-4}$	$1.56 \cdot 10^{-4}$	$1.78 \cdot 10^{-4}$	$9.85 \cdot 10^{-5}$
3	$2.36 \cdot 10^{-4}$	$2.10 \cdot 10^{-4}$	$2.30 \cdot 10^{-4}$	$2.56 \cdot 10^{-4}$	$2.15 \cdot 10^{-4}$
4	$3.14 \cdot 10^{-4}$	$2.80 \cdot 10^{-4}$	$3.04 \cdot 10^{-4}$	$3.34 \cdot 10^{-4}$	$3.02 \cdot 10^{-4}$
5	$3.93 \cdot 10^{-4}$	$3.50 \cdot 10^{-4}$	$3.78 \cdot 10^{-4}$	$4.12 \cdot 10^{-4}$	$3.64 \cdot 10^{-4}$
30	$2.36 \cdot 10^{-3}$	$2.10 \cdot 10^{-3}$	$2.23 \cdot 10^{-3}$	$2.37 \cdot 10^{-3}$	$2.28 \cdot 10^{-3}$
300	$2.36 \cdot 10^{-2}$	$2.10 \cdot 10^{-2}$	$2.22 \cdot 10^{-2}$	$2.35 \cdot 10^{-2}$	$2.34 \cdot 10^{-2}$

Table 5.1: Comparison of frequency estimates

a 95%-confidence interval of $[0.0211, 0.0265]$). It was to be expected, too: in chapter 3, we already saw that the prior distributions of the haplotype surveying method only have a big impact on the posterior distribution for rare haplotypes. As for the high-profile method, the pattern likelihood, which is a sum of a large number of sequence likelihoods, is dominated by those sequences in which the most frequent haplotypes appear in the right order, the most frequent symbols thus representing the haplotypes with the highest probabilities. These dominant sequence probabilities are maximized by the classical maximum likelihood estimator for the entire parameter space, so the high-profile estimate should be close to the database frequency.

For the rarest profiles, absent from the original database and thus occurring once in the extended database, the situation is very different: the highest frequency estimate, produced by the haplotype surveying method for $W = 0.14$, is almost twice as large as the high-profile estimate, while the counting method estimate lies in the middle.

Such a high W-value for an absent haplotype is not unthinkable; for instance, haplotype $14 - 14 - 29 - 25 - 11 - 13 - 13$, which only differs two repeats from the most common haplotype and has a weighted inverse distance of $W = 0.21$, does not occur in the database. Of course, this could mean that it is absent from the entire population and thus will never appear in forensic casework - this will be clear when data from cases become available.

For haplotypes with frequency 2, there is only a big difference between the estimate for the high-profile method on the one hand and the counting and haplotype surveying methods on the other hand. For haplotypes with three or more copies in the database, the deviations from the counting method estimate provided by any of the other methods are never more than 11%. Such small differences are thought to be irrelevant in most lawsuits.

5.1 Other countries

Before we decide which method to use, it may be insightful to have a look at the policies of the forensic laboratories in other European countries. To this end, we sent a questionnaire to the members of ENFSI, the European Network of Forensic Science Institutes, to find out how they report matches of Y-STR haplotypes, and which method they use to estimate the relevant population frequency. Only 10 out of 43 laboratories responded, but these responses already revealed that opinions on this relatively new technique vary widely. Four laboratories prefer to use the counting method (or report the number of findings in the database without converting this to a probability), three use the estimates from the YHRD website, and one lab (from Finland) only uses Y-STR typing for exclusion, because it only recently started using the technique. Two laboratories do not use Y-STR typing yet.

Most of the laboratories use the YHRD database, or a relevant subset of it, for their calculations. The British FSS and the Finns are the only ones who have a separate national reference database. Even in France, the YHRD numbers are reported, although this database contains almost no French data; the reason for this is that the European population is thought to be sufficiently homogeneous, and due to the large numbers of foreign inhabitants of and visitors to France, suspects can be of any population.

Together with an account of the frequency of a suspect's profile, some forensic experts remind the judge or jury that paternal relatives of the suspect usually have the same Y-STR profile. The German Bundeskriminalamt reports a local frequency, in addition to a list of other populations where the haplotype has been detected.

None of the respondents use their databases of Y haplotypes to search for possible suspects.

Chapter 6

Conclusion and discussion

In the previous chapters, I have presented and analyzed four methods for calculating a match probability for Y-STR haplotypes, through an estimation of the frequency of this haplotype in the European population. I will summarize the important aspects of these methods before advising NFI on which method to use in their reports.

The counting method has the advantage of providing the classical maximum likelihood estimate for the frequency of each individual haplotype. Although its estimates become unreliable for rare haplotypes, it is easy to explain in court and it is unbiased, if no information other than the frequency of the haplotype of interest in the database is taken into consideration. This leads to the key question: since so much more information is available, such as the genetic structure of the haplotype pool, the significant number of unique haplotypes in the database, and the theoretical background provided by population genetics, can we use this information to obtain a better estimate? All of the other methods try to use part of this extra information.

The haplotype surveying method, to begin with, tries to incorporate the genetic information into its estimate. Its main hypothesis is that Y-STR profiles that are genetically close to frequent profiles can be expected to have higher frequencies themselves. Indeed, the data show a relation between a profile's weighted inverse distance and its database frequency, but this relation is very weak: for each of the fifteen groups of haplotypes having roughly the same W-value, the standard deviation of the frequencies is much higher than the mean. Consequently, the prior distribution for a profile's frequency that we obtain from this regression analysis is very dispersed, and the likelihood of the data largely determines the resulting posterior distribution, so that the mean of that distribution is always close to the counting method estimate. Therefore, also considering the small errors that have been made in the assumptions (the beta distribution, the definition of W) and the calculations (the bias in $\sigma(W)$), I would not recommend using this method in its present form, despite its popularity in the international forensic community.

Charles Brenner suggests the use of another piece of information, namely the existence of a large number of haplotypes that have yet to be detected, indicated by the numerous unique haplotypes (singletons) in the database. According to this estimate, the observed profiles constitute 89% of the actual population, and we are bound to overestimate frequencies on average if we ignore this fact. The problem, however, is that we don't know which profiles are going to be overestimated. The singletons are the most likely candidates.

To see why, consider all profiles in the population with a frequency below $\frac{1}{n}$, where n is the database size. None of these is expected to be drawn (i.e. the expected frequencies are smaller than 1), but if there are many of them, some will be drawn by chance. The frequencies of all of these lucky ones will be overestimated, and the majority of them will turn up as singletons. But since in each forensic case we are only interested in estimating *one* profile's population frequency, and since we cannot determine whether this profile was one of the rare ones and was accidentally drawn, or rather was one that was frequent enough to be expected to be drawn, it is not clear how these considerations should affect our estimate. We would need to know how many of the rare profiles there are, and how rare they are exactly.

This is where the high-profile estimator comes in. This estimator attempts to reconstruct the entire set of population frequencies by picking that set that would be most likely to give rise to *all* observed database frequencies simultaneously, instead of each of the frequencies separately, like the counting method does. In the process, the high-profile estimator creates some extra, unobserved haplotypes and assigns probabilities to them, thus coming closer to the structure of the real population.

But the merit of the high-profile method is also its handicap. By considering all pairings of haplotypes to database frequencies, it may give a better estimate of the entire vector of population frequencies, but loses the connection between individual haplotypes and their frequencies. I have tried to resolve this by taking the most probable pairing, the one in which the highest probabilities are paired with the most frequent haplotypes; the estimate for a singleton then comes down to the mean of all frequencies that are linked to a singleton in this pairing. But there are other options: one could, for instance, take the highest frequency assigned to a singleton instead of the mean frequency, to be conservative; or one could consider all possible pairings, and then take as an estimate the average frequency a singleton has in all of these pairings, weighting each pairing by its corresponding sequence probability. This calls for more theoretical backing-up.

In my opinion, it is safe to await this research for the moment and not use the high-profile estimator. Although I do believe that this is the way to go, one should, when in doubt, choose a conservative estimate, and this means using the counting method. After all, for profiles occurring more than three times in the database, there is little difference between the estimates provided by the three most important estimators, and for the singletons and doubletons, there is such a high unreliability in all methods that one would probably rather refrain from stating a hard number as an estimated frequency altogether. Therefore, the best way to report a frequency to the court in my view is to give the number of copies observed in the relevant database, and the size of this database, and to divide them as an illustration. Furthermore, it is instructive to provide these numbers for several databases, to illustrate the spatial variation.

Now there is one very important issue that we have not yet paid attention to. Since we have used the European database for our calculations, we have calculated a match probability for an individual chosen at random from the European population (assuming that the database is representative for this population). However, in forensic cases, the population we are interested in, the suspect population, depends on the circumstances and is usually smaller than the European one, sometimes as small as a village. The village population cannot simply be regarded as a sample from the European population, since the inhabitants of the village are probably somewhat more related and their Y-haplotypes are therefore not independent. This is why some experts stress the fact that paternal relatives have the same haplotype: those paternal relatives often live in the suspect's neighbourhood and thus may belong to the suspect population, increasing the expected frequency of his haplotype in that population. The more generations the suspect's family has been living in the same area, the more paternal relatives there will be and the more serious this increase will be.
This bears consequences on the evidential weight of the matching profiles, depending on the situation. If the matching suspect is not related to the other members of the suspect population, for instance if he only recently migrated into the neighbourhood, or if the suspect population consists of the visitors of a bar in a big city, the suspect and the culprit (under the assumption that they are not the same person) can be viewed as two independent samples from the Dutch population. Analysis of molecular variance by Roewer et al. [2000] has shown that population stratification across western and central Europe is small enough for the European database to be also representative for the Dutch population. Indeed, the Dutch samples that we have do not exhibit large deviations from the estimated European frequencies, at least not for those haplotypes for which both the Dutch and European databases warrant reliable estimates.

6.1 Special cases

In most forensic cases where Y-STR typing is used, the DNA evidence will be a comparison between a suspect's complete Y haplotype and a complete Y haplotype left at the crime scene. If the two profiles don't match, the probability that the suspect is the source of the crime scene DNA, typing errors aside, is zero, and in this paper I have discussed what to do if they do match. However, there are cases in which the evidence is more complicated. I would like to discuss three of these special cases.

The first one is a case where only an incomplete haplotype can be obtained from the crime scene sample. There may be a number of explanations for this: either there is not enough DNA material to obtain a full profile, or the material has been damaged due to extreme circumstances, or it is very old. The sample can also be polluted, so that some alleles are no longer clearly visible. In such a case, the original full haplotype can be any haplotype that has the correct number of repeats at the loci that could be typed unambiguously. The various methods allow different ways of dealing with this kind of evidence.

First, the counting method would count all possible haplotypes in the database that reduce to the observed incomplete haplotype, and divide this number by the database size to obtain a frequency estimate. A similar approach would be best for the haplotype surveying method: the simplest solution is to sum the posterior means of all possible extensions of the incomplete haplotype. One could also rebuild the entire method using only the loci that could be typed in this particular case, but this would involve a lot of extra work.

The haplotype surveying website [Willuweit & Roewer 2007] does provide frequency estimates for incomplete haplotypes, but I would not recommend using them, since sometimes incomplete haplotype frequencies are estimated lower than some of their extensions to complete haplotypes (for instance, the estimate for $13-11-29-24-?-13-13$ is $2.50 \cdot 10^{-4}$, while the estimate for $13-11-29-24-11-13-13$ is $3.30 \cdot 10^{-4}$). Some incomplete haplotypes even yield negative estimates. Thus, the website calculations should only be used for complete (nine-locus!) haplotypes.

The other two methods require a different course of action, since they both incorporate unobserved haplotypes in their calculations. Because the crime scene haplotype could be one of these extra haplotypes, ignoring them would result in an underestimation of the match probability. This can be resolved by reducing all haplotypes to those loci that could be typed for the crime scene haplotype, effectively grouping together some of them, and then applying the Brenner and high-profile methods to these modified data. The resulting estimates will be higher than the sum of the original estimates for all possible complete haplotypes, since the pattern of haplotype frequencies will contain higher numbers and thus fewer extra profiles will be needed, leaving higher probabilities for the observed profiles.

The second case that I would like to consider is one with mixed profiles. If a sample contains DNA from two different donors, each locus will exhibit two alleles, unless the two contributors match at that locus. Suppose that a suspect's haplotype matches one of the crime scene alleles at each of the loci, then he cannot be excluded, but how should this evidence be evaluated? Fortunately, a mixture of Y-STR profiles is easier to evaluate than a mixture of autosomal profiles (see

Buckleton [2005]).

Denote by $\pi_1 \oplus \pi_2$ the mixture of profiles π_1 and π_2, e.g. $(14 - 14 - 30 - 24 - 11 - 14 - 14) \oplus (14 - 13 - 29 - 25 - 10 - 13 - 13) = (14, \{13, 14\}, \{29, 30\}, \{24, 25\}, \{10, 11\}, \{13, 14\}, \{13, 14\})$. Let A be the mixture found at the crime scene, s the suspect and π his profile, τ the unique profile such that $\pi \oplus \tau = A$, $\mathcal{D}$ the set of all possible profiles, and C_1 and C_2 the two contributors to the sample (assume, for simplicity, that they are unrelated). Then the relevant likelihood ratio can be expressed in the following way, along the lines of the Bayesian theory of chapter 2:

$$\frac{P(C_1 \oplus C_2 = A | s \notin \{C_1, C_2\})}{P(C_1 \oplus C_2 = A | s \in \{C_1, C_2\})} \tag{6.1}$$

$$= \frac{P(C_1 \oplus C_2 = A | s \notin \{C_1, C_2\}}{P(C_1 \oplus C_2 = A | C_1 = s)} \tag{6.2}$$

$$= \frac{\sum_{\substack{H_1, H_2 \in \mathcal{D} \\ H_1 \oplus H_2 = A}} P(C_1 \equiv H_1, C_2 \equiv H_2 | s \notin \{C_1, C_2\})}{P(C_2 \equiv \tau | C_1 = s)} \tag{6.3}$$

For the probability in the numerator, one needs to estimate the joint probabilities of all combinations of two haplotypes that constitute the observed mixed profile, which is a sum over 2^p combinations, where p is the number of loci that exhibit two different alleles in the mixture. For the probability in the denominator, one needs to estimate one frequency, namely the frequency of the unique profile τ that together with the suspect's profile forms the mixed profile. This reveals an interesting aspect of this type of evidence: if the mixed profile can be decomposed in several ways into two frequent haplotypes, while the haplotype that complements the suspect's is a very rare one, the above likelihood ratio can become bigger than one, in which case the evidence provided by the mixed profile is actually *in favour* of the suspect, despite his profile fitting into it. This phenomenon is caused by the fact that allele probabilities at multiple Y-STR loci are dependent. In autosomal evidence, it can only occur if allele frequencies are higher than $\frac{1}{2}$.

The third and final special case is the one in which the DNA evidence consists of both a Y-STR profile and an autosomal STR profile. Usually, the autosomal profile will be incomplete, since otherwise there would be little need for extra DNA evidence, the random match probability already being small enough. According to the results of the ENFSI questionnaire, the Slovenian laboratory multiplies the match probabilities associated with the two pieces of evidence, while the other laboratories that have encountered such cases are not as bold and report the two figures separately. The important issue here is whether the two profiles can be regarded as independent pieces of information or not. Adversaries of this assumption point out that people who share the same Y-STR profile usually have similar genetic backgrounds, and are thus more likely to have similar autosomal profiles, because their profiles have been sampled from the same genetic pool, in

which frequencies of autosomal alleles can be different from those in the larger population. One could correct for this effect by incorporating a subpopulation coefficient θ, as is common practice in autosomal frequency estimation [Balding 2005].
A recent study by Walsh et al. [2007] shows that autosomal and Y-STR alleles are sufficiently independent to warrant multiplication of the corresponding match probabilities. This study was conducted in the United States, so it would be good to do a similar study in Europe. Considering the history of migration to the United States from all continents, however, we would expect the American population to be more stratified than the European one and thus to be more prone to linkage disequilibrium between autosomal and Y-chromosomal loci.

6.2 Suggestions for future research

The most important subject of future research, in my opinion, should be the effect of small, geographically defined suspect populations on the frequencies of Y-STR haplotypes. There are two obvious directions in which this research could develop: first, one could use a theoretical population-genetic approach and analyze mutation-migration models for populations of various sizes, with various rates of migration from the larger population, possibly adapted to fit real case situations. This could provide information on how likely it is that a rare haplotype reaches a high frequency in a relatively isolated population. Another approach is an empirical one: by gathering data about small populations, one could also gain insight into these mechanisms. Peter de Knijff is currently working on this in the Netherlands, along with the construction of a larger Dutch database.
The estimation of European haplotype frequencies has been discussed extensively, but a number of issues have yet to be resolved before either of the more sophisticated methods can be used confidently. For the haplotype surveying method, the main hypothesis of W-dependent beta distributions for haplotype frequencies should be validated more properly. To this end, one could use a goodness-of-fit test (like the Kolmogorov-Smirnov test) to determine whether frequencies of haplotypes with the same W-value are indeed beta distributed, and then compare a W-dependent fit of such a distribution to the data to a non-W-dependent fit, to see if W is indeed a predictor of the frequency.
Lastly, I would recommend further research into the properties of the high-profile estimator, since it has some nice advantages over the classical maximum likelihood estimator, but important aspects like convergence behaviour or uniqueness have not been investigated satisfactorily. Nor is it clear how this estimator can be used to estimate the frequency of a particular profile.

Acknowledgments

A number of people have helped me to write this thesis. First and foremost I would like to thank my two thesis advisors, Erik van Zwet and Marjan Sjerps, for keeping me motivated during the past eight months, for supporting me every time I had some crazy idea, and for supplying me with all the articles and books I needed. I would also like to thank the other people at the Netherlands Forensic Institute that helped me one way or another, most notably Ate Kloosterman and Annabel Bolck.

The following persons were very kind to answer my questions about their publications and to give their opinions on my thoughts, for which I would like to thank them: Peter de Knijff, Manfred Kayser, Hans Metz, Charles Brenner, Michael Krawczak, Lutz Roewer, and Sascha Willuweit. They helped me to understand the various methods I described and to guide me in developing my own method. Finally, I would like to thank all members of ENFSI who responded to our questionnaire.

References

1. D.J. Balding, Weight-of-evidence for Forensic DNA profiles, Wiley (2005).

2. J. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press (2005).

3. B. Budowle, M. Adamowicz, X. G. Aranda, C. Barna, R. Chakraborty, D. Cheswick, B. Dafoe, A. Eisenberg, R. Frappier, A. M. Gross, Twelve short tandem repeat loci Y chromosome haplotypes: Genetic analysis on populations residing in North America, Forensic Sci. Int. 150 (2005) 1-15.

4. J.M. Butler, Forensic DNA typing, 2nd ed., Elsevier, Burlington USA (2005).

5. I.W. Evett, B.S. Weir, Interpreting DNA evidence: statistical genetics for forensic scientists, Sunderland, MA: Sinauer Associates (1998).

6. R. Fu, A.E. Gelfand, K.E. Holsinger, Exact moment calculations for genetic models with migration, mutation, and drift, Theor. Pop. Biol. 63 (2003) 231-243.

7. I.J. Good, The population frequencies of species and the estimation of population parameters, Biometrika Vol. 40 (1953), 237-264.

8. R.C. Griffiths, Allele frequencies in multidimensional Wright-Fisher models with a general symmetric mutation structure, Theor. Pop. Biol. 17 (1980) 51-70.

9. L. Gusmão, J.M. Butler, A. Carracedo, P. Gill, M. Kayser, W.R. Mayr, N. Morling, M. Prinz, L. Roewer, C. Tyler-Smith, P.M. Schneider, DNA Commisssion of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis, Forensic Sci. Int. 157 (2006) 187-197.

10. M.A. Jobling, C. Tyler-Smith, The human Y chromosome: an evolutionary marker comes of age, Nature Rev. Genet. 4 (2003), 598-612

11. M. Kimura, T. Ohta, Stepwise mutation model and distribution of allelic frequencies in a finite population, Proc. Natl. Acad. Sci. USA Vol. 75, No. 6 (1978) 2868-2872

12. M. Krawczak, Forensic evaluation of Y-STR haplotype matches: a comment, Forensic Sci. Int. 118 (2001) 114-115.

13. A. Orlitsky, N.P. Santhanam, K. Viswanathan, J. Zhang, On modeling profiles instead of values, ACM Int. Conference Proceeding Series Vol. 70 (2004a) 426-435.

14. A. Orlitsky, S. Sajama, N.P. Santhanam, K. Viswanathan, J. Zhang, Algorithms for modeling distributions over large alphabets, Int. Symposium on Information Theory, Proceedings (2004b) 304.

15. L. Pereira, M.J. Prata, A. Amorim, Mismatch distribution analysis of Y-STR haplotypes as a tool for the evaluation of identity-by-state proportions and significance of matches – the European picture, Forensic Sci. Int. 130 (2002) 147-155.

16. H.E. Robbins, Estimating the total probability of the unobserved outcomes of an experiment, Annals of Math. Statistics Vol. 39 (1968) 256-257.

17. L. Roewer, M. Kayser, P. de Knijff, K. Anslinger, A. Betz, A. Caglià, D. Corach, S. Füredi, L. Henke, M. Hidding, H.J. Kärgel, R. Lessig, M. Nagy, V.L. Pascali, W. Parson, B. Rolf, C. Schmitt, R. Szibor, J. Teifel-Greding, M. Krawczak, A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males, Forensic Sci. Int. 114 (2000) 31-43.

18. B. Walsh, A.J. Redd, M.F. Hammer, Joint match probabilities for Y chromosomal and autosomal markers, Forensic Science International Apr 19 2007.

19. S. Willuweit, L. Roewer, on behalf of the International Forensic Y Chromosome User Group, Y chromosome haplotype reference database (YHRD): Update, Forensic Science International: Genetics 1(2) (2007) 83-87.

20. S.G. Wright, The distribution of gene frequencies in populations, Proc. Natl. Acad. Sci. 23 (1937) 307-320.

Appendix

highprofile.r

```
# input: a vector pattern, containing the database counts of all
#    observed haplotypes, ordered from most frequent to rarest.
# output: a vector P, representing the high-profile distribution

n = sum(pattern) # length of the sequence
l=length(pattern) # number of different colours in the sequence
n1=sum(pattern==1) # number of singletons
colours=2*l # total number of different colours that we allow
pattern=c(pattern, rep(0,l)) # extend the pattern, for easy programming
P = seq(colours,1,by=-1)*2/(colours^2-colours) # initial prob. vector
runs = 2000000 # number of iterations in the Metropolis algorithm
runs2 = 20 # number of iterations in the EM algorithm
samp = 100 # to reduce runtime

run2 = 0
while(run2<runs2){
y = 1:length(P)
run = 0
Pnext = rep(0, length(P))
while (run < runs){
a = sample(1:l,1)
b = sample((1:length(P))[-a],1)
u = rexp(1,1)
if((pattern[a]-pattern[b])*(log(P[y[a]])-log(P[y[b]])) < u){
temp = y[a]
y[a] = y[b]
y[b] = temp
}
run = run + 1
if(run %% samp == 0){
for(i in 1:l)
Pnext[y[i]] = Pnext[y[i]] + pattern[i]
}
}
P = (Pnext[Pnext>0])*samp/(runs*n)
P = P[order(P, decreasing=TRUE)]
run2 = run2 + 1
cat(P[1:20],'\n')
}
```