Aalborg University

Department of Mathematical Sciences

PhD Thesis

Statistical Aspects of Forensic Genetics
Models for Qualitative and Quantitative STR Data

August 2010

Torben Tvedebrink

Department of Mathematical Sciences, Aalborg University,
Fredrik Bajers Vej 7 G, 9220 Aalborg East, Denmark

Department of Mathematical Sciences
Fredrik Bajers Vej 7G
DK-9220 Aalborg East
Denmark
Telephone: +45 99 40 88 00
Fax: +45 98 15 81 29
Web: http://www.math.aau.dk

Title:

Statistical Aspects of Forensic Genetics
Models for Qualitative and Quantitative STR Data

Author:
Torben Tvedebrink

Supervisor:
Poul Svante Eriksen
Department of Mathematical Sciences
Aalborg University

PhD Assessment Committee:

Prof. Thore Egeland, Dr. Scient
Institute of Forensic Medicine
University of Oslo

Prof. Dr. Peter M. Schneider
Institute of Legal Medicine
University of Cologne

Prof. Rasmus Waagepetersen, PhD
Department of Mathematical Sciences
Aalborg University

Thesis – number of pages: 183

Submitted: August 2010

Preface

Summary in English

This PhD thesis deals with statistical models intended for forensic genetics, which is the part of forensic medicine concerned with analysis of DNA evidence from criminal cases together with calculation of alleged paternity and affinity in family reunification cases. The main focus of the thesis is on crime cases as these differ from the other types of cases since the biological material often is used for person identification contrary to affinity.

Common to all cases, however, is that the DNA is used as evidence in order to assess the probability of observing the biological material given different hypotheses. Most countries use commercially manufactured DNA kits for typing a person's DNA profile. Using these kits the DNA profile is constituted by the state of 10-15 DNA loci which has a large variation from person to person in the population. Thus, only a small fraction of the genome is typed, but due to the large variability, it is possible to identify individuals with very high probability. These probabilities are used when calculating the weight of evidence, which in some cases corresponds to the likelihood of observing a given suspect's DNA profile in the population.

By assessing the probability of the DNA evidence under competing hypotheses the biological evidence may be used in the court's deliberation and trial on equal footing with other evidence and expert statements. These probabilities are based on population genetic models whose assumptions must be validated. The thesis's first two articles describe the "θ-correction" which compensate for possible population structures and remote coancestry that could affect the models' accuracy. The Danish reference database with nearly 52,000 DNA profiles, is analysed and the number of near-matches is compared to the expected numbers under the model.

A frequent event in connection with crime cases is the detection of more than one person's DNA in a sample from the crime scene. In such cases, the DNA profile is called a DNA mixture as it is not possible mechanically or chemically to separate the biological traces into its contributing parts. To ascribe an evidentiary weight to a DNA mixture, the quantitative part (comprised as signal intensities in a so-called electropherogram - EPG) of the result from biotechnological analysis is used. Two models for handling DNA mixtures are presented together with an efficient algorithm to separate the DNA mixture in the most probable contributing profiles. Furthermore, it is discussed how the quantitative part of the evidence is included in calculating the evidential weight.

In criminal cases, the biological traces are often found at crime scenes in conditions which can degrade and contaminate the DNA strand, which complicates the subsequent biochemical analysis. Furthermore, the amount of DNA may be limited which may challenge the sensitivity of the biotechnology applied in the analysis. Models to evaluate the degree of degradation and estimate the probability of an allelic drop-out are discussed in the thesis. Furthermore, it is exemplified how to incorporate the probability of degradation and drop-out when calculating the weight of evidence.

Finally, the thesis contains an article which deals with post-processing of the data after the signal is processed by PCR thermo cycler and detected by electrophoresis apparatus. Central is the detection of a signal-to-noise limit which currently is a fixed limit recommended by the manufacturer of the typing kit. This article discusses how this threshold can be determined from the noise such that it may be specific to each case and locus. Additionally two filters are presented that handle specific types of artifacts in the data generation process which are manifested as increased signals in the EPG.

Summary in Danish

Denne ph.d-afhandling omhandler statistiske modeller med anvendelse indenfor retsgenetik, som er den del af det retsmedicinske område som beskæftiger sig med analyser af dna-spor fra kriminalsager, samt beregning af påstået slægtskab i forbindelse faderskabs- og familiesamføringssager anvendt i retlig sammenhæng. Afhandlingen har et særligt fokus på kriminalsager, idet disse adskiller sig fra de øvrige sagstyper ved at det biologiske materiale ofte anvendes til personidentifikation i modsætning til beslægtethed.

Fælles for sagerne er dog, at dna bruges som bevis i forhold til at sandsynliggøre forskellige hypoteser fremsat i den respektive sag. I langt de fleste lande anvendes kommercielle dna-kit til at typebestemme en persons dna-profil. Disse kit fastlægger dna-profilen ud fra 10 til 15 dna-markører, som har en stor variation fra person til person i befolkningen. Således er det kun en brøkdel af genomet som typebestemmes, men grundet den store variabilitet er det muligt ud fra disse få markører at identificerer personer med meget høj sandsynlighed. Disse sandsynligheder anvendes til at udregne den bevismæssigevægt, som eksempelvis beskriver sandsynligheden for at observerer en given mistænkts dna-profil i befolkningen.

Ved at vurdere sandsynligheden for dna-beviset under konkurrerende hypoteser kan det biologiske bevis inddrages i rettens votering og domsafsigelse, på lige fod med øvrige beviser og ekspertudsagn. Disse sandsynligheder bygger på populationsgenetiske modeller, hvis antagelser må godtgøres. I afhandlingens to første artikler beskrives den såkaldte "θ-korrektion" som kompenserer for mulige befolkningsstrukturer og fjernt slægtskab, som kan indvirke på modellernes korrekthed. Blandt andet analyseres den danske referencedatabase med knapt 52.000 dna-profiler, hvor det undersøges, hvor meget disse dna-profiler adskiller sig fra hinanden, samt om antallet af nærmatches kan forklares ved hjælp af de anvendte modeller.

En ofte forekommende hændelse i forbindelse med kriminalsager er detektion af mere end én persons dna i en prøve fra et gerningssted. I sådanne tilfælde kaldes gerningsstedsprofilen en dna-mikstur, idet det ikke er muligt rent mekanisk eller kemisk at separere det biologiske spor i de bidragende dna-profiler. For at kunne tilskrive en bevismæssig vægt til en dna-mikstur, bruges den kvantitative del (bestående af signalintensiteten udtryk i et såkaldt elektroferogram - EPG) af resultatet fra de bioteknologiske analyser af dna-sporet. Der præsenteres to modeller til håndtering af dna-miksturer og en effektiv algoritme til at separere dna-miksturer i de mest sandsynlige bidragsprofiler. Endvidere diskuteres det, hvorledes den kvantitative del af beviset inddrages i udregningen af den bevismæssige vægt.

I kriminalsager er det biologiske spor ofte fundet på gerningssteder under forhold, som kan nedbryde og forurene dna-strengen, hvilket besværliggør den senere biokemiske analyse. Ydermere kan mængden af dna være begrænset, hvilket kan udfordre sensitiviteten af bioteknologien anvendt i dna-analyserne. Modeller til at vurdere graden af nedbrudthed, samt estimere sandsynligheden for et alleludfald i dna-analysen behandles i afhandlingen, samt eksempler på hvorledes dette indkoorporeres i den bevismæssige vægt præsenteres.

Endelig indholder afhandlingen en artikel som omhandler processeringen af de kvantitative data observeret fra EPG'et detekteret af elektroforesemaskinerne efter PCR-processen. Centralt er detektionen af en signal-støjgrænse som hidtil har været en fast anbefalet grænse fra producenten af det kommercielle kit. I artiklen diskuteres det hvorledes grænsen kan fastsættes ud fra støjniveauet, således den kan være specifik for hver sag og dna-markør. Der præsenteres to yderligere filtre til håndtering af særlige typer af artefakter som udtrykkes i EPG'et som forstærkede signaler.

Acknowledgements

My biggest thanks goes to my supervisor through five years Poul Svante Eriksen from whom I have learned so much. Thank you for inspiring discussions and proposing solutions to many of the problems I have worked on. For always being encouraging and reading all my manuscript drafts of dubious quality and for debugging R-code during the past many years.

I also would like to thank my very good friend and office mate Ege for great times and discussions with and without beers involved. Thanks to the staff and colleagues at the Department of Mathematical Sciences for a friendly and inspiring place to work, and in particular the Head of Department E. Susanne Christensen for coming to Oxford in the first place and convincing me to work with forensic genetics in my MSc thesis and giving me the opportunity to write this PhD thesis.

I also would like to thank Helle Smidt Mogensen and Niels Morling, for sharing their insights in forensic genetics. For always proposing interesting problems and providing data in order to put statistics into forensic genetics. More inspirational and committed collaborators are hard to find. Furthermore, I would like to thank the entire Section of Forensic Genetics for friendly discussions, for the staffs interest in my work and making my visits in Copenhagen pleasant and fruitful. Finally, thanks to the University of Copenhagen for co-founding my PhD position.

Thanks to the New Zealanders in forensic genetics: Bruce Weir, John Buckleton and James Curran. Bruce for hosting my stay at the University of Washington, Seattle, during the end of 2008. The discussions and work in Seattle initiated my interest in population genetics, substructures and IBD. John and James for inviting me to summerly Auckland in the cold European winter 2010 to collaborate on our common interests in forensic genetics. I look forward to our future meetings and discussions. In addition I would like to thank Kund Højsgaards Fond, Oticon Fonden and Christian og Ottilia Brorsons Rejselegat for yngre videnskabdsmænd og -kvinder who supported my travels to USA and New Zealand. I would also like to thank Ellen og Aage Andersen's Foundation for financial support during my PhD studies.

Finally, I would like to thank my family and friends who have put up with me and shown an interest in my research over the past years. Last but not least I would like to thank Tenna for her everlasting sympathy and love over the years - and those to come. For letting me focus on my work during the final period of my PhD project and accepting my 'distant' moments in the few times of higher enlightenment. Your great cooking skills is a constant inspiration to me.

Aalborg, August 2010 Torben Tvedebrink

Contents

CHAPTER 1

Introduction

Forensic genetics is about drawing conclusions from biological evidence related to various types of crimes and legal disputes. It is the task of the forensic geneticists to present the genetic evidence as scientific and impartial as possible. The scientific aspects comprises thorough investigation of the various components in the analysis process of biological evidence. The analysis consists of several sub-analyses handling specific tasks on the route from tissue or body fluid to data used for interpretation. Since there are many sources of variability and uncertainty, the interpreter must be able to quantify the amount of uncertainty and include this when reporting the evidential weight.

Evidence from a scene of crime is subject to more sources of variability than samples taken in relation to family disputes. In the former case issues of contaminated samples or degraded DNA due to non-optimal conditions raises problems for the typing technology. The DNA might be too damaged for analysis or it might only be possible to obtain results from a subset of the DNA markers used for identification yielding partial DNA profiles. In paternity disputes or family reunification cases the problems facing the forensic geneticists are mainly related to population genetics and pedigree analysis since in these cases the reference samples are often of high quality and in sufficient amounts such that the risks of contamination, allelic drop-out or degradation of the biological material are minimal. However, the tissue used for identification of body remains found in the debris from a mass disaster or in mass graves is often severely degraded.

1.1 Qualitative models

Even before it was possible to obtain DNA profiles, biological features or phenotypes were used for evidential calculations. The blood type of a child is determined by the blood types of the parents' blood types. Hence, this information may be used in paternity disputes, where the alleged father can be excluded if there are inconsistencies in the constitutions of the trio's blood types. However, the few possible states of the blood type implies that the power of discrimination is low since many men unrelated to the child will share blood type with the true father.

Hence, the more polymorphic and diverse the biological marker, the more informative and powerful it is for discriminating among individuals. The development of DNA markers has minimised the problem of low discriminating power. By selecting DNA markers on different chromosomes forensic geneticists have obtained a powerful tool for making statements about paternity, relatedness and identity. The prevailing DNA typing technology used in forensic is based on the short tandem repeat (STR) typing technique.

The STR repeat sequences used in forensic genetics are typically made up by motifs of four or five base pairs, e.g. the typical repeat motif for TH0 is given by TCAT (Butler, 2005, Table 5.2). This implies that for locus TH0 an allele designated "6" has this motif repeated consecutively six times, which if often denoted $[\text{TCAT}]_6$.

Excluding abnormalities, every individual has two alleles per locus - one maternal and one paternal. However, it is impossible to determine the origin of the alleles and they may possibly be identical (homozygote) which implies only one allelic type is detected. Otherwise two distinct alleles are observed (heterozygote) and in either case at least one of the individual's parents share minimum one allele with their common offspring, assuming no mutations.

The commercial STR kits genotype 10 to 15 autosomal STR loci each having 10 to 25 frequently occurring alleles in the Danish population. That is, the qualitative part of the DNA profile consists of a set of loci where the DNA profile is specified by the states of the alleles. The heterozygous DNA profile with the highest probability in the Danish population using the SGM Plus kit (Applied Biosystems) is reported in Table 1.1.

Table 1.1: The heterozygous DNA profile with the highest probability in the Danish population.

Locus	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
Alleles	15,16	16,17	11,12	17,20	13,14	29,30	14,15	13,14	6,9.3	21,22

Since the STR loci are located on different chromosomes the laws of inheritance suggest that the allelic distribution over loci multiply: $P(A_{1i_1}A_{1j_1},\ldots,A_{Li_L}A_{Lj_L}) = \prod_{l=1}^{L} P(A_{li_l}A_{lj_l})$, where A_{li_l} is the ith allele on locus l and L is the total number of typed STR loci. Using the allele probabilities estimated for the Danish population, the probability of observing the DNA profile of Table 1.1 when sampling a random person from the population is 1.327×10^{-10}.

When a crime is committed, DNA evidence is often considered in the court of law, when convicting a suspect guilty or innocent. Let respectively H_p and H_d denote the hypotheses relating to the guilt and innocence of the suspect, and $\mathcal{E}$ the evidence relevant for the hypotheses. Then the court is interested in posterior ratio $P(H_p|\mathcal{E})/P(H_d|\mathcal{E})$. However, such statements are impossible for the forensic geneticist to quantify since this involves the prior ratio $P(H_p)/P(H_d)$ which is unknown to the forensic expert. What can be evaluated by the expert witness is the likelihood ratio $P(\mathcal{E}|H_p)/P(\mathcal{E}|H_d)$ using a model for the occurrence of the evidence given that the hypothesis is true.

The likelihood ratio, *LR*, is the essential quantity in forensic genetics and this thesis discuss several ways to include more of the available information in its evaluation. Consider a crime case with an identified suspect. Let G_S denote the suspect's DNA profile and $\mathcal{E}_c$ the DNA stain obtained from the scene of crime, and assume that $\mathcal{E}_c$ is consistent with G_S. That is, all alleles in G_S are present in $\mathcal{E}_c$ which we denote $G_S \equiv \mathcal{E}_c$. The two competing hypothesis state respectively; H_p: "The suspect is the donor of the DNA stain" and H_d: "An unknown and to the suspect unrelated person is the donor of the DNA stain". The latter hypothesis is what is called a "random man"-hypothesis. Let G_U denote the DNA profile of the random man which assuming no typing errors implies that $G_U \equiv G_S$. In this case the *LR* is given by:

$$LR = \frac{P(\mathcal{E}|H_p)}{P(\mathcal{E}|H_d)} = \frac{P(\mathcal{E}_c, G_S|H_p)}{P(\mathcal{E}_c, G_S|H_d)} = \frac{P(\mathcal{E}_c|G_S)P(G_S)}{P(\mathcal{E}_c|U)P(G_S|G_U)P(G_U)} = \frac{1}{P(G_U|G_S)}$$

where $P(G_U|G_S)$ under some model assumptions is the probability of observing the crime scene profile at random in the population. Hence, the evidence enables the forensic geneticist to make statements like "The probability of observing this particular profile at random in the reference population is 1 in 1,000,000" or equivalently "the DNA evidence is 1,000,000 times more likely under H_p than under H_d". There is an ongoing debate in the forensic genetic community on which probability is of relevance to the court. In the recent decade the "match probability" (Balding, 2005) that takes subpopulation structures or common coancestry into account has become prevalent. That is, rather than considering profiles of the suspect and "random man" as independent, one computes the probability of the observed profile conditioned on the suspect's profile. Hence, using the "posterior distribution" of alleles rather than the "prior distribution", rare alleles are less extreme. This implies a more conservative evaluation of the evidence since one accounts for the possibility that an allele that is rare in an admixed population is more commonly observed in one of its subpopulations to which the suspect (and possibly the culprit) might belong.

Over the recent years the national databases of STR profiles have grown in size due to the success of forensic DNA analysis in solving crimes. With these vast numbers of profiles available, it is possible to test the validity and applicability of population models to forensic genetics (Weir, 2004, 2007; Curran et al., 2007; Mueller, 2008). Furthermore, the accumulation of DNA profiles implies that the probability of a random match or near match of two randomly selected DNA profiles in the database increases. If all pairs of profiles are compared to each other in the database this corresponds to $\binom{n}{2} = n(n-1)/2$ pairwise comparisons in a database with n DNA profiles. In the Danish DNA reference database there are approximately 52,000 DNA profiles which yield 1,351,974,000 pairwise comparisons. With these large number of comparisons it is

likely to observe DNA profiles that coincide on many loci which has concerned some commentators and raised questions about “overstating” the power of DNA evidence. Hence, it is important to demonstrate that the observed and expected number of matches are sufficiently close in order to retain the confidence in DNA typing in general and the population genetic models used for evidential calculations in particular.

1.2 Quantitative models

The commercial kits used for analysis of DNA evidence provide quantitative and qualitative information to the analyst. The qualitative information reports which alleles that are present in the data (like in Table 1.1), whereas the qualitative part gives information on peak intensities in terms of height and area of the peaks obtained from the electropherogram (EPG). An example of an EPG is given in Figure 1.1 where the peak intensities (peak heights and areas) are plotted in relative fluorescent units (rfu) against the base pair (bp) length. Peaks with a low bp value correspond to alleles with short amplicons (amplicons are made up by the primer binding site and STR repeat structure). The peak intensities are measured by a CCD camera where the observed intensity corresponds to the amount of light emitted from the fluorescence dye.

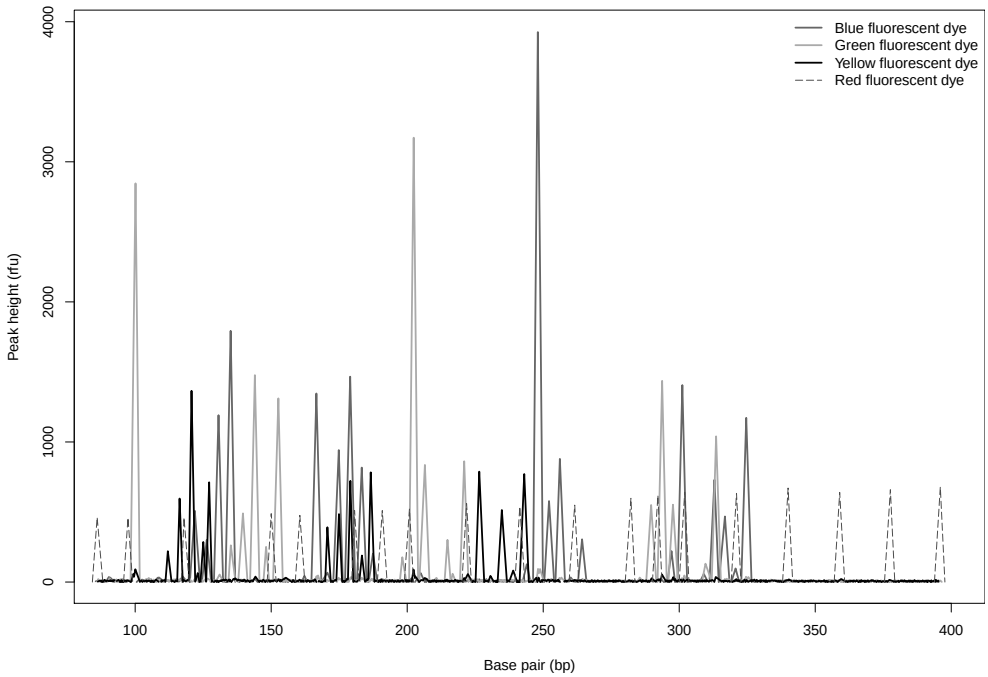

Figure 1.1: An example of an electropherogram (EPG) for the SGM-Plus kit (Applied Biosystems) with peak height (in rfu) plotted against the base pair (bp) length.

Thus the crime scene evidence, $\mathcal{E}_c$, consists of two components: The qualitative (or genetic) part, $\mathcal{G}$; and the quantitative part with peak intensities, $\mathcal{Q}$. The peak intensities reflect the amount of DNA contributed to the particular allele, and since the technology is indifferent to the origin of the various DNA fragments, the DNA amounts contributed to shared alleles add up. The resulting peak intensities are registered via a CCD camera that detects the light emitted from a fluorochrome attached to DNA molecules corresponding to a STR allele. A difference in electric potential forces the DNA molecules to move in the capillary, where the size difference of the molecules implies that some DNA fragments pass the CCD camera before others.

Since the length of the repeat sequences of the STR loci under investigation are overlapping, most commercial STR kits applies 3-5 different fluorochrome dyes in order to concurrently detect signals from multiple alleles and loci. One of these dyes contain DNA fragments of known length which are used for fragment size determination of the observed peak intensities. These fixed lengths are used to align the observed peak intensities to an allelic ladder which converts an observed fragment length to an allelic repeat number. For the SGM-Plus kit this size marker is given a red fluorochrome (represented by dashed lines in Figure 1.1).

In a single contributor DNA sample it is possible to observe one or two alleles per locus depending on whether the DNA profile is homozygous or heterozygous, respectively. However, when m DNA profiles contribute to the same sample it is possible to observe one to $2m$ alleles per locus, since the individuals may share all or no alleles. The peak intensities associated to the alleles reflect the amount of DNA contributed to that particular allele. Hence, in a two-person mixture alleles where the major component (the DNA profile with the largest amount of DNA contributed to the sample) contributes are often larger than those of the minor component. However, if the DNA profiles share alleles the peak intensities of the common alleles are approximately the sum of the contributions.

When assigning weight to the evidence under a given hypothesis the methodology needs to consider both parts of the data. This is particularly important when the data originate from a DNA mixture, since the quantitative evidence currently is the only way used to separate the observed alleles into contributing profiles. Often the peak intensities are used only to reduce the number of possible combinations entering the likelihood ratio. This approach is sometimes called the "binary model" in the forensic literature, e.g. by Bill et al. (2005); Buckleton et al. (2005). However, a more correct approach would be to attach a likelihood to each combination of profiles measuring the agreement between the observed peak intensities and the expected intensities under some model. Let the evidence $\mathcal{E} = (\mathcal{E}_c, \boldsymbol{K})$ where $\boldsymbol{K}$ are the known profiles associated to the crime, then the extended *LR* taking $\mathcal{Q}$ into account is given by:

$$LR = \frac{P(\mathcal{E}|H_p)}{P(\mathcal{E}|H_d)} = \frac{P(\mathcal{Q}, \mathcal{G}, \boldsymbol{K}|H_p)}{P(\mathcal{Q}, \mathcal{G}, \boldsymbol{K}|H_d)} = \frac{P(\mathcal{Q}|\mathcal{G}, \boldsymbol{K}, H_p)P(\mathcal{G}|\boldsymbol{K}, H_p)P(\boldsymbol{K}|H_p)}{P(\mathcal{Q}|\mathcal{G}, \boldsymbol{K}, H_d)P(\mathcal{G}|\boldsymbol{K}, H_d)P(\boldsymbol{K}|H_d)}, \tag{1.1}$$

where $P(\mathcal{Q}|\cdot)$ measures the agreement between the observed and expected peak intensities. Ideally the model for $P(\mathcal{Q}|\cdot)$ should take the entire EPG signal into account which includes the noise component (pictured as a "rug" close to 0 rfu in Figure 1.1), adjustment and correction for technical artefacts (stutters and pull-up effects, cf. below), detection of degradation (discussed in the end of this section), the genotypes of the contributors, etc.

However, evaluating the *LR* under such a model is computationally intense and complicated. That is, for each locus every pair of alleles constructed as a Cartesian product of the allelic ladder should be considered even though the peak height imbalances (ratio of peak heights) within and between loci were extreme. For practical purposes such an approach would be infeasible and too computational intense for standard case work. Hence, it is common to reduce $\mathcal{Q}$ to a smaller set of observations by using a criterion to separate the noise and signal into two parts, such that the number of possible combinations of DNA profiles decreases. A limit of detection is often used to discriminate between the noise and signal. However, such a threshold approach induces the risk of making wrong assignment of noise and signal, i.e. false positive and negative calls. In forensic genetics, these terms are commonly denoted drop-ins and drop-outs which refers to extra alleles in the signal not contributed by the true donors of the stain and missing alleles of the true donors being be below the limit of detection.

Let $\mathcal{Q}$ denote the part of the EPG that is classified as true signal. As mentioned above $\mathcal{Q}$ is currently the basis for separating DNA mixtures in its contributing components. That is, by defining a model for $\mathcal{Q}$ given a set of contributing profiles, it is possible to determine the goodness-of-fit between a hypothesised combination of DNA profiles and the observed peak intensities. Methods exist for modelling $P(\mathcal{Q}|\mathcal{G}, \boldsymbol{G}, H)$ of which some are more heuristic than statistical (Bill et al., 2005; Wang et al., 2006), but progress is made towards models based on statistical methods (Perlin and Szabady, 2001; Cowell et al., 2007a, b, 2010; Curran, 2008; Tvedebrink et al., 2010).

In cases where the amount of DNA contributed by the donor of the profile is low, there is a risk of the peak heights being below a limit of detection. The limit of detection is introduced in order to distinguish between noise and true signals. This may imply allelic drop-out which causes only a partial (or no) profile to be typed. Hence, a true contributing profile to an observed stain may have one or more alleles not present in the case sample. Not taking allelic drop-out into consideration could imply that the true donor is erroneously excluded from further consideration. In order to include the possibility for drop-out in the evidence evaluation it is necessary to be able to quantify this possibility in terms of a probability.

In contrast to drop-out which is "missing" alleles, the biotechnology used in the typing of DNA profiles may cause additional peaks to be present in the observed stain. The PCR process, which amplifies the DNA by making multiple copies of the present alleles, causes extra peaks in the position in front of the true peak. These peaks are called stutters and is due to mispairings between the Taq enzyme and amplicon. This creates a DNA product one repeat unit shorter than the true amplicon. Stutters may be produced in any cycle of the PCR process and a rule of thumb says that the stutter peak height is about 10-15% of the true peak height. This percentage is an overall value across alleles and loci, but shorter alleles tend to have lower stutter percentage than longer alleles.

Another systematic component caused by the typing technology are the so called pull-up (or bleed through) effects, where the light emitted from one fluorescent dye is detected in the spectra of a different fluorochrome. This implies false detection of peaks with similar fragment length as the parental peak, but on a different dye band. Furthermore, using a fixed limit of detection, of 50 rfu say, neglects important information about the noise level in a sample. If a peak in the interval 40 rfu to 49 rfu is observed, the fixed threshold-protocol determines this peak as undetected.

However, by using a model for the threshold, it might be reasonable to have a variable limit set such that e.g. 99.95% of all noise peaks are removed. This may for some cases imply a threshold as low as 25 rfu allowing for a more flexible analysis scheme which may be valuable for samples of low amounts of DNA.

When DNA is exposed to inhibitors such as chemicals, moisture, sunlight and heat, the DNA molecules are prone to degrade and the DNA strand damaged. This causes the results of the DNA investigation to have a characteristic profile with decreasing peak intensities as a function of the DNA fragment length. The longer the amplicon, the more likely it is that the peaks will have low emission values. This implies that the risk of allelic drop-outs increase for longer amplicons and may result in partial DNA profiles since some loci fail to produce any signal. Degraded biological material is pronounced in samples taken from the debris of mass disasters or mass graves.

1.3 Outline

The following seven chapters (Chapters 2-8) present the seven journal papers constituting this PhD thesis. The organisation of each chapter is such that the paper is presented in its journal form (including bibliography) followed by supplementary remarks about the results, how it relates to the previous chapters, further discussion and additional data analysis. As a consequence notation is not necessarily consistent between the chapters and some of the material is repeated in different chapters. On the other hand, the chapters may be read independently of each other. Each chapter has its own bibliography with the references used in there, and on the last pages of the thesis there is a complete list of all references. In chapters were there is a reference to supplementary material, e.g. as in journal papers, the material is available on-line at `http://people.math.aau.dk/~tvede/thesis`.

The order of the chapters is such that the number of factors considered in the evaluation of the evidence increases. First only the qualitative part of the data is considered in the likelihood ratio with the correction for population stratification effects. Later the quantitative data is added to the likelihood ratio where each model relaxes the assumptions made in the preceding chapters. Finally the last chapter combines the results and suggests topics for future research.

Chapter 2 discusses the topic of substructures in populations and how to account for this in evidential calculations. Concepts of identical-by-descent and subpopulations effects are common concepts from population genetics. The idea of measuring population stratification goes back to Wright (1951) who defined three quantities measuring the degree of relatedness between individuals, subpopulations and the total population. The model discussed in the chapter handles this from a statistical point of view by defining the correlation among individuals' DNA profiles as overdispersion and show how it is manifested in the so-called θ-correction used in forensic genetics.

Chapter 3 is an analysis of the Danish DNA profile reference database. By the beginning of 2009 the database included 51,517 unique DNA profiles typed on ten forensic autosomal STR loci. We investigated the methodology of Weir (2004, 2007) who made pairwise comparisons of every pair of DNA profiles in the database. We derived an efficient way to compute the expected number of matches and partial matches for a given θ, cf. above. Furthermore, in line with Curran et al. (2007) we extended the model to allow for closer familial relationships (full-siblings, first-cousins, parent-child and avuncular) and we derived expressions for the variance of the number of matches and partial matches in the database.

Chapter 4 is the first of five papers on the quantitative part of the data available from STR results. The paper is an extension of the work I did in my MSc thesis where the peak intensities of the EPG were modelled by a multivariate normal distribution. The challenging part of the model is the fact that the dimensions of the data vector (and sub-vectors hereof) vary among DNA mixtures due to the different number of shared alleles between individuals. An EM-algorithm was proposed for optimisation and we demonstrated that the model in fact is a mixed effects model.

Chapter 5 discusses a simpler and more operational model for DNA mixtures than the one from the previous chapter. In order to separate an observed DNA mixture into the contributing DNA profiles we derived a statistical model, which was suited for a greedy optimisation algorithm. The algorithm is very efficient, separating complex DNA mixtures in a few seconds. It is implemented as an on-line tool which provides valuable graphical output for further interpretation by the forensic geneticists.

Chapter 6 addresses an important question in forensic genetics and evidential calculations: Estimating the probability of allelic drop-out. We define a proxy for the amount of DNA contributed to a sample and use this quantity to derive an logistic regression model to estimate the probability of allelic drop-out.

Chapter 7 presents a methodology for filtering the quantitative data from STR results. The observed data is a conversion of emitted light from a fluorochrome detected by a CCD camera. This implies that the signal consists of a noise component and further systematic components, the so-called "pull-up effects" and "stutters". We demonstrate how to determine a floating threshold using distribution analysis of the noise component. Pull-up and stutter corrections were performed by regression analysis. The methodology decreases significantly the number of allelic drop-outs compared to the standard protocol.

Chapter 8 is a short communication on how to model degraded DNA in a simple and intuitive manner. Degraded DNA is a common problem in crime case samples since the biological material from which the DNA is extracted has often been exposed to non optimal conditions. Sunlight, humidity, bacteria and chemicals are some of the reasons for observing degraded DNA which complicate the succeeding analysis and interpretation. The model presented in the paper is used to modify the drop-out model discussed in Chapter 6 by adjusting the proxy for the amount of DNA taking the level of degradation into account.

Chapter 9 summarises the results from the proceeding seven chapters by forming a 'unifying' likelihood ratio. The terms in this likelihood ratio consist of:

$$P(\mathcal{Q}_{\text{mis}}|\mathcal{Q}_{\text{obs}}, \mathcal{G}_{\text{mis}}, \mathcal{G}_{\text{obs}}, \boldsymbol{G})P(\mathcal{Q}_{\text{obs}}|\mathcal{G}_{\text{mis}}, \mathcal{G}_{\text{obs}}, \boldsymbol{G})P(\mathcal{G}_{\text{mis}}, \mathcal{G}_{\text{obs}}|\boldsymbol{K}, \boldsymbol{G})P(\boldsymbol{K}|\boldsymbol{G})P(\boldsymbol{G}),$$

where $\boldsymbol{G}$ is a combination of DNA profiles consistent with the hypothesis under consideration. Furthermore, $\mathcal{Q}$ and $\mathcal{G}$ symbolises the quantitative and qualitative parts of the evidence, respectively. The first term, $P(\mathcal{Q}_{\text{mis}}|\cdot)$ evaluates the probability of allelic drop-out using the models of Chapters 6, 7 and 8, $P(\mathcal{Q}_{\text{obs}}|\cdot)$ is evaluated by one of the models for DNA mixtures (Chapters 4 and 5), while the last terms are evaluated using the θ-correction discussed in Chapters 2 and 3.

CHAPTER 2

Overdispersion in allelic counts and θ-correction in forensic genetics

Publication details

Co-authors: None

Journal: Theoretical Population Biology (In Press)

DOI: 10.1016/j.tpb.2010.07.002

Abstract:
We present a statistical model for incorporating the extra variability in allelic counts due to subpopulation structures. In forensic genetics, this effect is modelled by the identical-by-descent parameter θ, which measures the relationship between pairs of alleles within a population relative to the relationship of alleles between populations (Weir, 2007). In our statistical approach, we demonstrate that θ may be defined as an overdispersion parameter capturing the subpopulation effects. This formulation allows derivation of maximum likelihood estimates of the allele probabilities and θ together with computation of the profile log-likelihood, confidence intervals and hypothesis testing.

In order to compare our method with existing methods, we reanalysed FBI data from Budowle and Moretti (1999) with allele counts in six US subpopulations. Furthermore, we investigate properties of our methodology from simulation studies.

Keywords:
θ-correction; Forensic genetics; Subpopulation; Dirichlet-multinomial distribution; Maximum likelihood estimate; Confidence interval.

2.1 Introduction

Attaching probabilities to different levels of relatedness in paternity disputes or evaluating the weight of evidence in crime cases with biological traces present at the scene of crime are essential tasks in forensic genetics. To this purpose, the difference in the genetic constitution of individuals in the population is used to assess the probabilities of the evidence under competing hypotheses. Currently, 10 to 20 locations on the genome (loci) are investigated for identification purposes and an individual's DNA profile is made up by the different states (alleles) of the loci.

It is well known that allele frequencies may vary between ethnic groups, geographic remote populations and subpopulations. However, due to a common evolutionary past it is assumed that the allele frequencies of the subpopulations have a common mean, and that the variation between subpopulations is due to genetic sampling (Weir, 1996).

In forensic genetics, population structures are of great importance when the probability of observing a given suspect's DNA profile is assessed under various hypotheses. Budowle and Moretti (1999) published allele frequencies from six different US subpopulations (African American, Bahamian, Jamaican, Trinidad, Caucasian and Hispanic) for 13 CODIS Core STR loci. In this study, the authors obtained allele frequency estimates varying significantly across subpopulations. For example, the frequencies range from 6.9% (Hispanic) to 27.3% (Jamaican) for allele 28 in locus D21, indicating that a homozygote on this locus could be 16 times more likely in the Jamaican than in the Hispanic subpopulation (when assuming Hardy-Weinberg equilibrium). The ability to distinguish true genetic differences from sampling effects depends on the sample size. That is, testing the significance of such allele frequency differences depends on the database sizes, since the variance of the estimates scales inversely with the number of sampled DNA profiles.

In order to correct for subpopulation structure, Nichols and Balding (1991) suggested the "θ-correction" to be used when inferring the weight of evidence in forensic genetics. Our approach acknowledges the extra variability in the allelic counts and addresses this as overdispersion. The statistical model of the present paper has the same properties as the genetic model. We exploit results from the statistical literature in order to obtain maximum likelihood estimates (MLEs) of the allele frequencies and θ-parameter, and compute profile log-likelihoods for θ providing approximate confidence intervals.

The basic idea and principle of overdispersion in allelic counts formulated in Section 2.2 has previously been noted in the forensic literature, although not called overdispersion, by Balding (2005). However, the terminology of overdispersion (or heterogeneity) explicitly underlines that a simple assumption of the sampling distribution (multinomial distribution) is insufficient to model the data. By "overdispersion" it becomes more transparent to statisticians with limited knowledge in population genetics to appreciate the concept of variability between population groups. Hence, these rather specialised types of model are put into a more general statistical framework.

2.2 Overdispersion in allelic counts

Our set-up assumes that the allelic counts in a given subpopulation $\boldsymbol{X}$ follow a multinomial distribution with some unknown allele probabilities. Due to an evolutionary past, there exists some variation among different subpopulations in terms of allele probabilities. However, these allele probabilities have a common distribution across subpopulations with a mean and variance. For now, we just let $\mathbb{E}(\boldsymbol{P}) = \boldsymbol{\pi}$ be the mean of this distribution and $\mathbb{V}(\boldsymbol{P})$ its covariance matrix. Note that this parametrisation of $\mathbb{E}(\boldsymbol{P})$ implies that $\boldsymbol{\pi}$ are the allele probabilities in the reference population from which the subpopulations are assumed to have descended.

Let n be the number of alleles sampled from a given subpopulation with k alleles. Then the model can be formulated as

$$P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{P} = \boldsymbol{p}) = \binom{n}{\boldsymbol{x}} \prod_{j=1}^{k} p_j^{x_j}, \quad \text{where} \quad \binom{n}{\boldsymbol{x}} = \frac{n!}{\prod_{j=1}^{k} x_j!}, \tag{2.1}$$

is the multinomial coefficient. Thus $\boldsymbol{X}$ follows a multinomial distribution when conditioned on $\boldsymbol{P} = \boldsymbol{p}$. This implies that $\mathbb{E}(\boldsymbol{X}) = \mathbb{E}(\mathbb{E}\{\boldsymbol{X}|\boldsymbol{P}\}) = \mathbb{E}(n\boldsymbol{P}) = n\boldsymbol{\pi}$ from the assumption of $\mathbb{E}(\boldsymbol{P}) = \boldsymbol{\pi}$.

2.2.1 Dirichlet-multinomial distribution

In line with other authors (Lange, 1995b,a; Weir, 1996; Rannala and Hartigan, 1996; Balding, 2003), we assume the distribution of allele probabilities to be a Dirichlet distribution. The assumption of a Dirichlet distribution is based on theoretical arguments from population genetics together with the convenience that the Dirichlet distribution is the conjugate prior of the multi-

nomial distribution. The Dirichlet distribution has density function

$$f(p_1,\ldots,p_k;\gamma_1,\ldots,\gamma_k) = \frac{\Gamma(\gamma_+)}{\prod_{j=1}^k \Gamma(\gamma_j)} \prod_{j=1}^k p_j^{\gamma_j-1}, \tag{2.2}$$

where $\gamma_+ = \sum_{j=1}^k \gamma_j$. When assuming a Dirichlet distribution of $\boldsymbol{P}$, we can derive the marginal distribution of $\boldsymbol{X}$ by multiplying (2.2) and (2.1) and integrating over $\boldsymbol{p}$. The resulting distribution is called the Dirichlet-multinomial distribution (Johnson et al., 1997) or multivariate Pòlya distribution (from its relation to the Pòlya urn scheme, Green and Mortera (2009)) with density

$$P(\boldsymbol{X}=\boldsymbol{x}) = \binom{n}{\boldsymbol{x}} \frac{\Gamma(\gamma_+)}{\Gamma(n+\gamma_+)} \prod_{j=1}^k \frac{\Gamma\left(x_j+\gamma_j\right)}{\Gamma\left(\gamma_j\right)}. \tag{2.3}$$

Using the results of Mosimann (1962), the mean of X_j may be computed as $\mathbb{E}(X_j) = n\gamma_j/\gamma_+$, where γ_j/γ_+ is the mean of P_j, $\mathbb{E}(P_j) = \pi_j = \gamma_j/\gamma_+$. Furthermore, the covariance matrix of $\boldsymbol{X}$ is given by $\mathbb{V}(\boldsymbol{X}) = cn[\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top]$, where $c = (n+\gamma_+)/(1+\gamma_+)$ and $\boldsymbol{\pi}^\top$ is the transpose of $\boldsymbol{\pi}$. Hence, the covariance matrix of the Dirichlet-multinomial distribution is inflated by the factor c compared to an ordinary multinomial covariance.

The Dirichlet-multinomial distribution derived in (2.3) is almost identical to Eq. (8) in Curran et al. (1999) except for the multinomial coefficient, which is merely a constant with respect to the parameters of the model. Furthermore, by introducing θ as in Curran et al. (1999), $\gamma_+ = (1-\theta)/\theta$ or equivalently $\theta = (1+\gamma_+)^{-1}$, we may rewrite c in terms of θ:

$$c = \frac{n+\gamma_+}{1+\gamma_+} = (n+\gamma_+)\theta = n\theta + (1-\theta) = 1+\theta(n-1).$$

This implies that $\mathbb{V}(\boldsymbol{X}) = n[\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top][1+\theta(n-1)]$ which is identical to the variance in Curran et al. (1999). In Curran et al. (1999), this expression was derived by letting θ denote the identical-by-descent (IBD) parameter, whereas in the statistical model θ is an overdispersion parameter.

A direct implication from $\boldsymbol{X}$ being Dirichlet-multinomial distributed is that the vector of proportions $\tilde{\boldsymbol{P}} = \{\tilde{P}_j\}_{j=1}^k = \{X_j/n\}_{j=1}^k$ is an unbiased estimator of $\boldsymbol{\pi}$ with covariance matrix $n^{-1}[\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top][1+\theta(n-1)]$. When $\boldsymbol{X}$ follows a multinomial distribution, $\tilde{\boldsymbol{P}}$ is the maximum likelihood estimator of $\boldsymbol{\pi}$. However, under the Dirichlet-multinomial model this variance does not go to zero even for very large sample sizes n,

$$\lim_{n\to\infty} \left[\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top\right]\left(\theta + \frac{1-\theta}{n}\right) = \left[\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top\right]\theta,$$

as opposed to the asymptotic behaviour under the multinomial model where $\lim_{n\to\infty} n^{-1}[\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top] = \boldsymbol{O}$, with $\boldsymbol{O}$ being the null matrix.

Let $\boldsymbol{Y}^n = (Y_1,\ldots,Y_n)$ denote the vector of sampled alleles, of which $\boldsymbol{X}^n$ is the sufficient statistic, where the superscript is added to stress that n alleles were sampled. Consider the probability

$P(Y_{n+1} = j|\boldsymbol{Y}^n = \boldsymbol{y}^n)$, i.e. the probability of a future j allele given the alleles previously sampled:

$$\begin{aligned} P(Y_{n+1} = j|\boldsymbol{Y}^n = \boldsymbol{y}^n) &= \frac{\int f(\boldsymbol{p})P(Y_{n+1} = j|\boldsymbol{p}) \prod_{i=1}^n P(Y_i = y_i|\boldsymbol{p})\mathrm{d}\boldsymbol{p}}{\int f(\boldsymbol{p}) \prod_{i=1}^n P(Y_i = y_i|\boldsymbol{p})\mathrm{d}\boldsymbol{p}} \\ &= \frac{\Gamma(\gamma_j + x_j^{n+1})\Gamma(\gamma_+ + n)}{\Gamma(\gamma_j + x_j^n)\Gamma(\gamma_+ + n + 1)} \\ &= \frac{x_j^n\theta + (1-\theta)\pi_j}{1 + (n-1)\theta}, \end{aligned} \tag{2.4}$$

where we used $f(\boldsymbol{p})$ from (2.2) and $x_j^{n+1} = x_j^n + 1$. This expression emphasises that the probability of observing a future j allele only depends on the previous sampled alleles through the total allele count, n, and how many of these alleles were of type j, x_j^n. Hence, we also apply the notation $P(j|x_j^n)$ for this probability which is identical to $P_n(A) = (n_A\theta + \{1-\theta\}p_A)/(1 + \{n-1\}\theta)$ in the recursion equation of Balding and Nichols (1997, equation (1) where we changed their notation from F for θ), which is the probability of observing an A allele after n_A of n alleles being of type A.

2.2.2 Application to paternity testing

Forensic genetics is widely used in paternity disputes or when a person applies for a family reunification. In the setting of a paternity dispute, let H_1 be the hypothesis: "The alleged father is the true father" and H_2 the hypothesis: "A man unrelated to the alleged father is the true father". The paternity index (PI) is defined as $PI = P(E|H_1)/P(E|H_2)$, where the evidence, E, is the DNA profiles of the involved individuals, i.e. child, mother and alleged father.

In paternity testing, the θ-correction enters the PI through the assumption of correlated individuals in the population due to subpopulation structures (Balding and Nichols, 1995; Evett and Weir, 1998). Consider only one locus where a child's DNA profile is (ac) and its mother's profile is (ab). Assuming no mutations, the true father must pass on a c allele to the child. If the alleged father's DNA profile is (cd), the H_1-hypothesis implies that the parents (ab, cd) have offspring (ac). The probability of the evidence, given H_1, is computed as $P(E|H_1) = P(ac|ab, cd)P(ab, cd)$, where $P(ac|ab, cd)$ is the probability that a child is ac when its parents are (ab, cd), i.e. $P(ac|ab, cd) = \frac{1}{4}$, and $P(ab, cd)$ is the probability for observing alleles a, b, c and d in the population. The other hypothesis, H_2, claims that the child got its c allele from a man unrelated to the alleged father. Then the paternity index, as derived in Appendix 2.A.1, is given by

$$PI(\theta) = \frac{1 + 3\theta}{2[\theta + (1-\theta)\pi_c]}, \tag{2.5}$$

where $PI(\theta)$ is used to emphasise PI's dependence on θ. Table 1 and 2 in Balding and Nichols (1995) give the (reciprocal) $PI(\theta)$ for other parent-child scenarios (with θ denoted by F).

As an example, let us assume that this specific trio scenario is replicated for all S loci used for DNA profile testing. The consequence between applying $\theta > 0$ and and using the simple $PI(0) =$

$1/(2p_c)$ for independent profiles is very pronounced even for reasonably common alleles. If $\pi_c = 0.025$ and $\theta = 0.03$, then $PI(0.03) \approx 10$ while $PI(0) = 20$. Hence, the numerical difference between the two paternity indexes is for S independent DNA markers $\{PI(0)/PI(0.03)\}^S \approx 2^S$, which for the typical forensic typing kits with $S \geq 10$ yields a ratio of at least 1,000. That is, the evidential weight may decrease by several orders of magnitude by correcting for possible IBD or population stratification.

2.2.3 Application to DNA mixtures

When two or more individuals contribute to a biological stain, the observable DNA profile is a mixture of the various alleles contributing to the stain, and is therefore called a DNA mixture. In an m-person DNA mixture, it is possible to observe 1 to $2m$ alleles per locus, since the involved DNA profiles may share all or no alleles (see e.g. Tvedebrink et al., 2010, for a further discussion of DNA mixtures). Assume for a two-person mixture, e.g. a rape case, that we observe the alleles (abc) and that the victim's DNA profile is (ab) and the suspect's DNA profile is (cc). Then, in line with the paternity index, the likelihood ratio is defined as $LR = P(E|H_1)/P(E|H_2)$, where E is the evidence (abc) and the known DNA profiles and H_1 and H_2 is the prosecutor's and defence's hypotheses, respectively (in the literature H_p and H_d are commonly used for the same hypotheses). The hypothesis H_1 states "The victim and suspect constitute the DNA mixture" whereas H_2 acquits the suspect: "The victim and an unknown individual constitute the DNA mixture". Let $P(abc|ab, ij)$ be the probability of observing the crime scene stain (abc) given the mixture originates from genotypes ab and ij. When assuming no typing errors this probability is 1 if $(ij) \in \{(ac), (bc), (ca), (cb), (cc)\}$ and 0 otherwise. In line with the derivation of $PI(\theta)$ (see Appendix 2.A.1 for the details of $PI(\theta)$), we get

$$\begin{aligned} LR(\theta) &= \frac{P(abc|ab, cc)P(ab, cc)}{\sum_{ij} P(abc|ab, ij)P(ab, cc, ij)} \\ &= \frac{(1 + 3\theta)(1 + 4\theta)}{(7\theta + \{1 - \theta\}[2\pi_a + 2\pi_b + \pi_c])(2\theta + \{1 - \theta\}\pi_c)} \end{aligned} \tag{2.6}$$

In Figure 2.1, we have plotted the $LR(\theta)$ function for the DNA mixture above with π_a and π_b fixed at 0.1. The solid line represents the uncorrected $LR(\theta = 0)$ and the broken lines show the corrected $LR(\theta)$ for θ-values as described by the legend. The inserted plot shows the behaviour close to the value $\pi_c = 0.71$ where the effect of the θ-correction is reversed. We see that the effect of θ is minimal for common alleles and more pronounced for rare ones. Hence, in practice the larger θ is the more conservative the LR-estimates are.

The use of θ in evidential computations can be seen as a means to smoothing the allele probabilities over possible subpopulations and thereby adjusting for the uncertainty associated with unobserved or unobservable substructures in the larger database. This latent structure may be seen as a reason for overdispersion in statistical terms, i.e. inhomogeneity due to unobserved/unobservable variables.

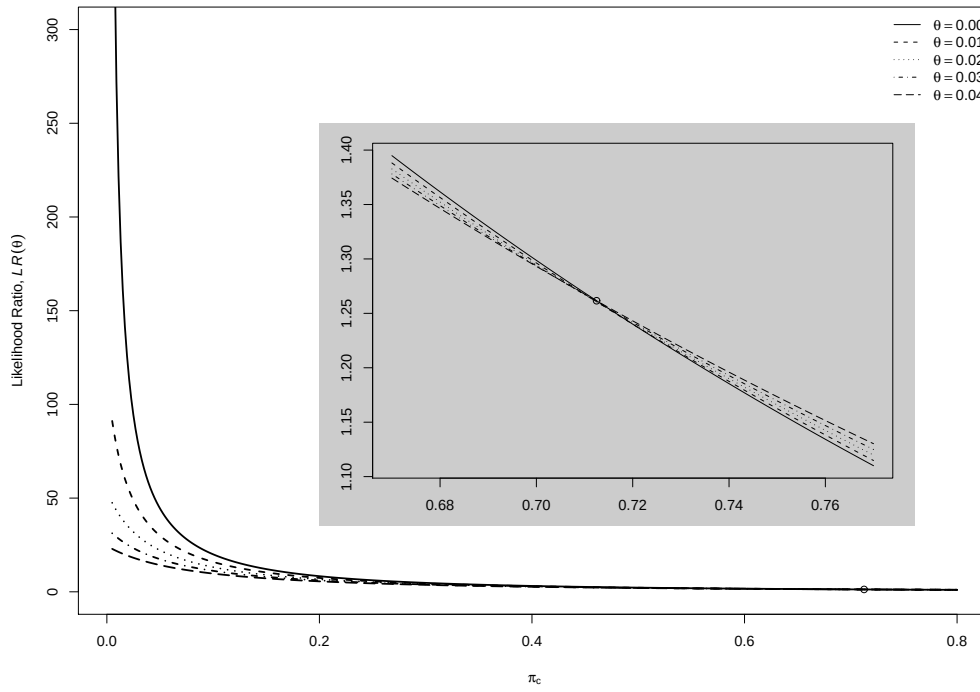

Figure 2.1: The effect of θ on $LR(\theta)$ for a single locus as exemplified. $LR(\theta)$ is plotted for various θ-values ranging from 0.00 (no subpopulation effect) to 0.04 (large subpopulation effect) against the allele frequency of the allele in question (here allele c) with the other probabilities (π_a and π_b) fixed at 0.1. Inserted is a blow-up of the curve near $\pi_c = 0.71$ ($\circ$ marks this point).

If the suspect or alleged father in the two situations considered above has a ethnicity or nationality that indicates that a specific database is representative for his genetic origin then allele frequencies estimated from this database are the most appropriate reference sample to use for evidential weight calculations. However, the database and the population that it resembles may be constituted by several subpopulations or groups, which causes this conceptual population to be heterogeneous. That is, geopolitical or tribal structures together with marital and religious preferences may induce genetic diversity causing overdispersion. Hence, a database that seems to be the most appropriate for a particular suspect may not be sampled on a sufficiently high resolution to obtain a homogeneous reference subpopulation. In fact, it may not even be possible to obtain samples with this property. Thus, genetic diversity and the resulting overdispersion in allele counts must be accounted for by the θ-correction.

2.3 Parameter estimation

Assume that we have allelic counts from N different subpopulations such that x_{ij} denotes the number of allele j in subpopulation i and that for each subpopulation $i, i = 1, \ldots, N$, there is a total of n_i counts, $\boldsymbol{n} = (n_1, \ldots, n_N)$. In addition, we assume that the subpopulations are independent, implying that the likelihoods of the counts from the subpopulations multiply.

The likelihood may then be derived by multiplying over the terms of (2.3). This likelihood implies differentiation of Γ-functions in order to solve the likelihood equations. A useful observation about the Γ-function is that

$$\prod_{r=1}^{y}\{\alpha + (r-1)\} = \alpha \times \cdots \times (\alpha + y - 1) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)}, \tag{2.7}$$

using the fact that $x\Gamma(x) = \Gamma(x+1)$. Hence, an equivalent way of expressing the distribution in (2.3) using the rising factorials of (2.7) is given as

$$P(\boldsymbol{X} = \boldsymbol{x}) = \binom{n}{\boldsymbol{x}} \frac{\prod_{j=1}^{k}\prod_{r=1}^{x_j}\{\pi_j(1-\theta) + (r-1)\theta\}}{\prod_{r=1}^{n}\{1 - \theta + (r-1)\theta\}}.$$

From this probability function we can compute the log-likelihood function $\ell(\boldsymbol{\pi}, \theta; \boldsymbol{x})$. Discarding the multinomial constant (which is a constant with respect to the parameters), the log-likelihood is

$$\ell(\boldsymbol{\pi}, \theta; \boldsymbol{x}) = \sum_{i=1}^{N}\sum_{j=1}^{k}\sum_{r=1}^{x_{ij}} \log\{\pi_j(1-\theta) + (r-1)\theta\} - \sum_{i=1}^{N}\sum_{r=1}^{n_i} \log\{1 - \theta + (r-1)\theta\}. \tag{2.8}$$

The corresponding likelihood equations, $\partial\ell(\boldsymbol{\pi}, \theta; \boldsymbol{x})/\partial(\boldsymbol{\pi}, \theta) = \boldsymbol{0}$, cannot be solved analytically for the parameters; hence numerical methods need to be invoked for parameter estimation. Let $\boldsymbol{\psi}$ denote the parameter vector $\boldsymbol{\psi} = (\boldsymbol{\pi}, \theta) = (\{\pi_j\}_{j=1}^{k-1}, \theta)$, since $\pi_k = 1 - \sum_{j=1}^{k-1}\pi_j$. A possible numerical method for solving the likelihood equations is Fisher-scoring, where the parameter estimates in each iteration are updated using $\hat{\boldsymbol{\psi}}_{(m+1)} = \hat{\boldsymbol{\psi}}_{(m)} + \{\mathcal{I}(\hat{\boldsymbol{\psi}}_{(m)})\}^{-1}\boldsymbol{u}(\hat{\boldsymbol{\psi}}_{(m)})$, where $\hat{\boldsymbol{\psi}}_{(m)}$ is the estimate in the mth iteration, $\boldsymbol{u}(\hat{\boldsymbol{\psi}}_{(m)})$ is the score function, $\partial\ell(\boldsymbol{\psi}; \boldsymbol{x})/\partial\boldsymbol{\psi}$, and $\mathcal{I}(\hat{\boldsymbol{\psi}}_{(m)})$ is the expected Fisher Information Matrix (FIM) both evaluated in $\hat{\boldsymbol{\psi}}_{(m)}$. Paul et al. (2005) derived exact expressions for the expected FIM entries, $\mathcal{I}(\boldsymbol{\psi})$. The results of Paul et al. (2005) imply that the expected FIM may be computed using expressions only involving the marginal distributions of X_j. Similar results were obtained by Neerchal and Morel (2005).

2.3.1 Computational considerations

Most of the methodology discussed in this section and subsections hereof have been implemented in the R-package `dirmult` available on-line in the CRAN repository at http://www.r-project.org (Tvedebrink, 2009).

Even though the expressions for the expected FIM, $\mathcal{I}(\boldsymbol{\pi}, \theta)$, given in (Paul et al., 2005, pp. 232) are compact, they cause the parameter estimation to be computationally inefficient. Numerical

work has shown that it is much more convenient to estimate the γ-parameters and transform the estimates, rather than estimate θ and $\boldsymbol{\pi}$ directly. The log-likelihood $\ell(\boldsymbol{\gamma}; \boldsymbol{x})$ is

$$\ell(\boldsymbol{\gamma}; \boldsymbol{x}) = \sum_{i=1}^{N} \sum_{j=1}^{k} \sum_{r=1}^{x_{ij}} \log\{\gamma_j + r - 1\} - \sum_{i=1}^{N} \sum_{r=1}^{n_i} \log\{\gamma_+ + r - 1\}, \tag{2.9}$$

where we used (2.3) and (2.7). The first-order and second-order derivatives of the log-likelihood $\ell(\boldsymbol{\gamma}; \boldsymbol{x})$ are given by

$$\frac{\partial \ell(\boldsymbol{\gamma}; \boldsymbol{x})}{\partial \gamma_j} = \sum_{i=1}^{N} \left\{ \sum_{r=1}^{x_{ij}} \frac{1}{\gamma_j + r - 1} - \sum_{r=1}^{n_i} \frac{1}{\gamma_+ + r - 1} \right\} \tag{2.10}$$

$$\frac{\partial^2 \ell(\boldsymbol{\gamma}; \boldsymbol{x})}{\partial {\gamma_j}^2} = \sum_{i=1}^{N} \left\{ \sum_{r=1}^{n_i} \frac{1}{(\gamma_+ + r - 1)^2} - \sum_{r=1}^{x_{ij}} \frac{1}{(\gamma_j + r - 1)^2} \right\} \tag{2.11}$$

$$\frac{\partial^2 \ell(\boldsymbol{\gamma}; \boldsymbol{x})}{\partial \gamma_j \partial \gamma_l} = \sum_{i=1}^{N} \sum_{r=1}^{n_i} \frac{1}{(\gamma_+ + r - 1)^2}, \tag{2.12}$$

where (2.10) gives the elements of the score function $\boldsymbol{u}(\boldsymbol{\gamma})$. Furthermore, this implies that the diagonal elements of the expected FIM, $\mathcal{I}(\boldsymbol{\gamma})$, are

$$\mathcal{I}(\gamma_j, \gamma_j) = \sum_{i=1}^{N} \left\{ \sum_{r=1}^{x_{ij}} \frac{P(X_{ij} \geq r)}{(\gamma_j + r - 1)^2} - \sum_{r=1}^{n_i} \frac{1}{(\gamma_+ + r - 1)^2} \right\},$$

for $j = 1, \ldots, k$, and the off-diagonal elements, $\mathcal{I}(\gamma_j, \gamma_l)$, equal (2.12). However, for most practical purposes using the observed FIM, $\mathcal{J}(\boldsymbol{\gamma})$, rather than the expected FIM, $\mathcal{I}(\boldsymbol{\gamma})$, in the Newton-Raphson scoring ensures much lower computational time. Numerical investigations indicate that the $\mathcal{J}(\boldsymbol{\gamma})$-implementation converges to the same extrema and much more quickly as the diagonal elements, $\mathcal{J}(\gamma_j, \gamma_j)$, for this matrix are as in (2.11), i.e. the terms $P(X_{ij} \geq r)$, $r = 1, \ldots, x_{ij}$, where $X_{ij} \sim$ Beta-Binomial$(\gamma_j, \gamma_+ - \gamma_j)$, need not be computed.

The inverse of the expected FIM is the asymptotic covariance matrix of the MLE. As our interest is in $(\boldsymbol{\pi}, \theta)$, we exploit that $\mathcal{I}(\boldsymbol{\pi}, \theta) = \Delta^\top \mathcal{I}(\boldsymbol{\gamma}) \Delta$, where $\{\Delta\}_{ij} = \{\partial \boldsymbol{\gamma} / \partial \boldsymbol{\psi}\}_{ij}$.

Simulations

Standard asymptotic theory assures that the MLE is the most efficient estimator. However, inference about θ depends mainly on the number of subpopulations sampled, N, and only to a minor degree on the subpopulation sample sizes, $\boldsymbol{n}$. Hence, in order to verify our implementation and the performance of the maximum likelihood estimator for different number of subpopulations, we simulated data with known allele frequencies, $\boldsymbol{\pi}$, and θ-value. When simulating the mth data matrix, $\boldsymbol{x}_m$, for $m = 1, \ldots, M$, we used the following sampling scheme:

1. Draw $\boldsymbol{p}'_{i,m} \sim$ Dirichlet$(\{\pi_j(1-\theta)/\theta\}_{j=1}^{k})$, $i = 1, \ldots, N$.
2. Draw $\boldsymbol{x}_{i,m} \sim$ Multinomial$(n_i, \boldsymbol{p}'_{i,m})$, $i = 1, \ldots, N$.
3. The mth data matrix is $\boldsymbol{x}_m = [\boldsymbol{x}_{1,m}, \ldots, \boldsymbol{x}_{N,m}]^\top$.

This ensures that the random variable $\boldsymbol{X}_i$, of which $\boldsymbol{x}_{i,m}$ is a realisation, follows a Dirichlet-multinomial distribution with parameters $\boldsymbol{\pi}$ and θ. Note that the concept of N subpopulations is a theoretical one. In practice only an overall database would exist which neglects the present substructure. However, the intension is to account for this partitioning using the θ-correction.

In Weir and Hill (2002), the authors argue that if the expectation of a ratio was the ratio of expectations then the method of moment (MoM) estimator, $\tilde{\theta}_{\text{MoM}}$, of Weir and Hill (2002, equation 5) was an unbiased estimator of θ:

$$\tilde{\theta}_{\text{MoM}} = \frac{\sum_{j=1}^{k}(\text{MSP}_j - \text{MSG}_j)}{\sum_{j=1}^{k}(\text{MSP}_j + (n_c - 1)\text{MSG}_j)},$$

where $n_c = (N-1)^{-1}\left(\sum_{i=1}^{N} n_i - n_+^{-1}\sum_{i=1}^{N} n_i^2\right)$ and $n_+ = \sum_{i=1}^{N} n_i$. The quantities MSG_j and MSP_j are two mean squares defined as

$$\text{MSP}_j = \frac{1}{N-1}\sum_{i=1}^{N} n_i(\tilde{p}_{ij} - \bar{p}_j)^2 \quad \text{and} \quad \text{MSG}_j = \frac{1}{\sum_{i=1}^{N}(n_i-1)}\sum_{i=1}^{N} n_i\tilde{p}_{ij}(1-\tilde{p}_{ij})$$

with $\tilde{p}_{ij} = x_{ij}/n_i$, $\bar{p}_j = n_+^{-1}\sum_{i=1}^{N} x_{ij}$. Even though the expectation does not satisfy the property mentioned above, the $\tilde{\theta}_{\text{MoM}}$-estimator seems to perform reasonably well on average.

More recently, Zhou and Lange (2010) has derived MM (Minorisationmaximisation) algorithms for some discrete multivariate distributions and among these the Dirichlet-multinomial distribution. The authors have provided Matlab scripts (on line supplementary material available at the website of Journal of Computational and Graphical Statistics) for estimating parameters in the MM set-up.

In the following we compare the MLE, MoM and MM estimates on simulated data using the relative frequencies in locus D13 from data published in Budowle and Moretti (1999) as $\boldsymbol{\pi}$ and $\theta = 0.03$. The box plot in Figure 2.2 show θ-estimates of 100 simulated datasets ($M = 100$) with sample sizes, n_i, of 200 and an increasing number of databases (increasing number of subpopulations, N).

From the box plot it is evident that the MLE has a lower variance, but also that on average the MoM and MM estimates are closer to the true value. However, as the number of databases increases so does the accuracy of the estimates, as one would expect. In addition to the accuracy of the estimation procedure, it is relevant to compare the computational speed and ease of implementation of the various methods. Naturally, the MoM estimator is the easiest to implement, and since no iterations are applied, "convergence" happens immediately. Both MLE and MM estimates are based on iterative procedures. Where several statistical tools exist for easy implementation of Newton-Raphson iterations, a little more code needs to be written for MM algorithms. However, the script-files of Zhou and Lange (2010) elegantly demonstrated how these obstacles can be handled in Matlab. We compared the computation times for the various iterative methods (Zhou and Lange, 2010, implemented simple and more advanced MM methods in their paper) and number of iterations needed to satisfy the convergence criteria. The MLE method implemented in R is always faster and needs fewer iterations for convergence compared

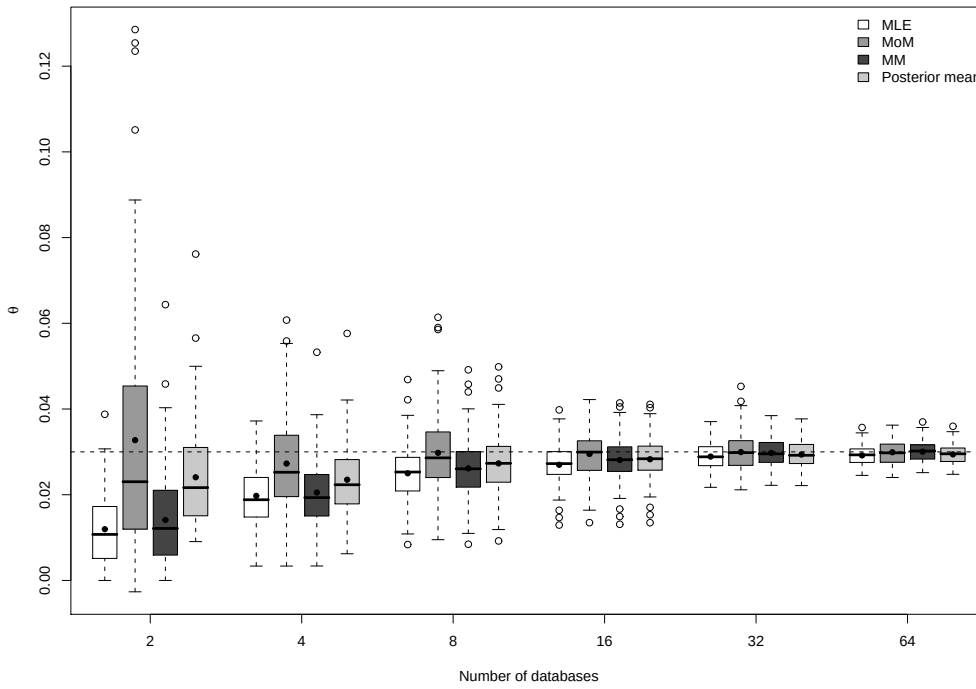

Figure 2.2: Box plots of 100 estimates based on simulated data with $\theta = 0.03$ for an increasing number of databases with a fixed number of observations per database ($n_i = 200$ for all i). White boxes are MLE, grey boxes are MoM estimates, dark grey boxes are MM estimates, and the light grey boxes are posterior means. The • indicates the average of the estimates within each block.

to the standard MM implementation. However, the more advanced MM updating schemes are more efficient than the MLE for small database counts. We tested the same algorithms on larger datasets (Danish and Greenlandic forensic databases of 20,000 and 2,000 DNA profiles). For these larger databases, the MLE implementation was 10 times faster than the specialised MM algorithms and up to 1,000 times faster than the standard MM implementation. However, this is only true when using the observed FIM, $\mathcal{J}(\boldsymbol{\gamma})$, while the computation of the expected FIM, $\mathcal{I}(\boldsymbol{\gamma})$, is very slow even for databases of moderate size.

Profile log-likelihood

From the box plots in Figure 2.2, there seems to be a tendency for the MLE to underestimate the θ-parameter. In order to investigate the reason for this behaviour and compute the confidence intervals for θ, we derived the profile log-likelihood, $\hat{\ell}(\theta) = \max_{\boldsymbol{\pi}} \ell(\boldsymbol{\pi}, \theta; \boldsymbol{x})$, for θ. That is, fixing θ at some value $\tilde{\theta}$ and finding the maximum likelihood value under this constraint. By fixing θ at $\tilde{\theta}$ we also fix γ_+ at $\tilde{\gamma}_+ = (1-\tilde{\theta})/\tilde{\theta}$. Hence, we are maximising the regular log-likelihood under the constraint that $\gamma_+ = \tilde{\gamma}_+$.

Since the analytical form of the log-likelihood is complicated, the only way to evaluate the profile log-likelihood is by numerical methods as for the maximum likelihood estimation. Applying a Lagrange multiplier, λ, we need to find the stationary points of $\tilde{\ell}(\boldsymbol{\gamma}) = \ell(\boldsymbol{\gamma}; \boldsymbol{x}) + \lambda(\tilde{\gamma}_+ - \gamma_+)$. The partial derivatives yield

$$\frac{\partial \tilde{\ell}(\boldsymbol{\gamma})}{\partial \gamma_i} = \frac{\partial \ell(\boldsymbol{\gamma}; \boldsymbol{x})}{\partial \gamma_i} - \lambda; \qquad \frac{\partial \tilde{\ell}(\boldsymbol{\gamma})}{\partial \lambda} = \tilde{\gamma}_+ - \gamma_+$$

which implies that the score function for this new system is $\boldsymbol{u}(\boldsymbol{\gamma}, \lambda) = (\boldsymbol{u}(\boldsymbol{\gamma}) - \lambda \mathbf{1}_k, \tilde{\gamma}_+ - \gamma_+)^\top$, where $\boldsymbol{u}(\boldsymbol{\gamma})$ is the score function from (2.10) and $\mathbf{1}_k$ is a k-dimensional vector of ones. The observed FIM, $\mathcal{J}(\boldsymbol{\gamma}, \lambda)$, is also almost preserved from the likelihood equations,

$$\mathcal{J}(\boldsymbol{\gamma}, \lambda) = \begin{bmatrix} \mathcal{J}(\boldsymbol{\gamma}) & -\mathbf{1}_k \\ -\mathbf{1}_k^\top & 0 \end{bmatrix},$$

where $\mathcal{J}(\boldsymbol{\gamma})$ is the observed FIM from Section 2.3.1. Hence, we may apply Newton-Raphson iterations in order to maximise $\ell(\boldsymbol{\gamma}; \boldsymbol{x})$ under the constraint $\gamma_+ = \tilde{\gamma}_+$. Alternatively this constrained optimisation problem could have been solved using (recursive) quadratic programming. However, for this particular log-likelihood function Newton-Raphson procedure works very well with Lagrange multipliers, and the existing code for maximisation is easily extended for handling the extra terms induced by the constraints.

In Figure 2.3 the profile log-likelihood for simulated data with $\theta = 0.03$ is plotted. Each panel is standardised such that the maximum value of $\hat{\ell}(\theta)$ is zero, $2[\hat{\ell}(\theta) - \hat{\ell}(\hat{\theta})]$. The intersection of the dotted line and the profile log-likelihood indicates a 95%-confidence interval for θ based on a χ^2_1-approximation of $-2\hat{\ell}(\theta)$. In each panel the associated MLE (marked by $\bullet$), MoM ($\circ$) and MM ($\triangle$) estimates are plotted together with the true θ-value ($\times$). In all six panels the true value is included in the confidence intervals. As one would expect, the width of the confidence intervals decreases as the number of datasets increases. There are profound arguments for using the χ^2-approximation of partial maximised log-likelihood as opposed to using asymptotic results relying on approximative normality of the MLE with a covariance matrix asymptotically equal to the inverse FIM (Barndorff-Nielsen and Cox, 1994).

From Figure 2.3, it is evident that the profile log-likelihood is skew for small numbers of databases. This pronounced departure from symmetry explains the bias of the MLE and MM estimate for small numbers of databases. Using a Bayesian approach, one may assume a uniform prior on θ. This implies that the posterior distribution of θ approximately equals the profile likelihood, $p(\theta|\boldsymbol{x}) \propto \exp[\hat{\ell}(\theta)]$. The posterior mean, $\mathbb{E}(\theta|\boldsymbol{x})$, may be evaluated using a numerical approximation,

$$\mathbb{E}(\theta|\boldsymbol{x}) = \int \theta p(\theta|\boldsymbol{x}) \mathrm{d}\theta \approx \frac{\sum_{i=1}^n \tilde{\theta}_i \exp[\hat{\ell}(\tilde{\theta}_i)]}{\sum_{i=1}^n \exp[\hat{\ell}(\tilde{\theta}_i)]}. \tag{2.13}$$

The n different θ-values used in the sums of (2.13) are the same as those used for computing the profile log-likelihood, e.g. equidistant points covering the 95%-confidence interval. Table 2.1 lists the posterior means and estimates for the data in Figure 2.3, where the data points used for computing each posterior mean lies within the 95%-confidence interval.

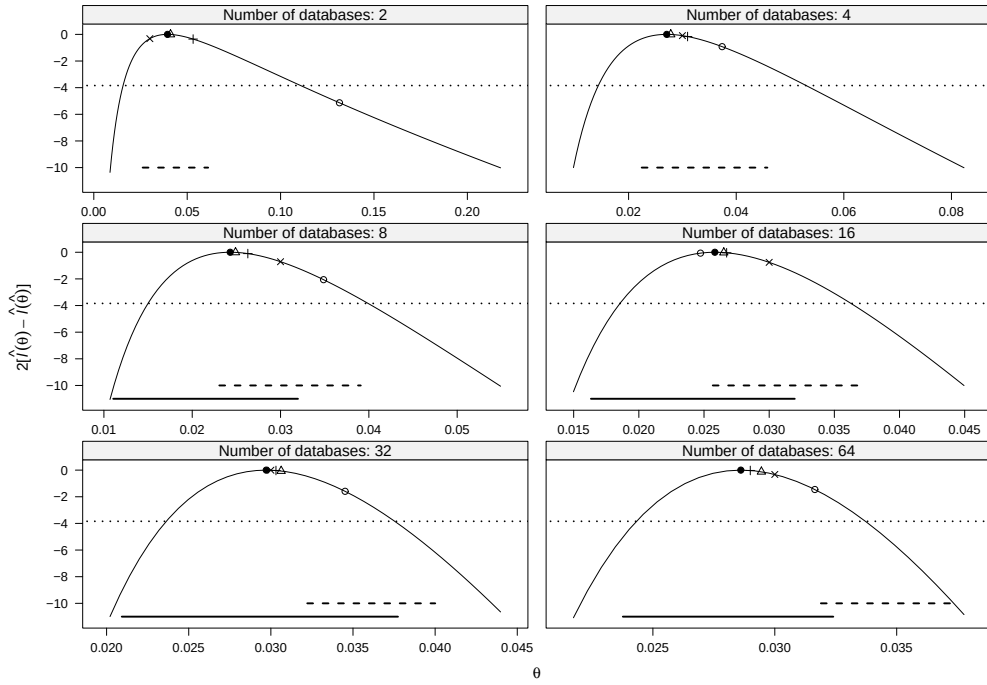

Figure 2.3: Profile log-likelihoods for simulated data for an increasing number of databases with $\theta = 0.03$ for all simulations (marked by $\times$). The MLE ($\bullet$), MoM ($\circ$), MM ($\triangle$) and posterior mean (+) are plotted together with a 95%-confidence interval (intersection of the dotted line and the profile log-likelihood curve). The horizontal dashed and solid lines represent bootstrap confidence intervals based on randomisation and cluster resampling, respectively.

Table 2.1: Posterior means and estimates for the data in Figure 2.3 ($\theta = 0.03$).

Number of databases	MLE	MoM	MM	Posterior mean
2	0.0395	0.1315	0.0411	0.0532
4	0.0271	0.0374	0.0278	0.0310
8	0.0243	0.0349	0.0249	0.0263
16	0.0258	0.0247	0.0265	0.0267
32	0.0297	0.0345	0.0306	0.0303
64	0.0286	0.0316	0.0295	0.0290

We see that the posterior mean estimate in most situations improves the MLE estimate (except for the first row) and reduces the amount of bias for small numbers of databases. In Figure 2.2 the light grey boxes (rightmost box whiskers for each stratum) represent the posterior means for the simulated data computed using θ-values within the 95%-confidence interval for the associated MLE. Table 2.1 indicates that the bias is reduced for the posterior means, with only a minor increment in the variance (see Figure 2.2).

A full Bayesian implementation with prior distributions on $\boldsymbol{\pi}$ and θ (or equivalently on $\boldsymbol{\gamma}$-parameters) was not pursued in this study. However, several authors (see e.g. Holsinger, 1999) have discussed estimation of θ (and other population genetics diversity measures) from a Bayesian perspective. We refer to the review paper by Holsinger and Weir (2009) for further results and discussions on Bayesian methodologies.

Bootstrapping confidence intervals

In addition to computing a confidence interval for θ using the χ^2_1-approximation of the profile log-likelihood, we also investigated the performance of bootstrap methods to construct the confidence intervals. However, there are some problems when bootstrapping clustered data in order to assess the variability of the intra-cluster correlation parameter θ.

Several studies (Davison and Hinkley, 1997; Ukoumunne et al., 2003; Fields and Welsh, 2007) indicate that special attention needs to be paid when one applies the bootstrap methodology to this problem. The general recommendation is to sample on a subpopulation (cluster) level rather than an individual (randomised) level due to the dependence structure implied by the intra-cluster correlation factor. In Figure 2.3, we have superimposed bootstrap confidence intervals (horizontal solid and dashed lines) based on both kinds of bootstrap regime. The general picture is that the cluster sampling underestimates θ (solid line - missing in first two panels due to few databases), whereas the randomised bootstrap provides overestimated values (dashed line).

From numerical studies we recommend the use of the profile log-likelihood method in order to estimate the confidence intervals for θ since this method is valid for any number of subpopulations in the data. This might not be surprising (Davison and Hinkley, 1997; Ukoumunne et al., 2003; Fields and Welsh, 2007). However, bootstrapping is often applied when assessing the variability of estimates but for θ this is inappropriate.

Significance test

Testing whether θ satisfies certain numerical properties is interesting since equality of θ across loci simplifies *PI* and *LR* computations. Further simplifications are possible if $\theta = 0$ is supported by data. This implies that there is no detectable difference among the databases, where the reasons for this may be small sample sizes (and thus large variation), or that the databases are as if sampled from a homogeneous population.

Samanta et al. (2009, Section 3: Hypothesis testing) derived hypothesis tests for inference about θ under various population assumptions. Here we initiate by testing for equality of θ_s for the various loci, $s = 1, \ldots, S$. The null hypothesis is $\theta_1 = \cdots = \theta_S = \theta'$ for some unknown θ',

where θ_s is the θ-value for locus s, with the alternative hypothesis specifying that at least one θ_s is different from θ'. In order to test this hypothesis, we evaluate

$$\ell(\{\boldsymbol{\pi}_s\}_{s=1}^S, \theta'; \boldsymbol{x}) = \sum_{s=1}^{S} \ell_s(\boldsymbol{\pi}_s, \theta'; \boldsymbol{x}_s) \tag{2.14}$$

where ℓ_s is the regular log-likelihood in (2.8) with $\theta_s = \theta'$ for all s. The test statistics is given by

$$-2\log Q = -2\left[\ell(\{\hat{\boldsymbol{\pi}}_s\}_{s=1}^S, \theta'; \boldsymbol{x}) - \sum_{s=1}^{S} \ell_s(\hat{\boldsymbol{\pi}}_s, \hat{\theta}_s; \boldsymbol{x}_s)\right],$$

and is approximately χ^2_{S-1}-distributed from the $S-1$ degrees of freedom (DoF). Details of finding stationary points of (2.14) are given in Appendix 2.A.2.

Furthermore, testing whether $\theta = 0$ is another interesting hypothesis test. Under the null hypothesis there is no evident substructure in the data. Having support for $\theta = 0$ implies that DNA profiles may be regarded as independent, which has a high influence on the estimation of the evidential weight (see Sections 2.2.2 and 2.2.3). The Dirichlet-multinomial model with $\theta = 0$ is equivalent to the simpler multinomial model. However, testing the hypothesis that $\theta = 0$ can not be based on asymptotic theory nor inferred from the inclusion/exclusion of zero in the confidence intervals from the profile log-likelihood since $\theta = 0$ lies on the boundary of the parameter space.

A possible method is to use a parametric bootstrap, where we simulate data $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_M^*$ under the null hypothesis $\theta = 0$. From these simulated data we estimate $\hat{\theta}_m^*$ and obtain an approximative distribution of $\hat{\theta}$ under the null hypothesis, which we apply in order to test the significance of $\theta \neq 0$ for the observed data, $\boldsymbol{x}$. Hence, the parametric bootstrap comprises two steps: (1) draw $\boldsymbol{x}_m^* \sim \text{Multinomial}(\{\boldsymbol{x}_{i+}\}_{i=1}^N, \{\boldsymbol{x}_{+j}\}_{j=1}^k / n_+)$ and (2) estimate $\hat{\theta}_m^*$.

By choosing M large, e.g. $M = 1000$, one gets M estimates of θ of which most should have an estimate smaller than $\hat{\theta}$ when the hypothesis $\theta = 0$ is false. An empirical p-value is computed by $\#\{\hat{\theta}_m^* > \hat{\theta}\}/M$, i.e. the ratio of the number of larger parametric bootstrap estimates to the total number of bootstraps.

2.4 Results

The paper of Budowle and Moretti (1999) presents allele frequencies of 13 CODIS Core STR loci in six US subpopulations. The data have previously been used to estimate the magnitude of θ used for forensic purposes; see e.g. Weir (2007). Henceforth we refer to these data as "FBI data".

Estimates of θ based on the MoM, MLE and MM are given in Table 2.2. There are some distinct differences between the $\hat{\theta}_{\text{MLE}}$ and $\tilde{\theta}_{\text{MoM}}$, with often a factor two in difference; furthermore , the standard errors are often very much smaller for the MLE than for the MoM estimates. The standard errors are asymptotic, where $SE(\tilde{\theta})$ is based on a Taylor series approximation by Li

Table 2.2: Locus-specific estimates of θ based on MLE, MoM, MM and posterior mean (PM). The confidence interval for the MLE is based on the χ^2_1-approximation of the profile log-likelihood.

Locus	$\tilde{\theta}_{\text{MoM}}$	$SE(\tilde{\theta})$	$\hat{\theta}_{\text{MLE}}$	$SE(\hat{\theta})$	95%-CI for $\hat{\theta}$	$\tilde{\theta}_{\text{MM}}$	PM
D3	0.0108	0.0085	0.0056	0.0020	(0.0028; 0.0110)	0.0057	0.0061
vWA	0.0107	0.0085	0.0053	0.0017	(0.0027; 0.0098)	0.0053	0.0056
FGA	0.0050	0.0051	0.0037	0.0010	(0.0021; 0.0061)	0.0037	0.0038
D8	0.0140	0.0106	0.0084	0.0024	(0.0049; 0.0145)	0.0085	0.0089
D21	0.0126	0.0097	0.0053	0.0013	(0.0031; 0.0086)	0.0053	0.0055
D18	0.0142	0.0107	0.0086	0.0019	(0.0056; 0.0133)	0.0087	0.0089
D5	0.0226	0.0157	0.0161	0.0042	(0.0097; 0.0276)	0.0163	0.0170
D13	0.0264	0.0180	0.0147	0.0040	(0.0088; 0.0254)	0.0149	0.0156
D7	0.0061	0.0056	0.0035	0.0013	(0.0015; 0.0072)	0.0036	0.0038
CSF	0.0050	0.0049	0.0091	0.0026	(0.0049; 0.0167)	0.0092	0.0097
TPOX	0.0306	0.0205	0.0248	0.0066	(0.0147; 0.0433)	0.0254	0.0263
TH01	0.0328	0.0217	0.0189	0.0054	(0.0110; 0.0340)	0.0193	0.0202
D16	0.0117	0.0091	0.0069	0.0023	(0.0036; 0.0131)	0.0070	0.0074

(Weir and Hill, 2002, pp. 730), and $SE(\hat{\theta}) = \{(\mathcal{I}^{-1})_{\theta,\theta}\}^{1/2}$ from Section 2.3.1. Standard errors of the MM estimates are not readily obtained from the Matlab scripts of the supplementary material of Zhou and Lange (2010), hence these are not provided in Table 2.2. The ratio $\tilde{\theta}_{\text{MoM}}/\hat{\theta}_{\text{MLE}}$ of the estimates in Table 2.2 repeats the pattern which was indicated by the plots in Figures 2.2 and 2.3. For most loci, the MoM estimate lies within the 95%-confidence interval. The MM estimates coincide with the MLE for all loci. The posterior means are for most loci close to the MLE, which is due to the rather symmetric shape of the profile log-likelihoods plotted in Figure 2.4, where the profile log-likelihoods for the FBI data are plotted together with the MLE (marked by $\bullet$), MoM ($\circ$), MM ($\triangle$) and posterior mean (+).

We tested the hypothesis of equality of θ for all loci in the FBI data. From Table 2.2, it is clear that there are differences among loci, but also some clustering of the estimates. In Table 2.3, we have listed the results from testing different hypotheses.

Table 2.3: Results from testing hypothesis of equality of θ for multiple loci.

Loci	$\overline{\theta}$	θ'	-2 log Q	DoF	p-value
All	0.0101	0.0090	62.8011	12	<0.0001
D5, D13, TPOX and TH01	0.0186	0.0183	2.2630	3	0.5196
Remaining loci	0.0063	0.0061	12.7175	8	0.1219

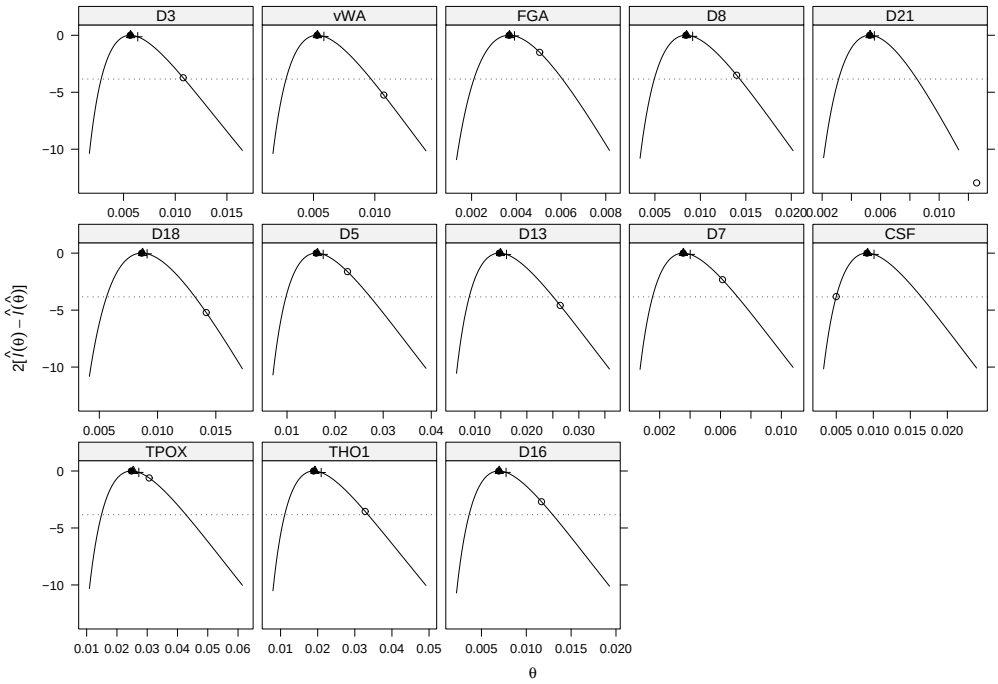

Figure 2.4: Profile log-likelihoods for the 13 CODIS loci from the FBI data of Budowle and Moretti (1999). The MLE is marked by •, MoM by ∘ and MM by △. For all loci the MLE and MM estimate coincide. For most loci, the MoM estimate lies within the MLE confidence interval. The + indicates the posterior mean.

The tests indicate that there are groups of loci with similar θ-values. The mean, $\bar{\theta}$, of the four loci (D5, D13, TPOX and TH01) with the largest θ-estimates in Table 2.2 is $\bar{\theta} = 0.0186$, and the mean of the remaining loci is $\bar{\theta} = 0.0063$. In both groups, the estimated $\theta' = 0.0183$ (95%-CI: $[0.0126; 0.0269]$) and $\theta' = 0.0061$ (95%-CI: $[0.0043; 0.0088]$) is almost equal to $\bar{\theta}$ (Table 2.3).

Furthermore, using the methodology described in Section 2.3.1 for testing if $\theta = 0$, the test yielded that for no loci was θ equal to zero. This was true for both the MLE and MoM estimates. However, the test based on the MLE is more powerful than using the MoM estimate. The estimated θ-values for the Caribbean subsample (subset of FBI data: Bahamian, Jamaican and Trinidad subpopulations) are given in Table 2.4 together with empirical p-values and 95%-confidence intervals under the null hypothesis, $H_0 : \theta = 0$.

In locus D8, the tests based on MLE rejects the null hypothesis, whereas the MoM test accepts that $\theta = 0$. For this locus (and D3, D21, D7, CSF, D16) the MoM estimate is negative, and so is the lower bound of the confidence intervals for all loci.

Conceptually, we could imagine that we only had observed a common Caribbean database with-

Table 2.4: θ-estimates for Caribbean sample (three databases from the FBI data) together with empirical p-values and 95%-confidence intervals when $\theta = 0$. For each locus, the first row is MLE and the second row MoM estimates. Note that for locus D8 the test based on MLE rejects the hypothesis while the MoM-based test does not.

Locus	θ-estimate	p-value	95%-Confidence interval	H_0-decision
D3	2.582×10^{-11}	0.099	$(3.975 \times 10^{-12}; 7.243 \times 10^{-11})$	Accept
	-1.440×10^{-3}	0.745	$(-2.755 \times 10^{-3}; 5.715 \times 10^{-3})$	Accept
vWA	3.904×10^{-4}	0.000	$(3.966 \times 10^{-12}; 9.639 \times 10^{-11})$	Reject
	3.488×10^{-3}	0.030	$(-2.226 \times 10^{-3}; 3.656 \times 10^{-3})$	Reject
FGA	3.944×10^{-12}	0.507	$(1.772 \times 10^{-12}; 1.470 \times 10^{-11})$	Accept
	6.496×10^{-4}	0.284	$(-2.021 \times 10^{-3}; 3.221 \times 10^{-3})$	Accept
D8	4.351×10^{-9}	0.010	$(3.973 \times 10^{-12}; 8.016 \times 10^{-11})$	**Reject**
	1.286×10^{-3}	0.214	$(-2.581 \times 10^{-3}; 4.404 \times 10^{-3})$	**Accept**
D21	3.510×10^{-12}	0.680	$(2.654 \times 10^{-12}; 1.711 \times 10^{-11})$	Accept
	-4.964×10^{-4}	0.567	$(-2.196 \times 10^{-3}; 3.710 \times 10^{-3})$	Accept
D18	6.262×10^{-4}	0.000	$(2.655 \times 10^{-12}; 2.230 \times 10^{-11})$	Reject
	6.657×10^{-3}	0.001	$(-2.066 \times 10^{-3}; 3.058 \times 10^{-3})$	Reject
D5	3.367×10^{-3}	0.000	$(3.964 \times 10^{-12}; 5.651 \times 10^{-11})$	Reject
	8.452×10^{-3}	0.000	$(-2.314 \times 10^{-3}; 4.449 \times 10^{-3})$	Reject
D13	1.405×10^{-11}	0.197	$(3.962 \times 10^{-12}; 9.433 \times 10^{-11})$	Accept
	3.693×10^{-3}	0.060	$(-2.449 \times 10^{-3}; 5.637 \times 10^{-3})$	Accept
D7	4.776×10^{-12}	0.712	$(3.964 \times 10^{-12}; 5.566 \times 10^{-11})$	Accept
	-1.062×10^{-3}	0.727	$(-2.337 \times 10^{-3}; 4.600 \times 10^{-3})$	Accept
CSF	4.399×10^{-12}	163	$(3.971 \times 10^{-12}; 8.619 \times 10^{-11})$	Accept
	-2.049×10^{-3}	0.924	$(-2.494 \times 10^{-3}; 4.102 \times 10^{-3})$	Accept
TPOX	1.478×10^{-3}	0.000	$(5.950 \times 10^{-12}; 2.149 \times 10^{-10})$	Reject
	7.890×10^{-3}	0.002	$(-2.533 \times 10^{-3}; 4.135 \times 10^{-3})$	Reject
TH01	8.026×10^{-12}	0.735	$(5.949 \times 10^{-12}; 1.429 \times 10^{-10})$	Accept
	6.922×10^{-4}	0.295	$(-2.700 \times 10^{-3}; 4.298 \times 10^{-3})$	Accept
D16	7.002×10^{-12}	0.704	$(3.976 \times 10^{-12}; 1.371 \times 10^{-10})$	Accept
	-1.925×10^{-3}	0.903	$(-2.403 \times 10^{-3}; 4.100 \times 10^{-3})$	Accept

out information on the specific island of origin. Thus for loci with $\theta = 0$ (see Table 2.4) this collapse of the observed databases would in principle not be a problem. However, for the other loci the present substructure would potentially cause the LR to be anti-conservative depending on a particular suspect's DNA profile and origin.

In Table 2.5, the estimated allele probabilities (π_j, for appropriate subscript j) are presented for each locus. Note that the estimated allele probabilities are estimates of the allele probabilities in the reference population from which each of the six subpopulations is assumed to have descended. Owing to lack of space, only alleles with integer values are presented, i.e. common alleles such as 9.3 in TH01 are not reported in Table 2.5.

2.5 Discussion

The model based on the Dirichlet-multinomial distribution has previously been discussed, for example, by Lange (1995b). However, the estimation methods suggested there relied on approximations of the trigamma-function, which were avoided here due to similar results as those of Paul et al. (2005).

The maximum likelihood estimation of parameters discussed in this paper is much more involved than those of the method of moment (MoM). However, the properties of the MLE ensure reduced variance of the estimates. In general, the $\boldsymbol{\pi}$-estimates based on MoM and MLE did coincide, indicating that the usual relative frequency estimate is adequate in order to obtain point estimates for the allele probabilities. However, as pointed out by Curran et al. (2002), the uncertainty of these point estimates needs to be carefully considered when assessing the weight of the evidence. If allele probabilities are estimated from limited databases the estimates of the rare alleles are subject to large standard errors. This may lead to overestimates of the (point estimates of) *LR* or *PI*.

Having a joint model for the allele probabilities and θ-parameter increases the belief in the estimates of the latter. However, since $\boldsymbol{\pi}$ may be estimated by the empirical probability $\bar{\boldsymbol{p}}$, the simpler one-dimensional maximisation problem $\max_\theta \ell(\theta, \bar{\boldsymbol{p}}; \boldsymbol{x})$ may be adequate for estimating θ and assessing its variance. Simulations have shown that this method underestimates θ even for large number of databases; hence this estimator is inefficient as opposed to the joint likelihood approach, which therefore is recommended for estimation.

Balding (2003, pp. 229) argues that one should expect variability of θ across the STR loci used in forensic genetics. This may be due to different mutation rates in the various loci and selection or "indirect selection" from linkage between the STR loci and genes/genetic regions subject to selection.

It is possible to test the hypothesis of equal θ across loci using our model. For the FBI data there were two groups of loci with common θ-estimates. Figure 2.1 showed that increased θ-values weaken the evidence in most cases. Hence, for a conservative evaluation of the evidence, it may be reasonable to use the largest θ-value. This supports the use of the upper 95%-confidence limit (see Table 2.2) of the θ-estimate, which in most cases does not disagree with the commonly used value 0.03 for θ (Phillips et al., 2010). Furthermore, Balding (2005, pp. 97) argues that "plug-in values (of θ) should tend to be towards the higher end of the range of plausible values" in order to incorporate uncertainty from higher-order terms of θ.

However, in paternity disputes it is not common practice to evaluate the evidence conservatively since in most circumstances these are civil lawsuits. Hence, in paternity cases it may be more ap-

Table 2.5: Estimates of $\boldsymbol{\pi}$ for each locus. The first line of each cell entry gives the MLE and the second line is the MoM estimate. In small font the associated standard errors ×10. Only integer-valued alleles are reported, for compactness. *: Add 10 units to each FGA allele designation. **: Add 20 units to each D21 allele designation.

Allele	Locus D3	vWA	FGA*	D8	D21**	D18	D5	D13	D7	CSF	TPOX	TH01	D16
5												.0063 .037	
												.0019 .023	
6					.0009 .009				.0024 .015	.0011 .011	.0389 .119	.1761 .227	
					.0005 .011				.0018 .018	.0005 .013	.0482 .419	.1734 .491	
7			.0008 .008		.0499 .080		.0066 .036	.0027 .021	.0129 .036	.0290 .072	.0227 .086	.3285 .282	
			.0005 .011		.0501 .226		.0131 .244	.0009 .014	.0122 .065	.0332 .292	.0164 .116	.3341 .901	
8			.0070 .027	.0055 .028	.1852 .145		.0254 .080	.0501 .113	.1626 .120	.0327 .077	.4161 .333	.1769 .227	.0344 .072
			.0089 .117	.0050 .068	.1892 .709		.0337 .294	.0492 .303	.1626 .394	.0425 .342	.4308 1.075	.1752 .556	.0310 .162
9			.0634 .081	.0078 .034	.1903 .146		.0279 .085	.0598 .124	.1202 .106	.0290 .072	.1525 .238	.1549 .215	.1568 .147
			.0625 .098	.0055 .038	.1882 .117		.0244 .173	.0746 .751	.1234 .396	.0271 .146	.1618 .801	.1453 .362	.1596 .582
10			.0819 .092	.0449 .088	.2098 .152	.0067 .031	.0755 .144	.0527 .116	.3228 .152	.2626 .197	.0726 .167	.0143 .060	.1135 .128
			.0833 .337	.0526 .389	.2165 .640	.0055 .053	.0661 .400	.0507 .279	.3216 .205	.2593 .156	.0645 .254	.0079 .062	.1136 .352
11		.0054 .025	.1202 .109	.0538 .097	.0760 .098	.0142 .048	.3017 .257	.2725 .240	.2234 .136	.2428 .192	.2439 .288		.2996 .186
		.0037 .033	.1220 .333	.0501 .167	.0750 .169	.0104 .076	.3055 .903	.2683 .437	.2230 .357	.2435 .367	.2334 .475		.2981 .163
12			.1841 .130	.1307 .147	.0180 .048	.0773 .116	.3536 .268	.3592 .260	.1244 .107	.3155 .208	.0514 .139		.2245 .169
	.0020 .014		.1845 .265	.1260 .190	.0154 .081	.0790 .338	.3529 .382	.3687 1.005	.1261 .373	.3178 .431	.0444 .256		.2324 .706
	.0014 .021		.1601 .123	.2524 .190	.0046 .023	.0806 .119	.1823 .215	.1442 .188	.0246 .050	.0686 .111	.0019 .019		.1447 .142
13	.0060 .027	.0086 .032	.1592 .253	.2584 .639	.0035 .034	.0890 .536	.1893 .667	.1370 .199	.0239 .093	.0636 .104	.0005 .010		.1427 .224
	.0062 .067	.0084 .098	.1627 .124	.2818 .197	.0064 .027	.0926 .127	.0172 .064	.0575 .121	.0052 .022	.0138 .048			.0236 .059
14	.0914 .109	.0768 .098	.1597 .350	.2822 .570	.0055 .071	.1029 .603	.0108 .073	.0502 .177	.0045 .032	.0098 .030			.0211 .114
	.0947 .338	.0749 .162	.1047 .102	.1677 .163	.0136 .041	.1504 .157	.0069 .037	.0014 .014		.0038 .023			.0028 .018
15	.3212 .177	.1533 .133	.1047 .228	.1672 .473	.0174 .163	.1461 .139	.0033 .028	.0005 .011		.0023 .024			.0014 .015
	.3235 .598	.1610 .633	.0417 .066	.0448 .088	.0026 .017	.1603 .162	.0014 .015						
16	.2931 .173	.2791 .166	.0432 .222	.0446 .248	.0030 .044	.1561 .431	.0005 .025						
	.2897 .395	.2804 .512	.0309 .057	.0096 .039	.0020 .014	.1502 .157							
17	.1908 .149	.2177 .153	.0317 .174	.0079 .079	.0015 .023	.1566 .518							
	.1893 .322	.2144 .375	.0092 .031	.0011 .011		.0967 .129							
18	.0832 .104	.1678 .138	.0089 .069	.0005 .011		.0979 .345							
	.0856 .438	.1685 .381	.0047 .021			.0707 .112							
19	.0078 .031	.0692 .093	.0040 .035			.0686 .287							
	.0067 .050	.0679 .139	.0008 .008			.0481 .092							
20		.0210 .051	.0005 .011			.0467 .248							
		.0192 .071	.0008 .008			.0228 .063							
21		.0011 .010	.0005 .026			.0204 .134							
		.0014 .025				.0110 .042							
22						.0099 .096							
						.0046 .025							
23						.0035 .034							
						.0011 .011							
24						.0005 .011							

propriate to use the locus-specific MLEs (or the common θ-values for groups of loci in Table 2.3) when computing $PI(\theta)$.

2.6 Conclusion

We have demonstrated how the genetic dependence caused by identical-by-descent assumption can be modelled as overdispersion from a statistical point of view. This allowed for maximum likelihood estimation of allele probabilities in the reference population, $\boldsymbol{\pi}$, and the identical-by-descent measure, θ. By using recent results from the statistical literature the FIM was computed analytically and confidence intervals based on profile log-likelihoods were provided.

Acknowledgements

I would like to thank my PhD supervisor Associate Professor Poul Svante Eriksen (Aalborg University, Denmark), Professor Niels Morling (University of Copenhagen, Denmark) and Professor Bruce S. Weir (University of Washington, USA) for comments and valuable discussions. I am thankful to Professor Weir for inviting me as visiting scientist to The Department of Biostatistics, University of Washington, which I was visiting while working on this paper. Furthermore, I would like to thank Associate Professor Esben Høg (Aalborg University, Denmark) and three anonymous reviewers for their comments, which have significantly improved the final version of this paper.

Appendix

2.A Mathematical details

In Appendix 2.A.1, we give some mathematical details on how to derive the paternity index, $PI(\theta)$, of (2.5), and Appendix 2.A.2 is about testing for equality of θ across loci.

2.A.1 Deriving paternity index (PI)

We demonstrate how to derive the paternity index, PI, in (2.5) using $P(Y_{n+1}=j|\boldsymbol{Y}^n=\boldsymbol{y}^n) = P(j|x_j^n)$ in (2.4). In a given locus, the child's profile is (ac) and the mother is heterozygous (ab), where c is different from a and b. Discarding the possibility of mutations, the true father needs to pass on a c allele to the child. Assume that the alleged father is heterozygous (cd), which implies $P(ac|ab,cd) = \frac{1}{4}$, i.e. under hypothesis H_1 the probability of the child's profile given its parents' profiles is $\frac{1}{4}$. The PI is determined by:

$$PI = \frac{P(ac,ab,cd|H_1)}{P(ac,ab,cd|H_2)} = \frac{P(ac|ab,cd)P(ab,cd)}{\sum_{i,j}^{k} P(ac|ab,cd,ij)P(ab,cd,ij)},$$

where (ij) denotes the profile of the true father under H_2 and summation is over all k alleles in the given locus. However, when omitting the possibility of mutations, unless i or j equals c the child can not be the true father's offspring, i.e. $P(ac|ab, cd, ij) = 0$ for (i, j) where $c \neq i$ and $c \neq j$. Hence, we fix $j = c$ and sum over all $i = 1, \ldots, k$, where $P(ac|ab, cd, cc) = \frac{1}{2}$ and $P(ac|ab, cd, ic) = \frac{1}{4}$ for all $i \neq c$ under H_2. This implies that the expression for the PI is given by

$$\begin{aligned} PI &= \frac{\frac{1}{4}P(ab, cd)}{\frac{1}{2}P(ab, cd, cc) + 2\sum_{i\neq c}\frac{1}{4}P(ab, cd, ic)} \\ &= \frac{P(ab, cd)}{2P(ab, cd, c)\Big[P(c|ab, cd, c) + \sum_{i\neq c} P(i|ab, cd, c)\Big]} = \frac{1}{2P(c|ab, cd)}, \end{aligned}$$

where the sum in square brackets by definition is one. Using the expression $P(j|x_j^n)$ in (2.4) with $\boldsymbol{x}^4 = (x_a^4, x_b^4, x_c^4, x_d^4) = (1, 1, 1, 1)$, we have,

$$PI = \frac{1}{2P(c|x_c^4)} = \frac{1 + (n-1)\theta}{2[x_c\theta + (1-\theta)\pi_c]} = \frac{1 + 3\theta}{2[\theta + (1-\theta)\pi_c]}.$$

2.A.2 Testing equality of θ for multiple loci

In order to find stationary points for the log-likelihood of (2.14), we use Fisher-scoring with Lagrange multipliers, $\boldsymbol{\lambda} = \{\lambda_s\}_{s=1}^S$, ensuring equal θ for all loci. Translating the common parameter θ' to γ' ensures computational simplicity. The observed FIM, $\mathcal{J}(\gamma)$, associated with (2.14) is

$$\mathcal{J}(\gamma) = \begin{bmatrix} [\mathcal{J}(\gamma_1)] & \boldsymbol{O}_{1,2} & \cdots & \boldsymbol{O}_{1,S} & \boldsymbol{g}_{k_1} \\ \boldsymbol{O}_{2,1} & [\mathcal{J}(\gamma_2)] & \cdots & \boldsymbol{O}_{2,S} & \boldsymbol{g}_{k_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{O}_{S,1} & \cdots & \boldsymbol{O}_{S,S-1} & [\mathcal{J}(\gamma_2)] & \boldsymbol{g}_{k_S} \\ \boldsymbol{g}_{k_1}^\top & \boldsymbol{g}_{k_2}^\top & \cdots & \boldsymbol{g}_{k_S}^\top & 0 \end{bmatrix}$$

where $\boldsymbol{O}_{s,t}$ is a $(k_s{+}1) \times (k_t{+}1)$-matrix of zeros,

$$[\mathcal{J}(\gamma_s)] = \begin{bmatrix} \mathcal{J}(\gamma_s) & -\mathbf{1}_{k_s} \\ -\mathbf{1}_{k_s}^\top & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{g}_{k_s} = \begin{pmatrix} \mathbf{0}_{k_s} \\ 1 \end{pmatrix}$$

Furthermore, the score function is

$$\boldsymbol{u}(\{\gamma_s, \lambda_s\}_{s=1}^S, \gamma') = (\{\boldsymbol{u}(\gamma_s) - \lambda_s\mathbf{1}_{k_s}, (\gamma' - \gamma_{s+})\}_{s=1}^S, \lambda_+)$$

where $\boldsymbol{u}(\gamma_s)$ is the score function of (2.10).

Bibliography

Balding, D. J. (2003). Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology 63*, 221–230.

Balding, D. J. (2005). *Weight-of-evidence for Forensic DNA Profiles*. Chichester, West Sussex: John Wiley & Sons, Ltd.

Balding, D. J. and R. A. Nichols (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica 96*, 3–12.

Balding, D. J. and R. A. Nichols (1997). Significant genetic correlations among caucasians at forensic DNA loci. *Heredity 78*(6), 583–589.

Barndorff-Nielsen, O. E. and D. R. Cox (1994). *Inference and Asymptotics*. Number 52 in Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Box, G. E. P. and N. R. Draper (1987). *Empirical model-builing and response surfaces*. Wiley.

Budowle, B. and T. R. Moretti (1999). Genotype profiles for six population groups at the 13 CODIS short tandem repeat core loci and other PCR-based loci. *Forensic Science Communications*.

Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution 23*(1), 72–84.

Cockerham, C. C. (1973). Analysis of gene frequencies. *Genetics 74*(4), 679–700.

Curran, J. M., J. S. Buckleton, C. M. Triggs, and B. S. Weir (2002). Assessing uncertainty in DNA evidence caused by sampling effects. *Science and Justice 42*(1), 29–37.

Curran, J. M., C. M. Triggs, J. S. Buckleton, and B. S. Weir (1999). Interpreting DNA mixtures in structured populations. *Journal of Forensic Science 44*(5), 987–995.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Evett, I. W. and B. S. Weir (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sunderland, MA: Sinauer Associates.

Fields, C. A. and A. H. Welsh (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society. Series B, Statistical methodology 69*(3), 369–390.

Green, P. J. and J. Mortera (2009). Sensitivity of inferences in forensic genetics to assumptions about founding genes. *Annals of Applied Statistics 3*(2), 731–763.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science 28*(706), 49–50.

Holsinger, K. E. (1999). Analysis of genetic diversity in geographically structure populations: A bayesian perspective. *Hereditas 130*, 245–255.

Holsinger, K. E. and B. S. Weir (2009). Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews. Genetics 10*(9), 639–650.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1997). *Discrete Multivariate Distributions*. Wiley.

Lange, K. (1995a). Applications of the Dirichlet distribution to forensic match probabilities. *Genetica 96*, 107–117.

Lange, K. (1995b). *Mathematical and Statistical Methods for Genetic Analysis* (2 ed.). Springer.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika 49*(1-2), 65–82.

Neerchal, N. K. and J. G. Morel (2005). An improved method for the computation of maximum likelihood estimates for multinomial overdispersion models. *Computational Statistics & Data Analysis 49*, 33–43.

Nichols, R. A. and D. J. Balding (1991). Effects of population structure on DNA fingerprint analysis in forensic science. *Heredity 66*, 297–302.

Paul, S. R., U. Balasooriya, and T. Banerjee (2005). Fisher information matrix for the Dirichlet-multinomial distribution. *Biometrical Journal 47*(2), 230–236.

Phillips, C., T. Tvedebrink, et al. (2010). Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel. *Forensic Science International: Genetics*. In Press.

Rannala, B. and J. A. Hartigan (1996). Estimating gene flow in island populations. *Genetical Research 67*, 147–158.

Samanta, S., Y.-J. Li, and B. S. Weir (2009). Drawing inferences about the coancestry coefficient. *Theoretial Population Biology 75*, 312–319.

Tvedebrink, T. (2009). *dirmult: Estimation in Dirichlet-Multinomial distribution*. R package version 0.1.3.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2010). Evaluating the weight of evidence using quantitative STR data in DNA mixtures. *Journal of the Royal Statistical Society. Series C, Applied statistics*. In Press.

Ukoumunne, O. C., A. C. Davison, M. C. Gulliford, and S. Chinn (2003). Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. *Statistics in Medicine 22*, 3805–3821.

Weinberg, W. (1908). Über den nachweis der vererbung beim menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg 64*, 368–382.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sinauer Associates, Inc.

Weir, B. S. (2007). The rarity of DNA profiles. *The Annals of Applied Statistics 1*(2), 358–370.

Weir, B. S. and C. C. Cockerham (1984). Estimating *F*-statistics for the Analysis of Population Structure. *Evolution 38*(6), 1358–1370.

Weir, B. S. and W. G. Hill (2002). Esimating F-statistics. *Annual Review of Genetics 36*, 721–750.

Wright, S. (1951). The genetical structure of populations. *Annals of eugenics 15*, 323–354.

Zhou, H. and K. Lange (2010). MM algorithms for some discrete multivariate distributions. *Journal of Computational and Graphical Statistics*. In Press.

2.7 Supplementary remarks

The methodology presented above for estimating θ and computing the profile log-likelihood has been applied in the publication by Phillips et al. (2010). I performed some of the computations of that paper using the `dirmult` package and made plots similar to those of Figures 2.3 and 2.4 (Fig. 4 in Phillips et al., 2010). Plots in higher resolution are available from my web page (`http://people.math.aau.dk/~tvede` under "List of publication").

In population genetics the Hardy-Weinberg equilibrium (HWE) constitute a fundamental point of reference. Proposed independently by Hardy (1908) and Weinberg (1908), the HWE states that assuming random mating, no selection, no mutations and infinite population size the probability of a diploid genotype is the product of allele probabilities, $P(A_iA_j) = 2p_ip_j$ and $P(A_iA_i) = p_i^2$. We know immediate from these assumptions that HWE fail to hold since no real world population satisfy these restrictions. However, quoting Box and Draper (1987, pp. 74): "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful" applies also to HWE. In fact testing for HWE is often done to test for data quality, where the test is performed on genetic data to detect possible over-representation of homozygotes due to typing errors.

Over the last 100 years since the publication of the Hardy-Weinberg principle several genetic models have been proposed to relax the assumptions mentioned above. One such attempt were Wright (1951) who defined the F-statistics (F_{ST}, F_{IT} and F_{SI}), which are measures of population differentiation (Holsinger and Weir, 2009). In forensic genetics the most interesting of the parameters is F_{ST} which measures the divergence between a subpopulation, S, and the total population, T. Cockerham (1969, 1973) showed that for most interesting assumptions made about the population structure and breeding patterns θ is identical to F_{ST} (Weir and Cockerham, 1984, pp. 1358). The use of the θ-correction alters the genotype probabilities $P(A_iA_j) = 2p_ip_j(1-\theta)$ and $P(A_iA_i) = p_i\theta + p_i^2(1-\theta)$, where the magnitude of θ controls the deviation from HWE.

In the following chapter, a paper discussing the θ-correction in relation to a DNA reference profile databases is presented. In that setting only one database is available, hence there is no point of reference to which extend a particular subsampled database differs in allelic constitution from another. Therefore different means of estimating θ needs to be considered. In the setting above θ was a measure of subpopulation structure in a larger database, whereas in the subsequent setting θ is a measure of correlation between gametes (within and between individuals). Hence, by making pairwise comparisons of all individuals in the database we may be able to quantify θ by analysing the difference between expected and observed counts of matching loci.

CHAPTER 3

Analysis of matches and partial-matches in Danish DNA reference profile database

Publication details

Co-authors: Poul Svante Eriksen[*], James Curran[†], Helle Smidt Mogensen[‡] and Niels Morling[‡]

[*] *Department of Mathematical Sciences*
Aalborg University
[†] *Department of Statistics*
University of Auckland
[‡] *Section of Forensic Genetics, Department of Forensic Medicine*
Faculty of Health Science, University of Copenhagen

Journal: Forensic Science International: Genetics (Under preparation)

Abstract:
In this paper we analyse the Danish reference database accumulated over approximately 40 years with 51,517 DNA profiles, which is close to 1% of the Danish adult population size. Each entry in the database is associated with a civil registration number such that twins are identified and potential near matches due to typing errors are removed.

We investigated the methodology of Weir (2004, 2007), and extensions by Curran et al. (2007) to allow for close relatives, who derived expressions for the expected number of matches and near matches in a database when every DNA profile is compared to all other profiles in the database. We extended the methodology by computing the covariance matrix of the summary statistic and used it to estimate the identical-by-descent parameter θ for the Danish database.

Keywords:
DNA database; θ-correction; Subpopulation; Close relatives; Covariance matrix.

3.1 Introduction

In order to accommodate the pressure from the legal community, Weir (2007) commented on the rarity of DNA profiles and in particular on the number of expected profile matches and near profile matches one should expect as the DNA databases increase in size. The fact that a pair of profiles matches at 9 out of 13 loci in an Arizonian database of 65,493 profiles (Troyer et al., 2001) is not unexpected. In fact Weir (2007) suggests that 163 of such pairs would be expected under his population genetic model with the coancestry parameter $\theta = 0.03$. However, if one compares the expected counts and observed counts in Weir (2004), it is evident that the expected number of partially-matching loci is much larger than what is observed. A possible explanation is that the population is subdivided which increases the number of homozygote profiles. That is, profiles that are homozygous are either similar or different, which is not captured in the model discussed by Weir (2004, 2007).

Mueller (2008) investigated the performance of simple population genetic models further. He also focused on the Arizona database and discussed how likely it was to observe the reported 122 pairs matching on 9 loci and 20 pairs matching on 10 loci out of 13 loci. By means of simulations he increased the complexity of the model to include five ethnic groups each with four possible subpopulations and a number of relatives. He concluded that in order to obtain sufficiently high probabilities for the observed counts, there needed to be between 1,000 and 3,000 pairs of full-siblings in a substructured population. Several other authors have discussed multi-locus matching and population structures influence on match probabilities, e.g. Lange (1993, 1995); Donnelly (1995b,a); Balding and Nichols (1995); Ayres (2000); Laurie and Weir (2003); Song and Slatkin (2007).

The main focus of this paper is the examination and validation of the model proposed by Weir (2004, 2007) and the modifications hereof by Curran et al. (2007) to allow for closely related individuals in the database. To this purpose we model and analyse the distribution of matches and partial-matches in the reference DNA database at the Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen.

3.2 Materials and methods

3.2.1 Data

The Danish reference DNA profile database contains 51,517 STR DNA profiles accumulated from 1971 to the beginning of 2009 typed at the 10 autosomal loci included in the SGM Plus kit (Applied Biosystems, CA, USA). The database constitute little more than 1% of the Danish adult population (approx 4 million people). Each entry in the database is associated with a civil registration number such that twins are identified and potential near matches due to typing errors are removed.

The database were analysed such that every profile were compared to any other profile in the database. For each pairwise comparison the number of matching (agreement on both alleles), m, and partially-matching loci (sharing exactly one allele), p were registered. Let G_i and G_j be two DNA profiles in the database. Then $M(G_i, G_j)$ is a 11×11-indicator matrix with zeros except for the (m, p)-entry corresponding to m and p matching and partially matching loci between profile G_i and G_j, respectively.

Hence, the summary statistic $M = \{M_{m/p}\}_{m,p}$ is formed by

$$M = \sum_{i=1}^{n-1} \sum_{j>i}^{n} M(G_i, G_j), \tag{3.1}$$

which corresponds to $N = \binom{n}{2} = n(n-1)/2$ pairwise comparisons of n DNA profiles. With the database size of $n = 51{,}517$ this results in $N = 1{,}326{,}974{,}886$ comparisons.

The result of analysing the Danish database with $n = 51{,}517$ DNA profiles is summarised in Table 3.1, where $M_{m/p}$ corresponds to the number of pairs with m matching loci and p partially-matching loci. From Table 3.1 we find that e.g. the number of pairs of profiles with 5 matching loci and 4 partially-matching loci out of ten autosomal loci is $M_{5/4} = 17{,}060$. Figure 3.1 shows a the summary statistic in an informative way where we have plotted the observed counts on $\log_{10}$-scale.

Two of the authors (T. Tvedebrink and J. Curran) implemented computationally efficient functions for constructing the M-table in the statistical software R (R Development Core Team, 2009). The `compare`-function from the `DNAtools`-package (Curran and Tvedebrink, 2010b) took less than 5 minutes to perform all 1,326,974,886 pairwise comparisons on a 2.50 GHz laptop computer. Most of the methodology in this paper has been implemented in the `DNAtools`-package together with specialised plotting functions. The package is described in more detail elsewhere (Curran and Tvedebrink, 2010a).

Table 3.1: Summary matrix M for the Danish reference DNA profile database with 51,517 DNA profiles. $M_{m/p}$ is the number pairs of profiles with m matching (where m is the row number) and p partially-matching (where p is the column number) loci. Owing to lack of space the font size is reduced for the least interesting part of the table (low number of matching loci).

M	0	1	2	3	4	5	6	7	8	9	10
0	906,881	8,707,969	37,632,872	96,157,037	160,570,778	182,820,115	143,627,613	76,852,119	26,786,782	5,486,572	501,671
1	1,100,493	9,484,061	36,229,766	80,292,877	113,733,413	106,635,954	66,164,365	26,183,818	5,992,415	604,900	
2	595,135	4,531,792	14,996,133	28,165,271	32,810,688	24,271,278	11,132,519	2,887,555	325,493		
3	188,146	1,237,733	3,467,281	5,353,738	4,913,791	2,683,854	805,798	103,305			
4	38,094	212,192	487,484	592,929	401,832	143,202	21,490				
5	5,114	23,490	42,459	37,933	17,060	3,100					
6	470	1,685	2,272	1,414	378						
7	26	96	91	64							
8	3	6	21								
9	0	0									
10	0										

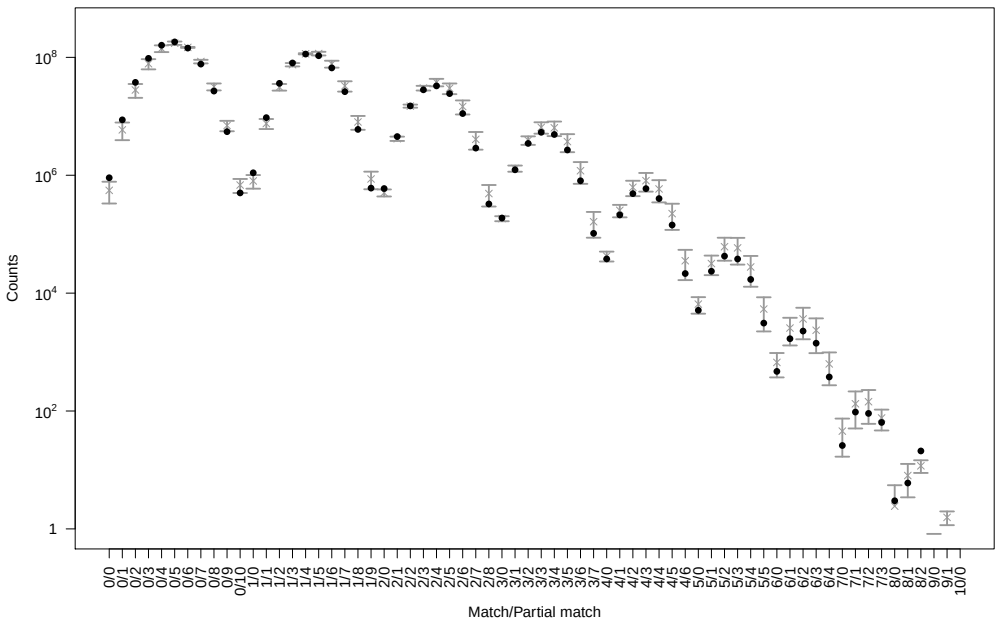

Figure 3.1: Plot of observed counts (marked by •) versus the number of matching and partially-matching loci (counts on $\log_{10}$-scale) for the Danish database. The superimposed points ($\times$) represents the expected counts (under the model described in Section 3.2.2) and the vertical bars indicate an approximative 95%-confidence interval computed by $N\boldsymbol{\pi} \pm 2\sqrt{\text{diag}\{\Sigma(\theta)\}}$ (see Sections 3.3 and 3.3.2).

3.2.2 Population genetic model

The model proposed by Weir (2007, 2004) defines for each of the L loci three probabilities $(P_{0/0}, P_{0/1}, P_{1/0})$, which are the probabilities for two randomly selected profiles sharing none, one or both alleles at a given locus (Weir denoted the probabilities P_0, P_1, P_2. The change of subscript will hopefully be clear in the following). The probabilities $P_{m/p}$ depends on the coancestry coefficient θ through the match probability equations (Nichols and Balding, 1991) that are derived using the recursion formula: $P(A_{i'}|\boldsymbol{x}^n) = [x^n_{i'}\theta + (1-\theta)p_{i'}]/[1+(n-1)\theta]$, which is the probability of observing an i' allele after having seen $x^n_{i'}$ alleles of type i' among n sampled alleles.

The expected values associated with the observed counts in M under this model is computed as $N\boldsymbol{\pi}$, where $\boldsymbol{\pi} = \{\pi_{m/p}\}_{m,p}$ is the matrix of probabilities for the match/partially-match events (m, p). The elements of $\boldsymbol{\pi}$, $\pi_{m/p}, m = 0, \ldots, L; p = 0, \ldots, L-m$, may be computed using recursion over loci: Let $\pi^{\ell}_{m/p}$ denote the probability based on ℓ loci, i.e. using only a subset of size ℓ of the L loci. Then the following equation denote how to compute $\pi^{\ell+1}_{m/p}$ for $\ell = 1, \ldots, L-1$:

$$\pi^{\ell+1}_{m/p} = P^{\ell+1}_{0/0}\pi^{\ell}_{m/p} + P^{\ell+1}_{0/1}\pi^{\ell}_{m/p-1} + P^{\ell+1}_{1/0}\pi^{\ell}_{m-1/p}, \tag{3.2}$$

where the "sum" of the subscripts for each term on the right hand side equals the subscript on the left hand side, and $P^{\ell}_{m/p}$ refer to the $P_{m/p}$ probabilities for the ℓth added locus. When either $m = 0$ and/or $p = 0$ we have these boundary equations:

$$\pi^{\ell+1}_{0/0} = P^{\ell+1}_{0/0}\pi^{\ell}_{0/0}, \quad \pi^{\ell+1}_{0/p} = P^{\ell+1}_{0/0}\pi^{\ell}_{0/p} + P^{\ell+1}_{0/1}\pi^{\ell}_{0/p-1} \quad \text{and} \quad \pi^{\ell+1}_{m/0} = P^{\ell+1}_{0/0}\pi^{\ell}_{m/0} + P^{\ell+1}_{1/0}\pi^{\ell}_{m-1/0},$$

where $\pi^1_{1/0} = P^1_{1/0}$, $\pi^1_{0/1} = P^1_{0/1}$ and $\pi^1_{0/0} = P^1_{0/0}$. These equations are easily implemented in computer software and efficiently compute the expected numbers for various θ-values.

Weir (2007) focused in his survey paper primarily on comparison between the observed counts and the expected number, $N\boldsymbol{\pi}(\theta)$, for different values of θ. However, as Curran et al. (2007) discussed one needs to consider normalisation of these differences for a proper comparison between the observed and expected counts. In this paper we show how to compute the covariance matrix of M in order to make a more rigorous comparison taking the correlation between cell counts into consideration.

Close relatedness

Weir (2007) showed that for a specified family relationship of a pairs of profiles, $P_{m/p}$ is updated using the probabilities, k_I, that the two individuals share I alleles identical-by-decent (IBD):

$$\tilde{P}_{0/0} = k_0 P_{0/0} \quad \tilde{P}_{0/1} = k_1(1-\theta)(1-S_2) + k_0 P_{0/1} \quad \text{and} \quad \tilde{P}_{1/0} = k_2 + k_1[\theta + (1-\theta)S_2] + k_0 P_{1/0},$$

where $S_2 = \sum_{i'=1}^{K} p^2_{i'}$ is the sum of squared allele probabilities at a given locus with K different alleles, and $\tilde{P}_{m/p}$ denote the probability that two individuals with the specified family relationship will match as m/p in a given locus. In order to compute $\tilde{\boldsymbol{\pi}}$, $P_{m/p}$ is replaced by $\tilde{P}_{m/p}$ in the (3.2). In Table 3.2 we have listed the five types of relatedness considered in this paper. The avuncular

class covers half-siblings, grandparent-grandchild and uncle-nephew (independent of gender) since these has identical $\boldsymbol{k}$-vector and are as such indistinguishable only using unlinked genetic markers.

Table 3.2: Probability of sharing I alleles IBD for the specified relationship (Weir, 2007, Table 4).

Relationship	Full-siblings	First-cousins	Parent-child	Avuncular	Unrelated
$\boldsymbol{k} = (k_2, k_1, k_0)$	(0.25, 0.5, 0.25)	(0, 0.25, 0.75)	(0, 1, 0)	(0, 0.5, 0.5)	(0, 0, 1)

The effect of these types of relatedness is represented graphically in Figure 3.2 where $\tilde{\pi}_{m/p}$ is plotted for the possible combinations of m and p for $\theta = 0.03$. Note that parent-child (marked by + in Figure 3.2) must share at least one allele per locus implying that $\tilde{\pi}_{m/p} = 0$ when $m + p \neq L$.

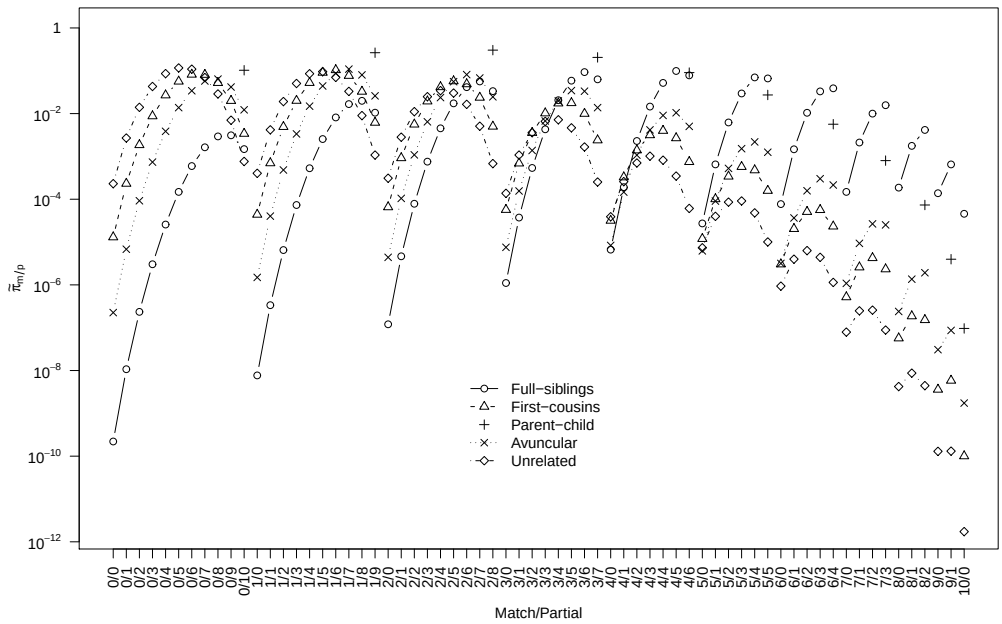

Figure 3.2: Effect on $\tilde{\boldsymbol{\pi}}$ for the five types of relatedness with $\theta = 0.03$. The legend explains the plot characters.

The inclusion of related pairs of profiles were investigated by Curran et al. (2007) using Australian data with Caucasian and Aborigine origin. Using that $E(\boldsymbol{\pi}) = E(E[\boldsymbol{\pi}|R]) = \sum_{r\in\mathcal{R}} E(\boldsymbol{\pi}|R = r)P(R = r)$ they computed expected number of matches by stratifying on close relationships, $\mathcal{R}$.

They formulated the model with $\mathcal{R}$ = {Full-siblings, First-cousins, Parent-child, Unrelated}:

$$\bar{\pi} = \alpha\tilde{\pi}^{\text{Full-siblings}} + \beta\tilde{\pi}^{\text{First-cousins}} + \delta\tilde{\pi}^{\text{Parent-child}} + \gamma\pi, \tag{3.3}$$

where $\gamma = 1-\alpha-\beta-\delta$ and the parameters refer to the fraction of the total comparisons that are made between pairs of full-siblings, first-cousins, parent-child and unrelated, respectively.

After fitting the model to the data, we have parameter estimates of the various parameters in the (3.3)-model. Thus we have an overall estimate of the probability that a random pair of profiles in the database has a certain familial relationship, e.g. the probability of two pairs of profiles originating from a pair of full-siblings in the Western Australia database is 6.91×10^{-6} (Curran et al., 2007, α-estimate in caption of Fig. 1).

These probabilities might be used in relation to crime cases where a suspect, S, declares that a close relative is the culprit, C. Let G_S be the suspect's profile (known to the investigator) and G_C the profile of the culprit (unknown, but may be identical to G_S). For some crime cases the defence may claim that the circumstances of the crime is such that the true offender is a close relative to S. Given a specific familial relationship, r, it is possible to compute the probability that S and C share the same DNA profile. We need to distinguish between the situation of G_S being heterozygous or homozygous, and let $P(G_C = A_{i'}A_{j'}|G_S = A_{i'}A_{j'}, R = r)$ and $P(G_C = A_{i'}A_{i'}|G_S = A_{i'}A_{i'}, R = r)$ denote these probabilities, where r is the specified familial relationship of C and S. Furthermore, the information about r, implies knowledge of $\boldsymbol{k}$ which gives these expression for the two probabilities:

$$\begin{aligned} P(G_C{=}A_{i'}A_{j'}|G_S{=}A_{i'}A_{j'}, R{=}r) &= k_2 + \frac{k_1}{2}\left[P(A_{i'}|A_{j'}, A_{i'}A_{j'}) + P(A_{j'}|A_{i'}, A_{i'}A_{j'})\right] + k_0 P(A_{i'}A_{j'}|A_{i'}A_{j'}) \\ &= k_2 + \frac{k_1}{2}\frac{2\theta+(1-\theta)(p_{i'}+p_{j'})}{1+2\theta} + 2k_0\frac{\theta^2+\theta(1-\theta)(p_{i'}+p_{j'})+(1-\theta)^2 p_{i'}p_{j'}}{(1+2\theta)(1+\theta)} \end{aligned} \tag{3.4}$$

$$\begin{aligned} P(G_C{=}A_{i'}A_{i'}|G_S{=}A_{i'}A_{i'}, R{=}r) &= k_2 + k_1 P(A_{i'}|A_{i'}, A_{i'}A_{i'}) + k_0 P(A_{i'}A_{i'}|A_{i'}A_{i'}) \\ &= k_2 + k_1\frac{3\theta+(1-\theta)p_{i'}}{1+2\theta} + k_0\frac{6\theta^2+5\theta(1-\theta)p_{i'}+(1-\theta)^2 p_{i'}^2}{(1+2\theta)(1+\theta)} \end{aligned} \tag{3.5}$$

If the suspect is not the true culprit, then the probability that $G_S \equiv G_C$ (share the same DNA profile) is given by $\tilde{\pi}_{10/0}$. For the five types of relatedness considered here, the probabilities are plotted in the right-most category in Figure 3.2 for $\theta = 0.03$.

3.3 Results

3.3.1 Simulations

We used the model discussed above to simulate DNA profile databases with known allele frequencies (the estimated allele frequencies from the Danish database) and various values for θ. For a specified number of DNA profiles, we used the recursive formula of Nichols and Balding

(1991) for individuals only remotely related $P(A_{i'}|\boldsymbol{x}^n) = [x^n_{i'}\theta + (1-\theta)p_{i'}]/[1+(n-1)\theta]$ to simulate alleles with a correlation governed by θ where $p_{i'}$ in the formula is the allele frequency of allele $A_{i'}$ and the vector $\boldsymbol{x}^n = (x^n_1, \ldots, x^n_K)$ is the sufficient summary statistic (Tvedebrink, 2010). In order to take close relationships among the individuals into consideration, we simulated the number of individuals with a specified relationship $\boldsymbol{n}_\mathrm{R} = (n_\mathrm{FS}, n_\mathrm{1C}, n_\mathrm{PC}, n_\mathrm{AV}, n - n_+)$, where all n_r are even numbers. The subscripts relates to full-siblings (FS), first-cousins (1C), parent-child (PC) and avuncular (AV). The last entry in $\boldsymbol{n}_\mathrm{R}$ refer to the remaining number of unrelated DNA profiles (UN). Since the comparisons $M(G_i, G_j)$ only considers pairs of profiles, the closely related DNA profiles are simulated in pairs such that:

1. Simulate the first relative R_1: $R_1 \sim P(A_{i'}A_{j'}|\boldsymbol{x}^n) = P(A_{i'}|\boldsymbol{x}^{n+1})P(A_{j'}|\boldsymbol{x}^n)$, where $\boldsymbol{x}^{n+1}=\boldsymbol{x}^n+\boldsymbol{e}_{j'}$ and $\boldsymbol{e}_{j'}$ is a vector of zeros except for a one in entry j'.
2. Simulate the number of alleles the second relative R_2 share IBD with R_1: $I \sim P(\boldsymbol{k})$.
3. Profile R_2 is simulated conditioned on the value of I:
   - $I = 0$: R_2 is generated unrelated to R_1: $R_2 \sim P(A_{k'}A_{l'}|A_{i'}A_{j'}, \boldsymbol{x}^n)$, and may be identical (by state) to R_1.
   - $I = 1$: The first allele of R_2 is drawn randomly from the alleles of R_1, e.g. $A_{i'}$ is sampled. The second allele is then sampled from $P(A_{k'}|A_{i'}, A_{i'}A_{j'}, \boldsymbol{x}^n)$.
   - $I = 2$: R_2 is identical to R_1. Note that only full-siblings has this possibility in our simulations.

By using this sampling scheme we make $n_r/2$ pairwise comparisons for relatedness on level r, since all other pairs of simulated relatives are mutually unrelated to each other. Hence, the known vector of $\boldsymbol{p}_r = \{P(R=r)\}_{r\in\mathcal{R}}$ is for each simulated database:

$$\boldsymbol{p}_r = \left(\frac{n_\mathrm{FS}}{n(n-1)}, \frac{n_\mathrm{1C}}{n(n-1)}, \frac{n_\mathrm{PC}}{n(n-1)}, \frac{n_\mathrm{AV}}{n(n-1)}, 1 - \frac{n_+}{n(n-1)}\right)$$

From the expressions above it is clear that for increasing database sizes the number of comparisons between relatives is $o(n^2)$. However the impact on M depends on the product of the matching probabilities and the fraction of comparisons, $\tilde{\boldsymbol{\pi}}_r p_r$. Mueller (2008) argued that the number of full-sibling pairs in the Arizonian database ($n = 65{,}493$) needed to be between 1,000 to 3,000 pairs. This gives that the fraction of pairwise comparisons attributed to full-siblings is between 4.73×10^{-7} and 1.42×10^{-6} for the Arizonian database.

In the formulation of Weir (2004, 2007) θ was assumed constant across loci. However, this need not to be the case due to different mutation rates, and possibly selection or *indirect* selection by linkage to other genes/markers subject to selection (Tvedebrink, 2010). In our simulations we used a constant θ across loci for simplicity. For each simulated database we estimated θ using five optimisation criteria:

$$C_1(\theta) = \sqrt{\sum (\boldsymbol{M} - N\boldsymbol{\pi}(\theta))} \qquad C_2(\theta) = \sum \frac{(\boldsymbol{M} - N\boldsymbol{\pi}(\theta))^2}{N\boldsymbol{\pi}(\theta)} \qquad C_3(\theta) = \sum \frac{|\boldsymbol{M} - N\boldsymbol{\pi}(\theta)|}{N\boldsymbol{\pi}(\theta)} \tag{3.6}$$

$$T_1(\theta) = \sum \frac{(\boldsymbol{M} - N\boldsymbol{\pi}(\theta))^2}{\mathrm{diag}\{\Sigma(\theta)\}} \qquad T_2(\theta) = \{\boldsymbol{M} - N\boldsymbol{\pi}(\theta)\}^\top \Sigma(\theta)^- \{\boldsymbol{M} - N\boldsymbol{\pi}(\theta)\}, \tag{3.7}$$

where summation is over the vector entries. The object functions in (3.6) were investigated by Curran et al. (2007) as a mean to compare the expected and observed counts. The authors argued

that numerical work indicated that $C_3(\theta)$ yielded good results since special emphasis is placed on the upper tail of the distribution (large number of matching loci). The functions in (3.7) uses the covariance matrix, $\Sigma(\theta)$, computed in this paper (cf. below). The first function, $T_1(\theta)$, does not take correlations into accounts, whereas $T_2(\theta)$ is a natural measure of similarity (a so called Mahalanobis-distance) incorporating the covariance matrix.

Let $\boldsymbol{M}$ be the M-matrix written in vector format (Appendix see 3.A for details on the transformation). We derived the expression for the variance of $\boldsymbol{M}$, $\Sigma(\theta)$, such that $T_2(\theta) = \{N\boldsymbol{\pi}(\theta) - \boldsymbol{M}\}^\top \Sigma(\theta)^- \{N\boldsymbol{\pi}(\theta) - \boldsymbol{M}\}$ may be compared for various values of θ in order to obtain the minimal $T_2(\theta)$. We use the generalised inverse of $\Sigma(\theta)$ since $\Sigma(\theta)$ is not of full rank due to the linear constraint $N = M_{+/+}$, where the "+"-notation indicates summation over the index. Let all the DNA profile identifiers, (i_1, i_2, i_3, i_4) be different, then the variance is computed as:

$$\Sigma(\theta) = \binom{n}{2}\mathbb{V}\left[\boldsymbol{M}(G_{i_1}, G_{i_2})\right] + 6\binom{n}{3}\mathbb{C}\left[\boldsymbol{M}(G_{i_1}, G_{i_2}), \boldsymbol{M}(G_{i_1}, G_{i_3})\right] + 6\binom{n}{4}\mathbb{C}\left[\boldsymbol{M}(G_{i_1}, G_{i_2}), \boldsymbol{M}(G_{i_3}, G_{i_4})\right], \tag{3.8}$$

where the covariances $\mathbb{C}\left[\boldsymbol{M}(G_{i_1}, G_{i_2}), \boldsymbol{M}(G_{i_1}, G_{i_3})\right]$ and $\mathbb{C}\left[\boldsymbol{M}(G_{i_1}, G_{i_2}), \boldsymbol{M}(G_{i_3}, G_{i_4})\right]$ are the most involved terms to compute since $\mathbb{V}\left[\boldsymbol{M}(G_{i_1}, G_{i_2})\right] = \text{diag}\{\boldsymbol{\pi}(\theta)\} - \boldsymbol{\pi}(\theta)\boldsymbol{\pi}(\theta)^\top$. The full details are given in Appendix 3.A.

Simulations of unrelated DNA profiles

We simulated 1, 000 databases for varying θ-values, $\theta \in \{0.00; 0.01; 0.02; 0.03; 0.04\}$ with 10, 000 DNA profiles per database. For each database we computed the summary statistic M and Figure 3.3 shows box-plots of the summary statistics on logarithmic scale for each m/p-category for θ=0.03. The superimposed vertical boxes (dark grey) represent an approximate 95%-confidence interval computed by $N\boldsymbol{\pi}(\theta) \pm 2\sqrt{\text{diag}\{\Sigma(\theta)\}}$, where the approximation rely on an approximation to normality for the counts. The performance of this approximation increases with the cell counts, i.e. the smaller the counts the less accurate is the approximation. The light grey boxes represent the 95% sample confidence interval based on the 2.5% and 97.5% quartiles in the distribution of the simulated values. Inserted is also the expected value ($\triangle$) for each category. It is evident that the median for most categories are identical to the expected value, except for cases with $N\pi_{m/p}$ small. Here, the box plot is of limited use since the observations are "all or nothing".

For each method the minimum was found by evaluating the function for θ_i on a fine grid of θ-values with step length 0.0001 for the interval [0, 0.12]. The box plot of Figure 3.4 compare the performance of the five measures of similarity between the observed and expected numbers for the various θ-values. In the box plot the known θ_0 is subtracted from the estimated $\hat{\theta}$ such that the box plot show the deviation of $\hat{\theta}$ from the true value.

The box plot of Figure 3.4 indicate that there is hardly no difference among the methods. However, the mean squared errors (MSE) in Table 3.3 show that the $T_2(\theta)$-method has a slightly better overall performance compared to the four other methods. Both the box plot and MSE show an increase in the deviation for increasing θ-values. This is due to the larger variability (from the higher correlation of the profiles) in the simulated data, and hence the available information for inference about θ decreases.

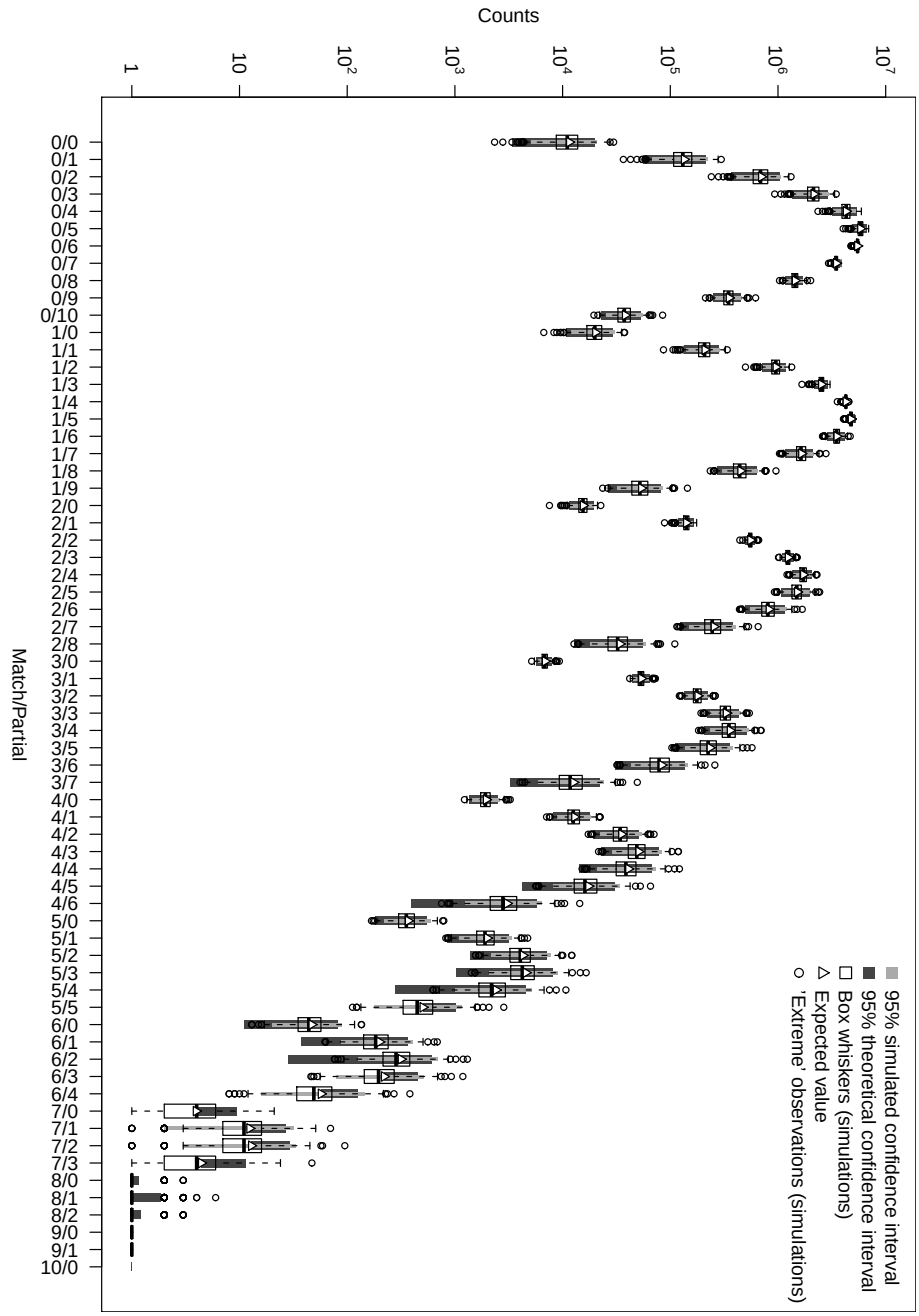

Figure 3.3: Box plots of the cell counts (on $\log_{10}$-scale) for the various categories for 1,000 simulated databases with 10,000 DNA profiles and $\theta = 0.03$. The legend explains the plot characters.

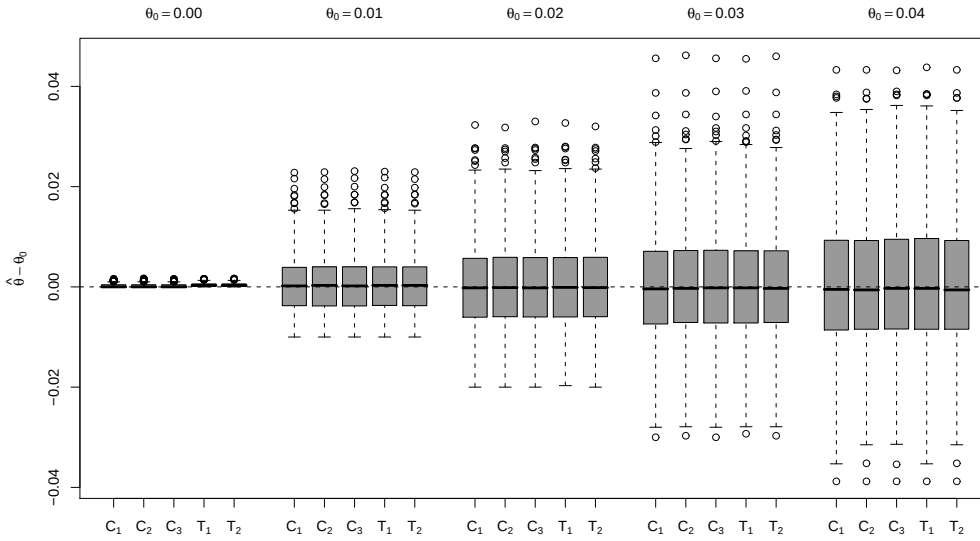

Figure 3.4: Comparisons of the performance of the object functions in (3.6) and (3.7).

Table 3.3: Mean square errors for the five different measures of similarity stratified on θ.

	$C_1(\theta)$	$C_2(\theta)$	$C_3(\theta)$	$T_1(\theta)$	$T_2(\theta)$
$\theta = 0.00$	1.072×10^{-7}	1.136×10^{-7}	1.078×10^{-7}	1.077×10^{-7}	1.205×10^{-7}
$\theta = 0.01$	3.432×10^{-5}	3.418×10^{-5}	3.457×10^{-5}	3.280×10^{-5}	3.264×10^{-5}
$\theta = 0.02$	7.509×10^{-5}	7.456×10^{-5}	7.601×10^{-5}	7.538×10^{-5}	7.460×10^{-5}
$\theta = 0.03$	1.213×10^{-4}	1.205×10^{-4}	1.231×10^{-4}	1.222×10^{-4}	1.208×10^{-4}
$\theta = 0.04$	1.711×10^{-4}	1.697×10^{-4}	1.730×10^{-4}	1.727×10^{-4}	1.702×10^{-4}
Overall	8.034×10^{-5}	7.977×10^{-5}	8.132×10^{-5}	8.061×10^{-5}	7.963×10^{-5}

Simulations including close relatives

The simulations in the previous section only considered remote relatedness trough allelic correlation governed by θ. However, most realistic reference DNA profile databases will contain DNA profiles from closely related individuals, e.g. brothers and father-son pairs. Hence, we also investigated the performance of the $C(\theta)$ and $T(\theta)$-functions for databases with pairs of close relatives. For each θ-value we simulated databases with the number of relatives as specified in Table 3.4.

Like in in the previous section we want to minimise the deviation between the observed and expected counts. However, for these simulations the expected value depend on θ and $\boldsymbol{p}_r$ through the expression: $\mathbb{E}(M;\theta,\boldsymbol{p}_r) = \sum_{r\in\mathcal{R}} P(R=r)\mathbb{E}(M|\theta;R=r) = \sum_{r\in\mathcal{R}} p_r N\tilde{\boldsymbol{\pi}}^r$, as discussed in relation to (3.3). Let $\tilde{C}(\theta)$ and $\tilde{T}(\theta)$ be as in (3.6) and (3.7), but with $N\boldsymbol{\pi}(\theta)$ replaced by $\sum_{r\in\mathcal{R}} p_r N\tilde{\boldsymbol{\pi}}^r$,

Table 3.4: The number of simulated relatives for the various θ-values with a total of 10,000 DNA profiles. The numbers in brackets are the relative frequency of pairwise comparisons between DNA profile with the specified relationship, i.e. the known $P(r)$-values.

Full-siblings	First-cousins	Parent-child	Avuncular	Unrelated
2,000 (2×10^{-5})	2,000 (2×10^{-5})	2,000 (2×10^{-5})	2,000 (2×10^{-5})	2,000 (0.99992)
5,000 (5×10^{-5})	1,000 (1×10^{-5})	1,000 (1×10^{-5})	1,000 (1×10^{-5})	2,000 (0.99992)
1,000 (1×10^{-5})	5,000 (5×10^{-5})	1,000 (1×10^{-5})	1,000 (1×10^{-5})	2,000 (0.99992)
1,000 (1×10^{-5})	1,000 (1×10^{-5})	5,000 (5×10^{-5})	1,000 (1×10^{-5})	2,000 (0.99992)
1,000 (1×10^{-5})	1,000 (1×10^{-5})	1,000 (1×10^{-5})	5,000 (5×10^{-5})	2,000 (0.99992)

then we seek $(\hat{\theta}, \hat{\boldsymbol{p}}_r) = \arg\min_{(\theta,\boldsymbol{p}_r)} \tilde{F}(\theta)$ for $\tilde{F}$ being either $\tilde{C}$ or $\tilde{T}$.

It should be noted that for consistency the variance of $\boldsymbol{M}$ should in this case be computed as $\tilde{\Sigma}(\theta) = \mathbb{E}(\mathbb{V}(\boldsymbol{M}|R)) + \mathbb{V}(\mathbb{E}(\boldsymbol{M}|R))$. However, we argue that the complexity and cost in computing $\tilde{\Sigma}(\theta)$ is far beyond the gain. Hence, when minimising with respect to $\tilde{T}_1(\theta)$ and $\tilde{T}_2(\theta)$ we use $\Sigma(\theta)$ in the computations.

The performance of the different optimisation measures is summarised in Figure 3.5 and Table 3.5. The pattern of larger variation of the θ-estimates for increasing θ_0 is repeated in the simulations with relatives. From Figure 3.5 there is a remarked spread in the estimates of P(First-cousins) for the $C_i(\theta)$-methods, $i = 1, 2, 3$. The MSE for $P(r)$ are generally smaller for $C_i(\theta)$ whereas $T_2(\theta)$ has smaller MSE for θ.

Assuming that the estimators of θ and $\boldsymbol{p}_r$ are unbiased, the expected values are given in Table 3.4 and the estimated variances in Table 3.5. Overall the mean is 10^{-5} while the standard errors are $\approx 10^{-4}$ indicating that not all parameters seem to be significant. Since the minimisation is computational intense, we dropped all close relationships but P(FS) and re-fitted the model. Naturally P(FS) overestimate the actual fraction of full-siblings since it needed to compensate for first-cousins, parent-child and avuncular. However, the estimate of P(Full-siblings) is for this reduced model significantly different from zero.

3.3.2 Danish database

The Danish reference DNA profile database was analysed using the described methods and gave the summary statistic presented in Table 3.1 and Figure 3.1. We have used the $T_2(\theta)$ method to estimate the θ and $\boldsymbol{p}_r$ for the Danish database. The minimum was obtained with $\hat{\theta} = 0.0107$ and $\hat{\boldsymbol{p}}_r$ as reported in Table 3.6. It is noteworthy that $\hat{\theta} = 0$ for all of the $C_i(\theta)$-methods. It seems rather unlikely that there is no effect of subpopulation after allowing for close relatives.

Note that the estimated P(Full-siblings) for $T_2(\theta)$ is about a factor 10 larger than 2×10^{-7} which is the approximate value obtained if one assumes that every individual of the Danish adult population has exactly one full-sibling. However, it is likely that the frequency of full-siblings is larger in the reference database than in the population due to various factors, e.g. the police's sampling

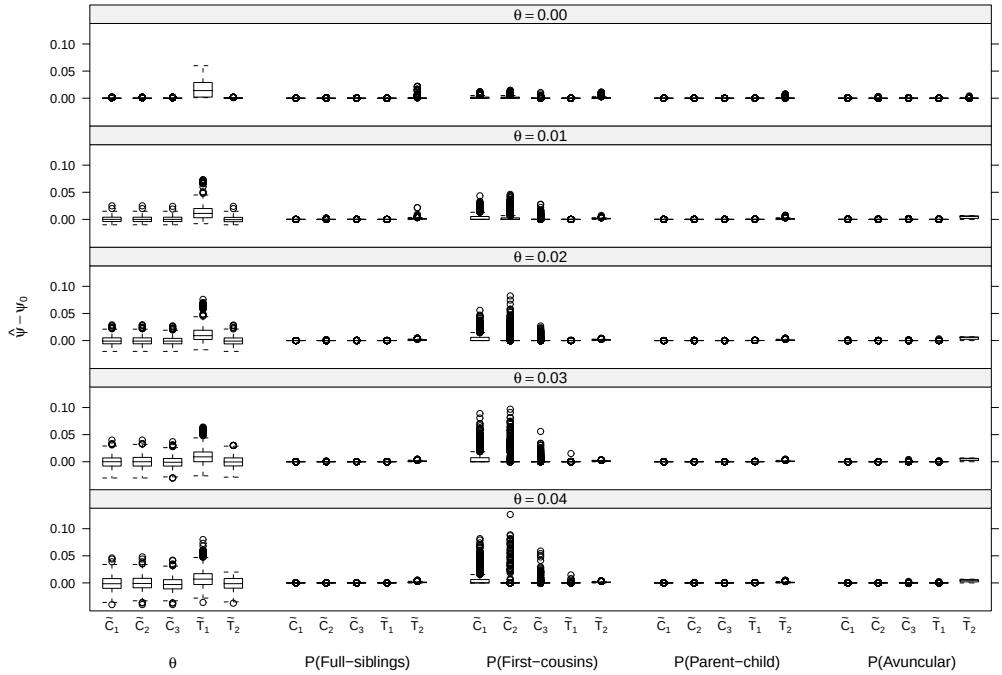

Figure 3.5: Box plot of the differences between ψ_0 and $\hat{\psi}$ (with ψ replaced for the relevant parameters) for various θ-values and number of relatives in the simulated databases.

criteria and social factors. Inserting these values in $\tilde{\pi}$ and $\Sigma(\theta)$ gives the expected values and covariance matrix, and given these quantities we computed marginal 95%-confidence intervals (superimposed in Figure 3.1).

The argument for using the θ-correction when assessing the evidential weight of a given DNA profile is to adjust for possible subpopulation effects in the population from which the suspect and profiles for estimating allele probabilities are drawn. A structured population causes the probability of observing a specific DNA profile to be heterogeneous, since the prevalence of its constituting alleles may be higher in some subpopulation relative to the entire population. Taking the argument further, one could argue that adjustment should be made for close relatedness between the suspect and "random man". Hence, when forming the likelihood ratio, *LR*, the hypothesis in the denominator could be H_d: "A man possibly related to the suspect is the true donor of the biological stain". The evaluation of $P(E|H_d)$ would then be a sum $\sum_{r\in\mathcal{R}} P(E|H_d, R = r)P(R = r)$, where $(H_d, R = r)$ concretises the specific relationship r between suspect and culprit.

Table 3.5: Mean squared errors (MSE) for various number of relatives stratified by θ-values.

θ	Parameter	$C_1(\theta)$	$C_2(\theta)$	$C_3(\theta)$	$T_1(\theta)$	$T_2(\theta)$
0.00	θ	1.354×10^{-7}	1.316×10^{-7}	1.539×10^{-7}	2.514×10^{-4}	1.202×10^{-7}
	P(FS)	1.008×10^{-11}	1.897×10^{-8}	1.683×10^{-9}	1.653×10^{-10}	5.238×10^{-6}
	P(1C)	4.813×10^{-6}	8.078×10^{-6}	5.762×10^{-7}	1.505×10^{-10}	3.630×10^{-6}
	P(PC)	6.835×10^{-11}	5.191×10^{-10}	1.362×10^{-10}	4.788×10^{-9}	7.962×10^{-7}
	P(AV)	1.835×10^{-8}	7.247×10^{-8}	6.895×10^{-9}	1.693×10^{-10}	7.217×10^{-8}
0.01	θ	3.221×10^{-5}	3.273×10^{-5}	3.049×10^{-5}	2.032×10^{-4}	2.984×10^{-5}
	P(FS)	2.878×10^{-11}	8.088×10^{-8}	8.669×10^{-10}	1.453×10^{-10}	1.995×10^{-6}
	P(1C)	3.825×10^{-5}	6.430×10^{-5}	7.402×10^{-6}	1.431×10^{-10}	6.560×10^{-7}
	P(PC)	1.708×10^{-10}	1.428×10^{-9}	2.053×10^{-10}	8.610×10^{-9}	1.052×10^{-6}
	P(AV)	7.710×10^{-9}	3.483×10^{-9}	1.590×10^{-8}	3.275×10^{-9}	6.465×10^{-6}
0.02	θ	7.006×10^{-5}	7.165×10^{-5}	6.542×10^{-5}	2.472×10^{-4}	6.853×10^{-5}
	P(FS)	4.029×10^{-11}	1.727×10^{-8}	1.131×10^{-9}	1.719×10^{-10}	6.694×10^{-7}
	P(1C)	7.252×10^{-5}	1.055×10^{-4}	9.885×10^{-6}	1.050×10^{-9}	3.598×10^{-7}
	P(PC)	1.976×10^{-10}	7.711×10^{-10}	2.924×10^{-10}	1.024×10^{-8}	5.935×10^{-7}
	P(AV)	4.547×10^{-9}	5.264×10^{-10}	1.201×10^{-8}	1.250×10^{-8}	5.160×10^{-6}
0.03	θ	1.232×10^{-4}	1.263×10^{-4}	1.153×10^{-4}	2.237×10^{-4}	1.213×10^{-4}
	P(FS)	5.841×10^{-11}	5.695×10^{-9}	7.739×10^{-10}	1.714×10^{-10}	7.768×10^{-7}
	P(1C)	1.688×10^{-4}	1.649×10^{-4}	1.854×10^{-5}	2.294×10^{-7}	3.728×10^{-7}
	P(PC)	2.294×10^{-10}	3.160×10^{-10}	3.701×10^{-10}	1.490×10^{-8}	7.187×10^{-7}
	P(AV)	3.361×10^{-10}	5.053×10^{-10}	5.101×10^{-8}	6.679×10^{-9}	4.757×10^{-6}
0.04	θ	1.661×10^{-4}	1.698×10^{-4}	1.532×10^{-4}	2.839×10^{-4}	1.463×10^{-4}
	P(FS)	8.469×10^{-11}	1.109×10^{-9}	1.195×10^{-9}	2.560×10^{-10}	1.063×10^{-6}
	P(1C)	1.886×10^{-4}	1.542×10^{-4}	2.180×10^{-5}	3.568×10^{-7}	5.585×10^{-7}
	P(PC)	2.318×10^{-10}	2.240×10^{-10}	4.195×10^{-10}	1.665×10^{-8}	1.028×10^{-6}
	P(AV)	5.348×10^{-10}	8.605×10^{-11}	1.799×10^{-8}	1.416×10^{-8}	5.296×10^{-6}

Table 3.6: Estimated values for the Danish database using various object functions.

Method	θ	P(Full-siblings)	P(First-cousins)	P(Parent-child)	P(Avuncular)
$C_1(\theta)$	0.0000	2.592×10^{-6}	8.413×10^{-9}	1.072×10^{-12}	1.930×10^{-9}
$C_2(\theta)$	0.0000	3.700×10^{-7}	5.100×10^{-7}	1.000×10^{-8}	4.600×10^{-7}
$C_3(\theta)$	0.0000	5.005×10^{-6}	3.534×10^{-7}	6.089×10^{-13}	2.475×10^{-7}
$T_1(\theta)$	0.0125	1.072×10^{-6}	4.573×10^{-8}	5.197×10^{-5}	5.930×10^{-9}
$T_2(\theta)$	0.0107	2.263×10^{-6}	1.757×10^{-7}	1.491×10^{-6}	5.882×10^{-9}

The problem of this approach would be to quantify $P(R = r)$ for a given suspect. One approach could be to take $\hat{p}_r$ as estimated from the database and then form a weighted sum in the denominator. By doing so for the Danish database with the estimated θ, frequencies for alleles and pairs of relatives we obtained LR and LR_r, where LR_r denotes the LR taking close relatives into account:

$$LR_r = \frac{P(E|H_p)}{P(E|H_d)} = \frac{P(E|H_p)}{\sum_{r\in\mathcal{R}} P(E|H_d, R = r)P(R = r)} = \frac{1}{\sum_{r\in\mathcal{R}} P(C|S, R = r)P(R = r)},$$

where $P(C|S, R = r)$ is computed by multiplying (3.4) and (3.5) over loci.

For each profile in the database we computed LR assuming that the profile was that of a suspect in single contributor crime case, i.e. $LR = 1/P(U|S)$ where $P(U|S)$ is the probability of observing an unknown profile (the defence hypothesis) given the suspect's profile. Similarly we computed LR_r under the same circumstances, except that the unknown profile may a close relative to S.

In Figure 3.6, we have plotted $\log_{10} LR_r$ against $\log_{10} LR$ and see that the relationship is close to linear: $\log_{10} LR_r = \beta + \alpha \log_{10} LR$. Estimating the parameters $(\hat{\alpha}, \hat{\beta}) = (0.115, 8.59)$ we obtain a simple formula to calculate LR_r from LR: $LR_r = 10^{8.59} LR^{0.115}$. In Figure 3.6, we have superimposed the predicted value (solid line) with the uncertainty represented by the predictive interval (dashed lines). The estimated mean and standard deviation of $\log_{10} LR/LR_r$ are respectively 3.128 and 0.97. Hence, an approximative confidence interval for the ratio is given as $10^{3.128\pm1.96\times0.97} = [27\,;106{,}955]$, i.e. taking close relatives into account decreases the LR with up to five orders of magnitude. The dominating contribution to the sum of $P(E|H_d)$ is that of full-siblings, $P(E|H_d, R = \text{FS})\hat{p}_{\text{FS}}$, which accounts for approximately 99.5% of LR_r. In Figure 3.2 this was also the category with the largest $\tilde{\pi}_{10/0}$. Hence, for practical purposes the only relevant type of close relatedness to include in LR_r is full-siblings since the decrease in $P(E|H_d, R)$ for the remaining types of relatives is minimal relative to $\hat{p}_r$. Furthermore, previous we saw that the model only including full-siblings and unrelated increased $P(\text{Full-siblings})$. Thus, this would decrease LR_r further yielding a more conservative evaluation of the evidence.

3.4 Discussion

It is evident from the analysis of the Danish reference DNA profile database that a θ-correction close to 1% is sufficient to capture the effects from substructure among the typed DNA profiles. Furthermore, did the analysis indicate the presence of close relatives in the database. A fact that were known beforehand, but the number of close relatives were unknown. However, the significance of the estimated probabilities, $\hat{\boldsymbol{p}}_r$, were not assessed implying some of them may be zero.

It is unknown whether it makes sense to present the LR_r in court since often the judge and jury are more interested in the LR for a specific relationship rather than a mean over common relationships with numerical impact on $P(E|H_d)$. However, LR_r may be used in order to accommodate for the fact that "the unrelated man" may in fact be a unknown close relative to the suspect.

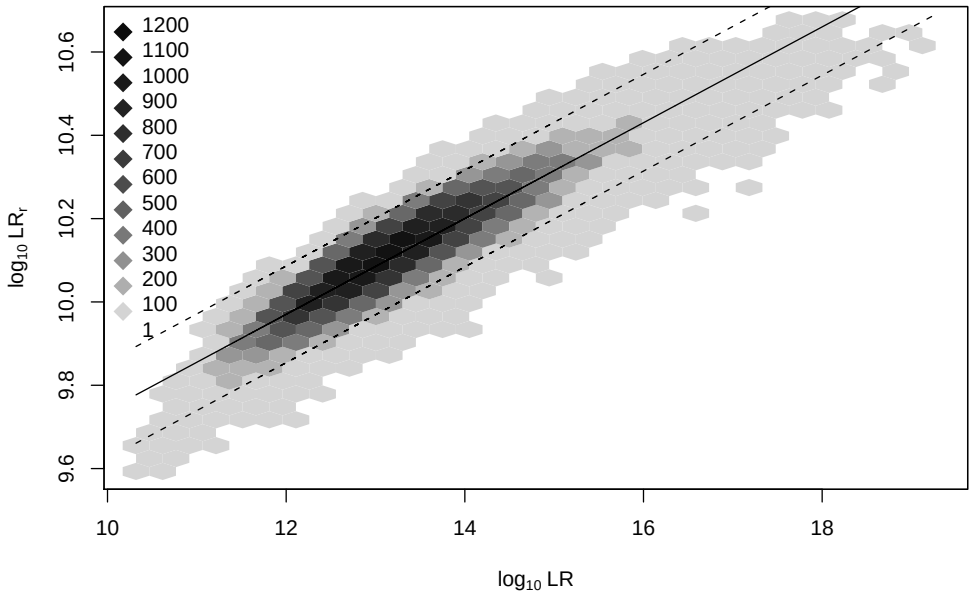

Figure 3.6: Relationship between LR and LR_r with a predictive interval superimposed (solid line: mean, dashed lines: predictive limits). The shaded hexagons indicate bin counts.

3.5 Conclusion

The main objective with the work presented in this paper were to analyse the Danish reference DNA profile database of 51,517 different individuals. This was to accommodate the fact that at some point two apparently unrelated individuals will share DNA profiles for all ten loci in the Danish population. If a specified relationship is determined it is straight forward to calculate the probability of identical DNA profiles, however, one still needs to account for remote coancestry for both related and unrelated pairs of profiles.

Furthermore, only modelling the expected value or calculating the mean is never satisfactory in statistics. A measure of precision or variability is needed in order to discuss the extremity of an observation relative to the expectation under a given model. Hence, deriving and computing the covariance matrix of M was essential. However, as the simulations exemplified that there was no pronounced improvement by using the Mahalanobis distance, $T_2(\theta) = [M - N\boldsymbol{\pi}(\theta)]^\top \Sigma(\theta)^- [M - N\boldsymbol{\pi}(\theta)]$, rather than the $C(\theta)$-functions for estimating θ.

Acknowledgements

The authors would like to thank Ms. Kirstine Kristensen and Ms. Line Maria Irlund Pedersen both from The Section of Forensic Genetics, University of Copenhagen) for their assistance in verifying the familial relationships of the twins in the database, and validating some near matches due to typing errors.

Appendix

3.A Derivation and computation of the variance

In order to compute the variance of the summary matrix, we use the definition of variance and covariance for random variables. First, note that $M(G_i, G_j)$ may be listed as a vector: $M(G_i, G_j) \to M(G_i, G_j)$, where the mapping operates on the m/p values: $f(m, p; L) = m[(L + 1) + (m - 1)/2] + (p + 1)$, where L is the total number of loci. Next, we expand the expression $\mathbb{V}(M) = \Sigma(\theta)$:

$$
\begin{aligned}
\Sigma(\theta) &= \mathbb{V}\left[\sum_{i=1}^{n-1}\sum_{j>i}^{n} M(G_i, G_j)\right] \\
&= \sum_{i=1}^{n-1}\sum_{j>i}^{n} \mathbb{V}\left[M(G_i, G_j)\right] + 6\sum_{i=1}^{n-2}\sum_{j>i}^{n-1}\sum_{k>j}^{n} \mathbb{C}\left[M(G_i, G_j), M(G_i, G_k)\right] \\
&\quad + \sum_{i=1}^{n-1}\sum_{j>i}^{n}\sum_{k\neq\{i,j\}}^{n-1}\sum_{\substack{l>k \\ l\neq\{i,j\}}}^{n} \mathbb{C}\left[M(G_i, G_j), M(G_k, G_l)\right] \\
&= \binom{n}{2}\mathbb{V}\left[M(G_{i_1}, G_{i_2})\right] + 6\binom{n}{3}\mathbb{C}\left[M(G_{i_1}, G_{i_2}), M(G_{i_1}, G_{i_3})\right] + 6\binom{n}{4}\mathbb{C}\left[M(G_{i_1}, G_{i_2}), M(G_{i_3}, G_{i_4})\right]
\end{aligned}
$$

where (i_1, i_2, i_3, i_4) in the last line relates to any of the DNA profiles in the database as long as they are different profiles. We go from the first to second line by expanding the sum and observe that $\mathbb{C}[M(G_i, G_j), M(G_i, G_k)] = \mathbb{C}[M(G_i, G_j), M(G_j, G_k)] = \mathbb{C}[M(G_i, G_k), M(G_j, G_k)]$ since $M(\cdot, \cdot)$ is symmetric. The sum over the last term in the expansion, $\mathbb{C}[M(G_i, G_j), M(G_k, G_l)]$ with all profile indexes different, also contain several symmetries implying the weights in the final expression. In order to compute the covariances , we need to compute

$$
\mathbb{E}\left[M(G_i, G_j)M(G_i, G_k)^\top\right] \qquad \text{and} \qquad \mathbb{E}\left[M(G_i, G_j)M(G_k, G_l)^\top\right],
$$

respectively, given that the DNA profile indexes i, j, k and l are all different.

For computing $\mathbb{E}\left[M(G_i, G_j)M(G_i, G_k)^\top\right]$ we need to account for the fact that profile G_i enters in both pairwise comparisons. Hence, we need to condition on G_i when deriving the probabilities $\pi_{m/p,\tilde{m}/\tilde{p}} = \sum_{i',j'} P(m/p, \tilde{m}/\tilde{p} | G_i = A_{i'}A_{j'})P(A_{i'}A_{j'})$ for all combinations of $m/p, \tilde{m}/\tilde{p}$, where m/p relates to the number of matches/partial-matches of G_i and G_j, with a similar definition of $\tilde{m}/\tilde{p}$ for profiles G_i and G_k.

As for the mean we use a recursion formula over loci to compute $\pi_{m/p,\tilde{m}/\tilde{p}}$. However, in this setting there are nine terms on the right hand side:

$$
\begin{aligned}
\pi^{\ell+1}_{m/p,\tilde{m}/\tilde{p}} = {} & \pi^{\ell}_{m/p,\tilde{m}/\tilde{p}} P^{\ell+1}_{0/0,0/0} + \pi^{\ell}_{m/p-1,\tilde{m}/\tilde{p}} P^{\ell+1}_{0/1,0/0} + \pi^{\ell}_{m-1/p,\tilde{m}/\tilde{p}} P^{\ell+1}_{1/0,0/0} + \\
& \pi^{\ell}_{m/p,\tilde{m}/\tilde{p}-1} P^{\ell+1}_{0/0,0/1} + \pi^{\ell}_{m/p,\tilde{m}-1/\tilde{p}} P^{\ell+1}_{0/0,1/0} + \pi^{\ell}_{m/p-1,\tilde{m}/\tilde{p}-1} P^{\ell+1}_{0/1,0/1} + \\
& \pi^{\ell}_{m-1/p,\tilde{m}/\tilde{p}-1} P^{\ell+1}_{1/0,0/1} + \pi^{\ell}_{m/p-1,\tilde{m}-1/\tilde{p}} P^{\ell+1}_{0/1,1/0} + \pi^{\ell}_{m-1/p,\tilde{m}-1/\tilde{p}} P^{\ell+1}_{1/0,1/0}.
\end{aligned}
$$

When one or more of the subscripts are zero there are similar boundary conditions for $\pi_{m/p,\tilde{m}/\tilde{p}}$ as those specified in Section 3.2.2. The probabilities $P_{m/p,\tilde{m}/\tilde{p}}$ are found by considering the events separately. For each configuration of $(m/p, \tilde{m}/\tilde{p}) \in \{(x_0/y_0, x_1/y_1) : (x_i, y_i) \in \{0,1\} \wedge 0 \leq x_i + y_i \leq 1\}$ we compute the probabilities:

$$P_{m/p,\tilde{m}/\tilde{p}} = P(m/p, \tilde{m}/\tilde{p}) = \sum_{i',j'} P(m/p, \tilde{m}/\tilde{p} | G_i = A_{i'}A_{j'}) P(A_{i'}A_{j'})$$

Each of the probabilities in the sums are expanded such that the events specified by m/p and $\tilde{m}/\tilde{p}$ are satisfied, e.g. $m/p = 1/0$ and $\tilde{m}/\tilde{p} = 1/0$ implying that both profile G_j and G_k matches the profiles of G_i on that particular locus:

$$\begin{aligned} P(1/0, 1/0) &= \sum_{i',j' \neq i'} P(A_{i'}A_{j'}, A_{i'}A_{j'} | A_{i'}A_{j'}) P(A_{i'}A_{j'}) + \sum_{i'} P(A_{i'}A_{i'}, A_{i'}A_{i'} | A_{i'}A_{i'}) P(A_{i'}A_{i'}) \\ &= 2 \sum_{i,j \neq i} P(A_{i'}A_{i'}A_{j'}A_{j'} | A_{i'}A_{j'}) P(A_{i'}A_{j'}) + \sum_{i'} P(A_{i'}A_{i'}A_{i'}A_{i'} | A_{i'}A_{i'}) P(A_{i'}A_{i'}) \\ &= 4 \sum_{i,j \neq i} P(A_{i'}A_{i'}A_{i'}A_{j'}A_{j'}A_{j'}) + \sum_{i'} P(A_{i'}A_{i'}A_{i'}A_{i'}A_{i'}A_{i'}). \end{aligned}$$

From the recursive formula $P(A_{i'}|\boldsymbol{x}^n) = [x^n_{i'}\theta + (1-\theta)p_{i'}]/[1 + (n-1)\theta]$, we see that the denominator do not depend on the total number of sampled alleles. Hence, for a probability like $P(A_{i'}A_{j'}A_{k'}A_{j'}A_{i'}A_{i'})$ that involves six alleles, the denominator will always be $\prod_{n=1}^{5}(1 + n\theta)$. Hence, to keep the formulae simple, we only consider the numerator in the following derivations. First, we observe that:

$$\begin{aligned} P(A_{i'}A_{j'}A_{k'}A_{j'}A_{i'}A_{i'}) &= P(A_{i'}|A_{j'}A_{k'}A_{j'}A_{i'}A_{i'}) P(A_{j'}A_{k'}A_{j'}A_{i'}A_{i'}) \\ &= [\theta(\beta_{i'} - 1) + (1-\theta)p_{i'}] P(A_{j'}A_{k'}A_{j'}A_{i'}A_{i'}) \\ &= \theta(\beta_{i'} - 1) P(A_{j'}A_{k'}A_{j'}A_{i'}A_{i'}) + (1-\theta)p_{i'} P(A_{j'}A_{k'}A_{j'}A_{i'}A_{i'}), \end{aligned} \tag{3.9}$$

where $\beta_{i'}$ counts the number of i' alleles in the expression on the left hand side. Now, the term $\theta(\beta_{i'} - 1)P(A_{j'}A_{k'}A_{j'}A_{i'}A_{i'})$ follows a similar expansion as the left hand side of (3.9). However, the latter term of (3.9) involves $p_{i'}$ which needs to be taken into account when evaluating $P(A_{j'}A_{k'}A_{j'}A_{i'}A_{i'})$. By following the recursion to the end, that is when the left hand side of (3.9) is, say, $P(A_{i'}A_{j'}) = P(A_{i'}|A_{j'})P(A_{j'}) = [(\beta_{i'} - 1)\theta + (1-\theta)p_{i'}]p_{j'} = (1-\theta)p_{i'}p_{j'}$ we end up with terms of the form $a_0\,\theta^{a_1}(1-\theta)^{a_2}p_1^{\alpha_1}\cdots p_K^{\alpha_K}$ for some constants $\boldsymbol{a} = (a_0, a_1, a_2)$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. The values of $\boldsymbol{a}$ and $\boldsymbol{\alpha}$ is build up during the recursion, hence determining the actual value is only a matter of bookkeeping.

Furthermore, consider the case where the product of allele probabilities is $p^2_{i'}p^2_{j'}p^2_{k'}$ where the indexes are different. A first step would be to replace $p^2_{k'} = S_2 - p^2_{i'} - p^2_{j'}$ and sum over $p^2_{i'}p^2_{j'}(S_2 - p^2_{i'} - p^2_{j'})$ for $i' \neq j'$. However, such calculations are very cumbersome to do by hand and from the equation below we see that there is a lot of repeated structure that may be exploited:

$$\sum_{i',j',k'}^{\neq} p^2_{i'}p^2_{j'}p^2_{k'} = \sum_{i',j'}^{\neq} p^2_{i'}p^2_{j'}(S_2 - p^2_{i'} - p^2_{j'}) = S_2 \sum_{i',j'}^{\neq} p^2_{i'}p^2_{j'} - \sum_{i',j'}^{\neq} p^4_{i'}p^2_{j'} - \sum_{i',j'}^{\neq} p^2_{i'}p^4_{j'},$$

where the notation imply summation over different values of the indexes. Rewriting the expression above with the powers replaced by the $\boldsymbol{\alpha}$-parameters we get this more general expression:

$$\sum_{i,j,k}^{\neq} p_{i'}^{\alpha_{i'}} p_{j'}^{\alpha_{j'}} p_{k'}^{\alpha_{k'}} = \left[\sum_{k'} p_{k'}^{\alpha_{k'}}\right] \sum_{i,j}^{\neq} p_{i'}^{\alpha_{i'}} p_{j'}^{\alpha_{j'}} - \sum_{i,j}^{\neq} p_{i'}^{\alpha_{i'}+\alpha_{k'}} p_{j'}^{\alpha_{j'}} - \sum_{i,j}^{\neq} p_{i'}^{\alpha_{i'}} p_{j'}^{\alpha_{j'}+\alpha_{k'}}$$

where all $\boldsymbol{\alpha}$-parameters were 2 in the previous example. The formula can be programmed in a computer as a recursion formula. Hence, in contrast to the simpler situations only involving a pair of DNA profiles where a few equations give the necessary probabilities (Weir, 2004, 2007), we let the computer compute the expectations $\mathbb{E}[\boldsymbol{M}(G_{i_1}, G_{i_2})\boldsymbol{M}(G_{i_1}, G_{i_3})^\top]$ and $\mathbb{E}[\boldsymbol{M}(G_{i_1}, G_{i_2})\boldsymbol{M}(G_{i_3}, G_{i_4})^\top]$. We have implemented efficient functions in R to compute these and other expectations implying that variances is computed within 10 to 30 seconds on a 2.5 GHz laptop computer for each θ-value. In order to get a impression of the structure in the matrix we have plotted a heat-map of the correlation matrix $\Omega(\theta)$ computed by:

$$\Omega(\theta) = \text{diag}\Big[1/\sqrt{\text{diag}\{\Sigma(\theta)\}}\Big]\,\Sigma(\theta)\,\text{diag}\Big[1/\sqrt{\text{diag}\{\Sigma(\theta)\}}\Big]$$

In Figure 3.7 we have plotted $\Omega(0.03)$ in grey-scale colours. However, the on line supplementary material has a coloured animation showing the change in pattern in $\Omega(\theta)$ for $\theta = [0, 0.001, \ldots, 1]$.

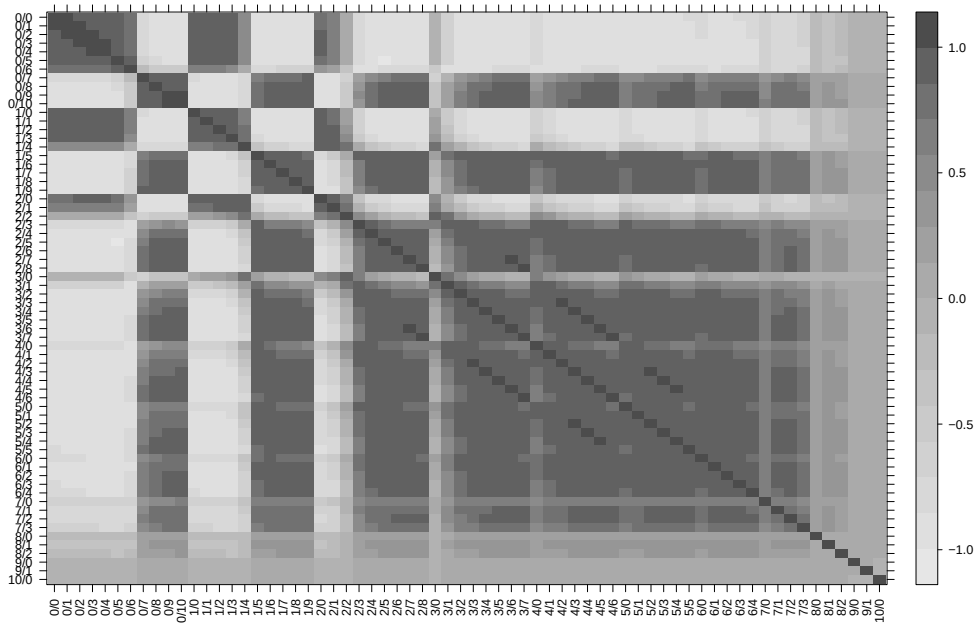

Figure 3.7: Graphical representation of the correlation matrix $\Omega(\theta)$ computed for $\theta = 0.03$ and $n = 10,000$.

Bibliography

Ayres, K. L. (2000). A two-locus forensic match probability for subdivided populations. *Genetica 108*, 137–143.

Balding, D. J. and R. A. Nichols (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica 96*, 3–12.

Curran, J. M. and T. Tvedebrink (2010a). DNAtools - a R package for forensic DNA database analysis. *Journal of Computational Statistics*. Manuscript in preparation.

Curran, J. M. and T. Tvedebrink (2010b). *DNAtools: Statistical functions for analysing forensic DNA databases*. R package version 0.1.

Curran, J. M., S. J. Walsh, and J. S. Buckleton (2007). Empirical testing of estimated DNA frequencies. *Forensic Sciences International: Genetics 1*, 267–272.

Donnelly, P. (1995a). Match probability calculations for multi-locus DNA profiles. *Genetica 96*, 55–67.

Donnelly, P. (1995b). Nonindependence of matches at difference loci in DNA profiles: quantifying the effect of close relatives on the match probability. *Heredity 75*, 26–34.

Lange, K. (1993). Match probabilities in racially admixed populations. *American Journal of Human Genetics 52*, 305–311.

Lange, K. (1995). Applications of the Dirichlet distribution to forensic match probabilities. *Genetica 96*, 107–117.

Laurie, C. and B. S. Weir (2003). Dependency effects in multi-locus match probabilities. *Theoretical Population Biology 63*, 207–219.

Mueller, L. D. (2008). Can simple populations genetic models reconcile partial match frequencies observed in large forensic databases? *Journal of Genetics 87*(2), 101–107.

Nichols, R. A. and D. J. Balding (1991). Effects of population structure on DNA fingerprint analysis in forensic science. *Heredity 66*, 297–302.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Song, Y. S. and M. Slatkin (2007). A graphical approach to multi-locus match probabilitiy computation: Revisiting the product rule. *Theoretical Population Biology 72*, 96–110.

Troyer, K., T. Gilroy, and B. Koeneman (2001). A nine STR locus match between two apparent unrelated individuals using AmpFℓSTR Profiler Plus™ and COfiler™. *Proceedings of the Promega 12th International Symposium on Human Identification*.

Tvedebrink, T. (2010). Overdispersion in allelic counts and θ-correction in forensic genetics. *Theoretical Population Biology*. In Press.

Weir, B. S. (2004). Matching and partially-matching DNA profiles. *Journal of Forensic Science 49*(5), 1–6.

Weir, B. S. (2007). The rarity of DNA profiles. *The Annals of Applied Statistics 1*(2), 358–370.

3.6 Supplementary remarks

It is relevant to be aware of and acknowledge the power of the DNA typing technology and its role in the society. To most people DNA evidence is thought of as flawless and superior to any other sort of evidence. However, due to the very nature of DNA profiles there is a possibility that *a* pair of apparently unrelated individuals share *a* DNA profile. As pointed out by Weir (2007, pp. 360-361) this is related to the 'birthday problem' where one computes the probability that at least two individuals out of n have the same unspecified birthday. The fact that $n = 23$ gives more than 50% probability of at least two individuals sharing birthday is surprising to many at first glance. However, this is due to the fact that the birthday is not specified. Similarly, when computing the probability that any two individuals share DNA profile the actual alleles of their common profile is not specified. If the profile were specified the computed probability would in fact be the match probability of two DNA profiles. When summing over the possible DNA profiles we obtained $N\pi_{L/0}(\theta)$, which was the expected number of pairs of individuals with identical DNA profiles. For the allele probabilities estimated from the Danish reference database we obtain $\pi_{10/0}(\theta) \approx (\alpha_0 + \alpha_1\theta)^{\alpha_2}$ which for non-negative parameters, $\hat{\boldsymbol{\alpha}} = (0.13, 0.87, 14.71)$, is a monotonic increasing function. That is, the probability increase with θ, i.e. the more heterogeneous the population is, the larger is the probability of coinciding DNA profiles.

However, this fact does not imply that DNA profiling is overrated nor that the weight of evidence reported in court is overstated. When using the *LR*-approach the reported evidential-value relates to *the specific* DNA profile of a suspect. The pairwise comparisons of each pair in the DNA database were used to validate the population genetic model. The diagnostics presented above indicated that the differences between the observed and expected counts were not too extreme, and thus we may still have confidence in the models used for reporting the evidential weight in court.

CHAPTER 4

Evaluating the weight of evidence using quantitative STR data in DNA mixtures

Publication details

Co-authors: Poul Svante Eriksen[*], Helle Smidt Mogensen[†] and Niels Morling[†]

[*] *Department of Mathematical Sciences*
Aalborg University
[†] *Section of Forensic Genetics, Department of Forensic Medicine*
Faculty of Health Science, University of Copenhagen

Journal: Applied Statistics (In Press)

DOI: 10.1111/j.1467-9876.2010.00722.x

Abstract:
The evaluation of results from mixtures of DNA from two or more persons in crime case investigations may be improved by taking not only the qualitative but also the quantitative part of the results into consideration. We present a statistical likelihood approach to assess the probability of observed peak heights and peak areas information for a pair of profiles matching the DNA mixture. Furthermore, we demonstrate how to incorporate this probability into the evaluation of the weight of the evidence by a likelihood ratio approach.

Our model is based on a multivariate normal distribution of peak areas for assessing the weight of the evidence. Based on data from analyses of controlled experiments with mixed DNA samples, we exploited the linear relationship between peak heights and peak areas, and the linear relations of the means and variances of the measurements. Furthermore, the contribution from one individual's allele to the mean area of this allele is assumed to be proportional to the average of peak height measurements of alleles, where the individual is the only contributor.

For shared alleles in mixed DNA samples, it is possible to observe only the cumulative peak heights and areas. Complying with this latent structure, we used the EM-algorithm to impute the missing variables based on a compound symmetry model. The measurements were subject to intra- and inter-locus correlations not depending on the actual alleles of the DNA profiles. Due to factorisation of the likelihood, properties of the normal distribution and use of auxiliary variables, an ordinary implementation of the EM-algorithm solved the missing data problem.

Keywords:
STR DNA mixture; Forensic genetics; Missing data; EM-algorithm; Compound symmetry model; Multivariate normal distribution

4.1 Introduction

4.1.1 DNA mixtures

The model presented in this paper is intended to be used in forensic genetics when facing DNA data from biological stains with more than one contributor (see Gill et al. (2006) for a detailed description of the DNA mixture problem). This specific problem has received increasing interest from both forensic geneticists and statisticians over the last decade, e.g. Evett and Weir (1998); Gill et al. (1998, 2006); Perlin and Szabady (2001); Bill et al. (2005); Cowell et al. (2007a).

When a crime has been committed, biological stains are often found at the scene of crime. DNA is present in almost all human cells and by using biochemical procedures, forensic geneticists are able to extract the DNA from the body fluids for further analysis. In many cases, more than one individual has contributed to a stain, which is then called a DNA mixture. Mixtures of DNA often appear in relation to crime cases, e.g. rape cases with one or more rapists, and cases involving violence. DNA may be extracted from semen obtained by a vaginal swab or from blood present on the victim's clothing.

In crime casework today, there is an international consensus to investigate DNA from short tandem repeat regions - STRs. The STR regions are situated between the coding regions in the DNA. The polymorphism of an STR region mainly results from differences in the number of repeated sequences. This leads to variations in the total lengths of the STR regions from person to person. In many European countries, ten STR systems and the sex-specific marker amelogenin are routinely investigated in crime cases by means of the SGM Plus STR kit (Applied Biosystems). The loci are located on different chromosomes. This is generally assumed to be sufficient to ensure statistical independence of alleles at different loci.

For the most common STR technologies used in forensic DNA analyses, the alleles are read from an electropherogram (pictured in Fig. 4.4) as peaks on a given scale. This makes two types of data available: qualitative allele type data, determined by the position of the peak (measured in DNA base pairs), and quantitative peak intensity data summarised by the height and area of the peak (measured in relative fluorescence unit, rfu). The set of observed alleles is termed a DNA profile.

The shaded cones in Fig. 4.4 show a typical picture of a DNA mixture comprising ten STR loci (denoted D3, vWA, ..., FGA in Fig. 4.4) used in forensic genetics. The peak height and area associated with an allele reflect the amount of DNA contributed to that particular allele. The potential peak positions of some loci overlap, which makes it necessary to use different fluorescent dyes (the different rows in Fig. 4.4 correspond to blue (D3, vWA, D16, D2), green (D8, D21, D18) and yellow (D19, TH0, FGA) fluorescent dyes) with a subsequent spectral deconvolution of the signal (Butler, 2005).

Depending on the DNA profiles mixed in the sample, the number of alleles present for each locus in a two-person DNA mixture ranges from one to four alleles since an individual may be either homozygous (carrying two identical copies of the same allele) or heterozygous (two different alleles), and the individuals may share one or both alleles. This implies that the amount of DNA contributed to each allele varies and the peaks are therefore expected to vary in height and area. In this paper, we present a statistical model for the peak areas for a given pair of profiles while taking into account the variable dimension (sub-vectors of dimension one to four for different loci) of the measurements.

4.1.2 Evaluating the weight of evidence

A complete DNA investigation is a very effective tool for excluding individuals who are not very closely related to the person from whom the stain material originated. A match between complete DNA profiles of a stain and a person is very strong evidence for the assumption that the stain came from that person compared with the assumption that the stain came from a random person. The weight of DNA evidence can be calculated in each case based on assumptions about the setting and knowledge of the distribution of the DNA characteristics in the relevant population. The weight of evidence from DNA investigations is generally accepted in almost all countries in which DNA investigations are used.

Methods are available to estimate the weight of the evidence of the qualitative results (Balding and Nichols, 1994; Evett and Weir, 1998). However, we do not have good mathematical methods to take into consideration the quantitative aspects of the DNA results in order to answer questions like: Can the two DNA profiles in a crime scene DNA mixture be identified based on the strength of the DNA results? Are the strengths of the various DNA results in a crime scene mixed DNA profile (that seems to consist of a major and a minor DNA profile) compatible with the hypothesis that the DNA comes from two persons with known DNA profiles?

Estimating the weight of evidence in forensic sciences is often done in terms of a likelihood ratio, which is the ratio of the probability of the evidence, $\mathcal{E}$, under two competing and mutually distinct but not exhaustive hypotheses. In the literature these two hypotheses are often denoted H_p and H_d for the "prosecutors hypothesis" and "defence hypothesis" respectively (Evett and Weir, 1998). Even though the hypotheses may have different origin than those of the prosecutor and defence, we apply the notation of H_p and H_d in this paper to denote the two disjoint events claimed in the hypothesis, i.e. the likelihood ratio is given by $LR = P(\mathcal{E}|H_p)/P(\mathcal{E}|H_d)$, where large values of LR support the H_p-hypothesis. For example in case of a rape the H_p-hypothesis may be: "The victim and the suspect are the contributors to the stain", whereas the H_d-hypothesis states: "The victim and an unknown individual unrelated to the suspect are the contributors to the stain". We denote the crime scene evidence from the mixture $\mathcal{E}_c = (\mathcal{G}, \mathcal{Q})$, where $\mathcal{G}$ denotes the qualitative allele information, and $\mathcal{Q}$ represents the quantitative peak information as measurements of peak heights and areas. The most frequent way to assess the probability $P(\mathcal{E}|H)$ is by solely using the qualitative information $\mathcal{G}$ in terms of allele probabilities. In DNA mixtures, however, this may discard important quantitative information of the DNA evidence. Thus, the probability of the evidence $\mathcal{E}$ given a hypothesis H needs to include both parts of the evidence $\mathcal{G}$ and $\mathcal{Q}$.

We define G_V, G_S and G_U to be the profiles of the victim, the suspect and a potential unknown and unrelated contributor, respectively. Both hypotheses H_p and H_d in our rape example are formulated such that they are consistent with $\mathcal{G}$, i.e. all alleles in $\mathcal{G}$ are accounted for and only alleles in $\mathcal{G}$ appear in the included profiles G_V, G_S and G_U. When fixing only one profile, G'', of a two-person mixture, the consistency with $\mathcal{G}$ induces the set $\mathcal{C} = \{G' : (G', G'') \equiv \mathcal{G}\}$, which are all profiles, G', that together with G'' are consistent with $\mathcal{G}$. If the H_p-hypothesis claims $\mathcal{G}$ to be a mixture of G_V and G_S, H_p:(G_V, G_S), while the H_d-hypothesis claims it is a mixture of G_V and G_U, H_d:(G_V, G_U), the likelihood ratio is

$$LR = \frac{P(\mathcal{E}_c, G_S, G_V|H_p)}{P(\mathcal{E}_c, G_S, G_V|H_d)} = \frac{P(\mathcal{Q}, \mathcal{G}, G_S, G_V|H_p)}{P(\mathcal{Q}, \mathcal{G}, G_S, G_V|H_d)} = \frac{P(\mathcal{Q}|\mathcal{G}, G_S, G_V, H_p)P(\mathcal{G}, G_S, G_V|H_p)}{P(\mathcal{Q}|\mathcal{G}, G_S, G_V, H_d)P(\mathcal{G}, G_S, G_V|H_d)},$$

where G_S and G_V enter as evidence as these are determined from the case circumstances. Let $\mathcal{C}_d = \{G_U : (G_V, G_U) \equiv \mathcal{G}\}$ be the set of unknown profiles that together with G_V are consistent with $\mathcal{G}$, then $P(\mathcal{G}|G_V, G_U) = 1$ for $G_U \in \mathcal{C}_d$ and 0 otherwise, i.e. the set of possible unknowns under H_d. We expand the denominator of the LR using hypothesis H_d,

$$P(\mathcal{E}, G_S, G_V|H_d) = \sum_{G_U \in \mathcal{C}_d} P(\mathcal{Q}|\mathcal{G}, G_S, G_V, G_U)P(\mathcal{G}, G_S, G_V, G_U),$$

Table 4.1: The four DNA profiles used in the controlled pairwise two-person mixture experiments.

	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
A	14,18	17,19	12,14	20,24	10,13	30.2,32.2	13,13	12,13	8,9	20,22
B	15,16	14,16	10,12	17,25	13,16	30,30	13,13	14,15	6,9	19,23
C	15,16	15,17	11,11	19,25	8,12	29,31	15,17	13,13	6,8	23,24
D	16,19	15,17	10,12	23,25	13,13	28,30	12,16	13,15	6,7	20,23

where $P(\mathcal{Q}|\mathcal{G}, G_S, G_V, G_U) = P(\mathcal{Q}|G_V, G_U)$ and $P(\mathcal{G}, G_S, G_V, G_U) = P(G_S, G_V, G_U)$ due to $(G_V, G_U) \equiv \mathcal{G}$ and H_d is assumed. Hence,

$$P(\mathcal{E}, G_S, G_V|H_d) = \sum_{G_U \in \mathcal{C}_d} P(\mathcal{Q}|G_V, G_U)P(G_S, G_V, G_U).$$

Similar arguments apply to the numerator of *LR*, and assuming independence between the profiles involved, i.e. unrelated individuals such that $P(G_S, G_V, G_U) = P(G_S)P(G_V)P(G_U)$, the final *LR* expression is:

$$LR = \frac{P(\mathcal{Q}|G_S, G_V)}{\sum_{G_U \in \mathcal{C}_d} P(\mathcal{Q}|G_V, G_U)P(G_U)}, \tag{4.1}$$

where the factors $P(G_S)P(G_V)$ have cancelled out. The numerator $P(\mathcal{Q}|G_S, G_V)$ of (4.1) assesses the probability of observing the quantitative information given that the mixture consists of genetic material from the profiles G_S and G_V. The denominator equals the mean value of the quantitative likelihood among the pairs of profiles that are consistent with the genetic trace. If we assume $P(\mathcal{Q}|G_S, G_V) = P(\mathcal{Q}|G_V, G_U)$ for all G_U, i.e. the observed quantitative information has equal probability for all profiles paired with G_V, then (4.1) reduces to the usual likelihood ratio as in Evett and Weir (1998), since $P(\mathcal{Q}|G_S, G_V)$ and $P(\mathcal{Q}|G_V, G_U)$ then cancel each other in (4.1). The assumption that the profiles G_S, G_V and G_U are independent is a rather strong. The so-called "θ-correction" incorporates the correlation from shared ancestry (Balding and Nichols, 1994) and closer familial relationships induces further correlation of the genetic profiles. However, for the purpose of introducing the factorisation of the qualitative and quantitative evidence the assumption used in (4.1) is adequate.

The objective of the present paper is to develop a methodology and an adequate statistical model to describe $P(\mathcal{Q}|G', G'')$, where both of the true profiles G' and G'' are known. This comprises a mathematical formalism of inter-locus dependencies of the quantitative evidence, the relationships between a sample's peak heights, peak areas, and the amount of DNA contributed to the individual peak.

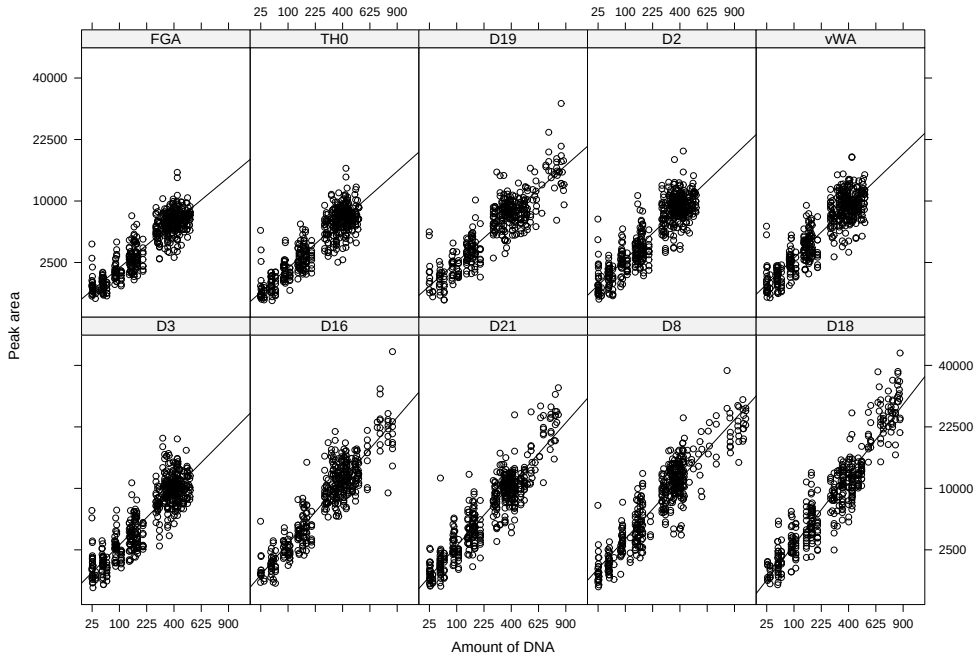

Figure 4.1: Proportionality of peak areas and amounts of DNA of square root transformed data.

4.2 Material and methods

4.2.1 Experimental data

The assumptions made as to the amplification behaviour of mixed DNA samples were based on data exploration of controlled experiments conducted at The Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen. The experiments consisted of pairwise two-person mixtures in various mixture ratios of the four profiles in Table 4.1. The data were prepared as described in Tvedebrink et al. (2009). The assumptions made did not contradict the assumptions made in e.g. Cowell et al. (2007a):

1. proportionality of the peak areas and the amount of DNA in the sample,
2. linearity of the observed peak areas and peak heights,
3. proportionality of the means and variances of peak areas.

These assumptions were supported by the plots in Fig. 4.1 and Fig. 4.2, which were based on data from the experiments described in Tvedebrink et al. (2009). The validity of the last assumption was emphasised by fitting a linear model: $\text{Area}/\sqrt{\text{DNA}} = \beta_s \sqrt{\text{DNA}} + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and with β_s being a locus specific proportionality factor. Graphical inspections show no systematic dependence of squared residuals and DNA.

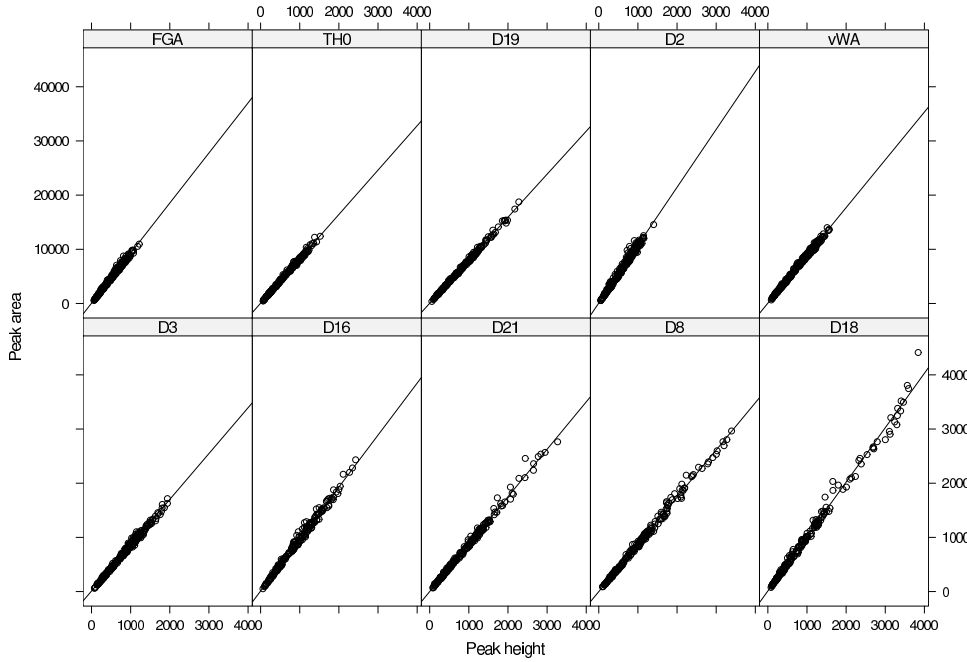

Figure 4.2: Proportionality of peak heights and peak areas. The proportionality factor depends on loci.

4.2.2 Model description

In a DNA mixture of profiles C and D (Table 4.1) we would observe peaks for alleles 15, 16 and 19 in locus D3. The peaks are expected to vary in height and area due to the different amounts of DNA contributed to the alleles, e.g. both profiles contribute to the peak of allele 16. For identical alleles we assume that the peak areas of each individual are additive resulting in an observable cumulative vector of peak areas, $\boldsymbol{M}$. Similarly, for homozygous profiles, the contribution to the observable peak area is the sum of two identical peak areas.

The unobservable peak areas, $\boldsymbol{A}$, from each individual are input for modelling the observable quantitative data, $\mathcal{Q}$, from DNA traces for the assessment of $P(\mathcal{Q}|G_V, G_P)$, where G_V and G_P are the profiles of the victim and true perpetrator, respectively. We use the EM-algorithm in addressing the DNA mixture problem because it can be formulated as a missing data problem (Little and Rubin, 2002). The model is derived for two-person mixtures but can be extended to cope with more than two contributors.

In the following, we let $\mathcal{S}$ denote the set of loci and S the number of loci used for identification, i.e. $|\mathcal{S}| = S$. For parameter estimation, we have access to data from C mixtures of known profiles.

The amount of DNA contributed to the mixture by person k, $k = 1, 2$, was modelled by $H^{(k)}$. This is a sum of the observed peak heights with person k as the only contributor divided by a sum

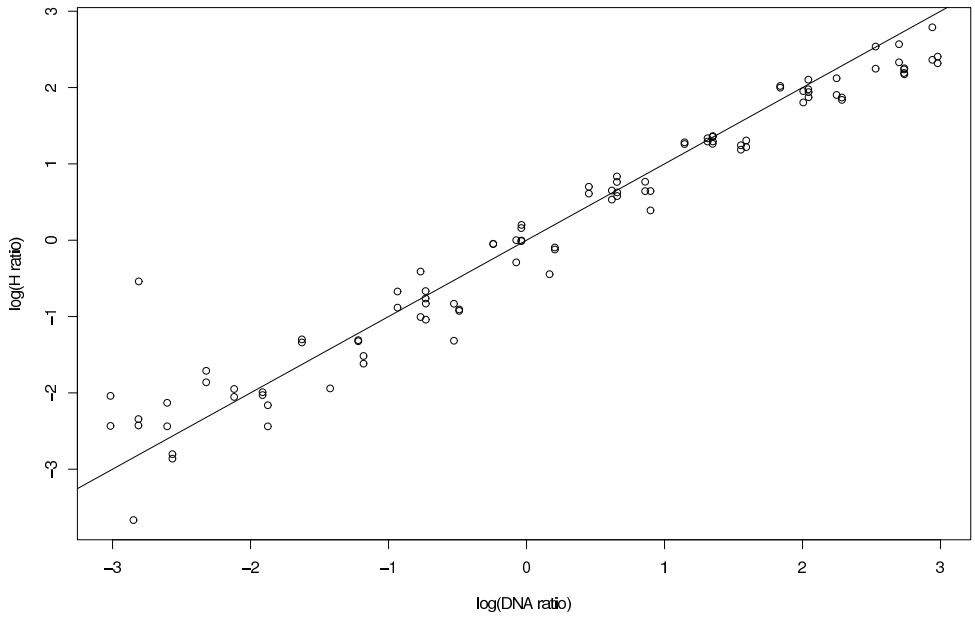

Figure 4.3: Linearity of the H-ratio and DNA-ratio with the identity line ($y = x$) superimposed. The outlier at $(-2.77, -0.55)$ was due to entry error of the laboratory.

of indicators with value two and one for alleles from loci where person k is homozygous and heterozygous, respectively. Let $h_i^{(k)}$ be the ith peak height with person k as the only contributor, $H^{(k)} = \{n_{\text{het}}^{(k)} + 2n_{\text{hom}}^{(k)}\}^{-1} \sum_{i=1}^{n^{(k)}} h_i^{(k)}$, where $n^{(k)} = n_{\text{het}}^{(k)} + n_{\text{hom}}^{(k)}$ is the number of person ks alleles from heterozygous, $n_{\text{het}}^{(k)}$, and homozygous loci, $n_{\text{hom}}^{(k)}$, and person k is the only contributor. Thus, $H^{(k)}$ is an estimate of the average peak height associated with person ks alleles.

Fig. 4.3 shows a plot of the ratio $H^{(1)}/H^{(2)}$ against the DNA ratio reported by the laboratory. The data demonstrate that it is reasonable to use $H^{(k)}$ as a proxy for the amount of DNA contributed by person k. Furthermore, for each pair of profiles, the quantities $H^{(1)}$ and $H^{(2)}$ can be computed using only the peak height observations.

We assumed independence among the components of $\boldsymbol{A}$ and that they followed a normal distribution with both mean and variance proportional to the amount of DNA. The components of $\boldsymbol{A}$ are $A_{s,i}^{(k)}$ for person k, locus s and allele i. We have $A_{s,i}^{(k)} \sim \mathcal{N}(\alpha_s H^{(k)}, \sigma_s^2 H^{(k)})$ which implies the same distribution of both alleles of locus s for person k.

The parameters α_s and σ_s, $s \in \mathcal{S}$, are locus dependent and shared for all cases, $c = 1, \ldots, C$. This parameterisation ensures the proportionality of the mean and variance and that both are proportional to the amount of DNA modelled by $H^{(k)}$. Since $H^{(k)}$ is the same for all loci, the αs ensure that the amplification efficiency may vary between loci. Hence, the magnitude of α_s reflects the emission intensity of locus s. Furthermore, the variation modelled by σ_s^2 can be interpreted as the data preprocessing variation of the STR allele signals, e.g. variability from

pipetteting the samples.

The relation between $\boldsymbol{M}$ and $\boldsymbol{A}$ is expressed as a linear transformation, T, adding together peak areas from identical alleles, and an additional error term related to the measurement error, $\boldsymbol{M} = T\boldsymbol{A} + \varepsilon$. For the measurement errors, ε, we assume independence of $\boldsymbol{A}$ and multivariate normal distribution with some dependencies within and across loci. We denote the covariance of ε as $\text{Cov}(\varepsilon) = \Omega$. Let the dimension of $\boldsymbol{M}$ be $n = \sum_{s\in\mathcal{S}} n_s$, where n_s, $1 \leq n_s \leq 4$, is the number of observed alleles in locus s. The transformation, T, is an $n \times 4S$-block diagonal matrix with block matrices T_s with 0 and 1 entries according to the profiles in the mixture. For each locus, s, we sort the unobservable peak areas, $\boldsymbol{A}_s$, by allelic number of each person, whereas $\boldsymbol{M}_s$ is sorted by allelic number. For a mixture of profile B and D from Table 4.1, the genotypes in locus $s = \text{D3}$ are $P_s^{(1)} = (15, 16)$ and $P_s^{(2)} = (16, 19)$, and the associated matrix T_s is

$$\boldsymbol{M}_s = (M_{s,15}\,,\, M_{s,16}\,,\, M_{s,19})^\top = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \left(A_{s,15}^{(1)}\,,\, A_{s,16}^{(1)}\,,\, A_{s,16}^{(2)}\,,\, A_{s,19}^{(2)}\right)^\top + \varepsilon_s,$$

adding together the entries in $\boldsymbol{A}_s$ that relates to the same allele, i.e. allele 16.

The number of allelic measurements within each locus varies from case to case since different pairs of profiles will share a different number of alleles. A mixture of person A and B would have $n_{\text{D3}} = 4$, and B and C has $n_{\text{D3}} = 2$ (see Table 4.1). Not only will the number of alleles vary, the specific alleles present in a given mixture depends on the profiles in the mixture, e.g. A and B give alleles $\{14, 15, 16, 18\}$, and B and C give $\{15, 16\}$. This makes it difficult to incorporate a covariance structure covering all allele combinations.

We standardised the residual, ε, by the observed peak heights, $\boldsymbol{h} = (\boldsymbol{h}_s)_{s\in\mathcal{S}}$ with $\boldsymbol{h}_s = (h_{s,i})_{i=1}^{n_s}$, by defining the scaled residual, $\tilde{\varepsilon} = (\tilde{\varepsilon}_s)_{s\in\mathcal{S}}$, where $\tilde{\varepsilon}_s = (\varepsilon_{s,i}/\sqrt{h_{s,i}})_{i=1}^{n_s}$. To make the model operational, we assumed a compound symmetry model for the covariance of $\tilde{\varepsilon}$, $\text{Cov}(\tilde{\varepsilon}) = \tilde{\Omega}$ and that this does not depend on the specific alleles in the mixture. The only case specific adjustment made was to make the dimensions of the compound symmetry concordant with the number of observed peaks for each locus. The compound symmetry structure of $\tilde{\Omega}$ implies that sub-vectors of $\tilde{\varepsilon}$ share some properties with respect to the scaled covariance $\tilde{\Omega}$. There are three different types of correlation in our setting:

- Different loci ($s \neq t$): $\text{Cov}(\tilde{\varepsilon}_{s,i}, \tilde{\varepsilon}_{t,j}) = \nu_{st}$.
- Same locus, different alleles ($s = t, i \neq j$): $\text{Cov}(\tilde{\varepsilon}_{s,i}, \tilde{\varepsilon}_{s,j}) = \nu_{ss}$.
- Same locus, same allele ($s = t, i = j$): $\text{Cov}(\tilde{\varepsilon}_{s,i}, \tilde{\varepsilon}_{s,i}) = \text{Var}(\tilde{\varepsilon}_{s,i}) = \nu_{ss} + \tau_s$.

Hence, we can parameterise $\tilde{\Omega}$ by $\boldsymbol{\tau} = \{\tau_s\}_{s\in\mathcal{S}}$ and $\Lambda = \{\nu_{st}\}_{s,t\in\mathcal{S}}$. The interpretation of ν_{st} is that the correlations between observations at different loci depend only on the loci and not on the specific alleles present on each locus. Similarly, the correlation between alleles on the same locus, s, is independent of the specific alleles, whereas for identical elements, the covariance corresponds to the variance, and the addition of τ_s allows for a larger variance than that given by the intra-locus covariance.

4.2.3 Implementation of the EM-algorithm

In order to handle the latent structure of $\boldsymbol{A}$ and the associated missing data problem, we used the EM-algorithm to impute the missing observations and estimate the parameters in the conditional distribution of $\boldsymbol{A}$ given $\boldsymbol{M}$. However, since the dimensions of $\boldsymbol{M}$ and sub-vectors hereof varied from case to case, we obtained a likelihood, which was not very well suited for the implementation of the EM-algorithm. The problem was solved by introducing appropriate auxiliary variables.

This allowed for an implementation of the EM-algorithm in the usual full exponential family framework with the constraint that the ν_{ss}-parameters should be positive, i.e. this method implies positive intra-locus covariances. However, the inter-locus covariances ν_{st} are not constrained. The parameters estimated using the EM-algorithm are not case specific but reflect the distribution of the quantitative STR DNA in the laboratory.

Appendices 4.A and 4.B give mathematical details on the model and the implementation of the EM-algorithm.

4.3 Impact on the likelihood ratio

As mentioned in Section 4.1.2, both the qualitative and quantitative evidence need to be evaluated for proper use of the available information from a crime scene. The probability $P(\mathcal{Q}|G', G'')$ in the likelihood ratio of (4.1) is evaluated by using the fitted model to calculate $L(\boldsymbol{M}|G', G'') = |\Sigma_{(G',G'')}|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{M} - \boldsymbol{\mu}_{(G',G'')})^\top \Sigma^{-1}_{(G',G'')}(\boldsymbol{M} - \boldsymbol{\mu}_{(G',G'')})\}$ of (G', G'') and thus yielding the observed signal $\boldsymbol{M}$, whereas $P(G')$ as usual is assessed using the allele frequencies (Evett and Weir, 1998).

Consider a more complicated case with no identified victim where the crime scene stain is assumed to be a mixture of two DNA profiles, e.g. DNA extracted from a cigarette butt found at the scene of crime. Then, given a suspect profile G_S, the H_p-hypothesis claims the stain to be a mixture of the suspect and an unrelated unknown profile, H_p:(G_U, G_S), whereas the H_d-hypothesis states it is a mixture of two unrelated unknown profiles, H_d:(G_{U_1}, G_{U_2}). We form two sets $\mathcal{C}_p = \{G_U : (G_S, G_U) \equiv \mathcal{G}\}$ and $\mathcal{C}_d = \{(G_{U_1}, G_{U_2}) : (G_{U_1}, G_{U_2}) \equiv \mathcal{G}\}$, consistent with each hypothesis. Similar arguments as used for obtaining (4.1) imply that *LR* is:

$$LR = \frac{P(\mathcal{E}, G_S|H_p)}{P(\mathcal{E}, G_S|H_d)} = \frac{\sum_{G_U \in \mathcal{C}_p} L(\boldsymbol{M}|G_U, G_S)P(G_U)}{\sum_{(G_{U_1}, G_{U_2}) \in \mathcal{C}_d} L(\boldsymbol{M}|G_{U_1}, G_{U_2})P(G_{U_1})P(G_{U_2})}.$$

Note that the sum in the denominator involves $7^{S_2}12^{S_3}6^{S_4}$ terms, where S_i is the number of loci with i observed peaks. This follows from the fact that there are 7, 12 and 6 possible combinations for two, three and four alleles to be assigned to two individuals, respectively. However, this often yields an intractable number of combinations, where only a limited number of pairwise profiles actually have a likelihood value, $L(\boldsymbol{M}|G_{U_1}, G_{U_2})$, large enough to have numerical impact on *LR*.

Table 4.2: Data stratified according to STR locus.

Locus	Dye	Allele	Height	Area
D3	Blue	15	1135	10301
D3	Blue	16	1031	9405
vWA	Blue	14	371	3365
vWA	Blue	15	921	8654
vWA	Blue	16	395	3610
vWA	Blue	17	804	7382
D16	Blue	10	485	4913
D16	Blue	11	2110	21651
D16	Blue	12	417	4304
D2	Blue	17	196	2121
D2	Blue	19	700	7713
D2	Blue	25	951	11209
D8	Green	8	774	7052
D8	Green	12	1006	9297
D8	Green	13	344	3166
D8	Green	16	291	2675
D21	Green	29	774	7152
D21	Green	30	789	7240
D21	Green	31	982	9174
D18	Green	13	593	6455
D18	Green	15	1002	10758
D18	Green	17	865	9458
D19	Yellow	13	1614	13532
D19	Yellow	14	211	1849
D19	Yellow	15	182	1647
TH0	Yellow	6	797	6894
TH0	Yellow	8	505	4334
TH0	Yellow	9	198	1751
FGA	Yellow	19	173	1606
FGA	Yellow	23	880	8720
FGA	Yellow	24	647	6682

4.3.1 Example

We illustrate that the inclusion of the quantitative peak information, $\mathcal{Q}$, is important when evaluating the weight of evidence in a mixture. In the example, we demonstrate the properties of our approach when the data of Table 4.2 are observed.

In order to limit the number of profiles in *LR*, we applied the guidelines of Bill et al. (2005). These guidelines evaluate each mixture using heuristic rules about peak height balances and mixture proportions. The authors define the heterozygote balance *Hb* as the ratio of two non-shared peaks of an assumed heterozygous profile, and provide estimators of mixture proportions within each locus, $\hat{M}_x^s$. If a two-person mixture is to pass the guideline criteria, it must satisfy $3/5 \leq Hb \leq 5/3$ and $\hat{M}_x - 0.35 \leq \hat{M}_x^s \leq \hat{M}_x + 0.35$, where $\hat{M}_x = S^{-1} \sum_{s \in \mathcal{S}} \hat{M}_x^s$ is an estimate of the overall mixture proportion. We used ± 0.25 as limits on $\hat{M}_x^s$ which resulted in 860 pairs satisfying the heuristic rules of Bill et al. (2005).

However, instead of assigning equal weight to all these pairs, we evaluate $L(\boldsymbol{M}|G', G'')$ for each pair of profiles. As mentioned in Bill et al. (2005), this approach will not yield the correct *LR* as all possible combinations should be weighted by their associated $L(\boldsymbol{M}|G', G'')$-value. This attempt to evaluate the *LR* aims at including more of the available information and thus yielding a better approximation to the actual *LR*, since each pair of profiles has its own weight reflecting how well it fits the quantitative data.

In the example, we demonstrate the effect of including the quantitative information in the evidence evaluation for three different suspect profiles. The suspect profiles used in the example

Table 4.3: Profiles of the suspects (a)-(c), unknowns and best matching pairs of profiles ($\star$) in example of Section 4.3.1. For all the suspects, only one unknown matches the chosen suspect among the 860 combinations. In loci where the suspect combination differs from the best matching combination in part ($\star$), allelic numbers are in bold font.

Locus		D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
(a)	Suspect	15,16	14,16	10,12	**17**,**17**	13,16	**29**,**31**	**15**,**15**	14,15	**9**,**9**	**19**,**19**
	Unknown	15,16	15,17	11,11	**19**,**25**	8,12	**30**,**31**	**13**,**17**	13,13	**6**,**8**	**23**,**24**
(b)	Suspect	15,16	14,16	10,12	17,25	13,16	**30**,**30**	**17**,**17**	14,15	6,9	**19**,**19**
	Unknown	15,16	15,17	11,11	19,25	8,12	**29**,**31**	**13**,**15**	13,13	6,8	**23**,**24**
(c)	Suspect	15,16	14,16	10,12	17,25	13,16	**30**,**30**	**13**,**15**	14,15	6,9	19,23
	Unknown	15,16	15,17	11,11	19,25	8,12	**29**,**31**	**15**,**17**	13,13	6,8	23,24
($\star$)	Minor	15,16	14,16	10,12	17,25	13,16	29,29	13,13	14,15	6,9	19,23
	Major	15,16	15,17	11,11	19,25	8,12	30,31	15,17	13,13	6,8	23,24

are given in Table 4.3, together with the unknown profile G_U that maximises $L(\boldsymbol{M}|G_S, G_U)$ for each suspect profile, G_S. For each suspect profile, only one of the 860 pairs of profiles satisfies $(G_U, G_S) \equiv \mathcal{G}$ which implies a product of $L(\boldsymbol{M}|G_S, G_U)$ and $P(G_U)$ in the numerator for each suspect profile, and 860 terms in the sum of the denominator of which the combination of "Minor" and "Major" of Table 4.3, part ($\star$) has the largest quantitative likelihood value. Throughout the example, the main focus will be on the suspect of part (a) in Table 4.3, with comparisons to the results obtained using the suspects of part (b) and (c).

In Fig. 4.4 and Fig. 4.5, the observed quantitative peaks, ▲, are plotted together with the expected peaks, △, for the profiles of part (a) and ($\star$) of Table 4.3, respectively. The expected peaks are given by $\hat{\boldsymbol{M}} = T\hat{\boldsymbol{\mu}}$, where T and $H^{(k)}$ in $\hat{\mu}_{s,k} = \hat{\alpha}_s H^{(k)}$ are computed for the specific pair of profiles. It is clear from Fig. 4.4 that the imbalances induced by the suspect combination in part (a) imply substantial deviation from the observed data for loci D2, D21, D18, TH0 and FGA. These are also the loci where the two pairs of profiles of part (a) and ($\star$) in Table 4.3 differ.

First, we make a non-quantitative evaluation of the *LR* using only allele probabilities for the suspect of part (a). Since there is only one combination among the 860 that includes this suspect, the likelihood ratio $LR = P(G_U)/[\sum P(G_{U_1})P(G_{U_2})]$, where the sum in the denominator is over the set $\mathcal{C}_d$, but here this set consists of 860 combinations satisfying $Hb \in [3/5\,;\,5/3]$ and $\hat{M}_x^s \in [\hat{M}_x \pm 0.25]$ for computational simplicity. This yields a non-quantitative likelihood ratio, $LR_{\mathcal{G}}$, estimate of $4.527{\times}10^{13}$, which is very strong evidence in favour of the hypothesis that the suspect is a contributor to the stain.

The dominating values of the quantitative likelihood in the numerator and denominator are given by $L(\boldsymbol{M}|G_S^{(a)}, G_U^{(a)}) = 5.9{\times}10^{-119}$ and $L(\boldsymbol{M}|G_{U_1}^{(\star)}, G_{U_2}^{(\star)}) = 5.57{\times}10^{-100}$ respectively. A large difference in the quantitative likelihood values was expected from the difference in fit to the observed peaks pictured in Figs. 4.4 and 4.5. Thus, including the quantitative evidence, the quantitative likelihood ratio estimate, $LR_{\mathcal{GQ}}$, decreased by a factor 10^{17} to $7.63{\times}10^{-4}$ which is strongly in favour of the suspect not having contributed to the stain.

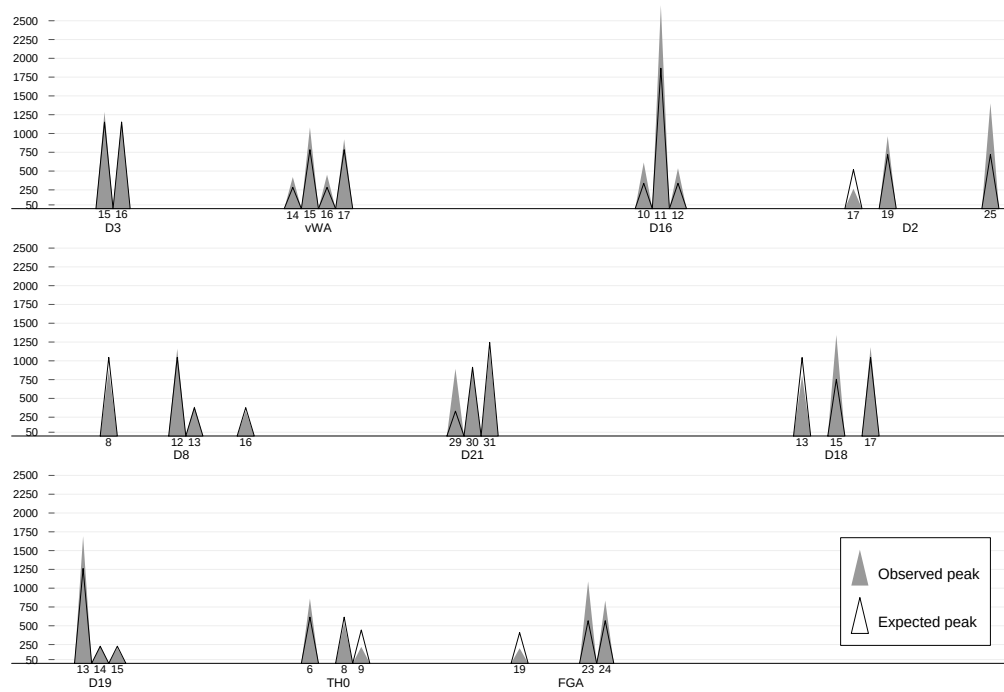

Figure 4.4: Observed, ▲, and expected peaks, △, assuming a two-person mixture of the suspect and unknown in Table 4.3, part (a). Abscissa: Basepair (bp) values computed using the allelic number and STR locus, Ordinate: Peak heights in rfu.

Table 4.4: Likelihood ratios for the three different suspects in Table 4.3. Here, $LR_{\mathcal{G}}$ and $LR_{\mathcal{GQ}}$ denote the non-quantitative and quantitative likelihood ratios, respectively, and $LR_{\mathcal{GQ}}/LR_{\mathcal{G}}$ is the relative change in the weight of the evidence. The allele frequencies used in the calculations were provided by The Section of Forensic Genetics, University of Copenhagen.

	$LR_{\mathcal{G}}$	$LR_{\mathcal{GQ}}$	$LR_{\mathcal{GQ}}/LR_{\mathcal{G}}$
Suspect (a)	4.527×10^{13}	7.630×10^{-4}	1.685×10^{-17}
Suspect (b)	4.216×10^{13}	5.185×10^{8}	1.230×10^{-5}
Suspect (c)	3.596×10^{13}	9.744×10^{13}	2.710

Together with similar computations for the suspects of parts (b) and (c), this information is given in Table 4.4. Here, we see that for suspects of part (b) and (c), the change in the weight of evidence is a moderate decrease and small increase, respectively. Note that part (b) differs from the best matching pair of profiles in three loci (D21, D18 and FGA) and part (c) in the two loci D21 and D18.

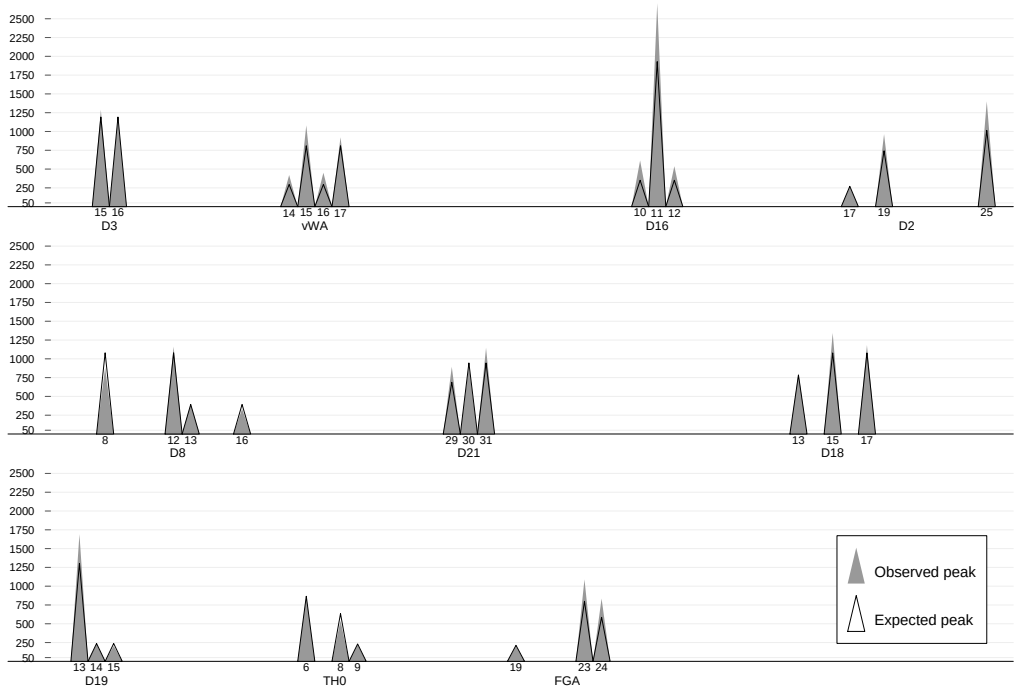

Figure 4.5: Observed, ▲, and expected peaks, △, assuming a two-person mixture of the minor and major profiles in Table 4.3, part (⋆). Abscissa and ordinate as in Fig. 4.4.

The non-quantitative likelihood ratio estimates, $LR_{\mathcal{G}}$, of Table 4.4 will in many legal systems point towards conviction of any of the suspects. When including the quantitative information, we see that the change in the weight of evidence may add further to the evidence against the suspect (as in part (c)), or may decrease the likelihood ratio estimate such that it provides strong evidence in favour of the suspect (part (a)), however, also situations in between these two extremes will occur (part (b)). This example shows that, even when a person's genotype matches the genetic stain, imbalanced STR DNA profiles judged by the observed quantitative data may speak strongly in favour of the suspect. However, weighing each pair of genotypes by the associated quantitative likelihood-value may add further to the evidence against the suspect when the suspect's profile only causes a few or small imbalances with respect to the observed peaks.

4.4 Parameter estimation

The EM-algorithm and the specific expressions as derived in Appendix 4.B were implemented in the statistical software package R (R Development Core Team, 2009). In order to validate the implementation, we simulated peak area data given the peak heights from controlled experiments and known model parameters. After 30,000 iterations, the parameter estimates were close to the

true values indicating a successful implementation of the fitting algorithm.

In order to estimate the model parameters, we used a training set consisting of results of investigations of DNA mixtures from 71 controlled experiments conducted at The Section of Forensic Genetics, University of Copenhagen. These 71 cases were chosen such that all alleles from the contributing profiles were present in the data, i.e. no drop-out events occurred (see Tvedebrink et al., 2009, for discussion on allelic drop-out). The algorithm was executed using several different sets of initial values. For each set, we ran 30,000 iterations of the EM-algorithm all converging to the same parameter estimates.

In order to monitor the convergence of the EM-algorithm, we computed the deviance after each iteration. After 1,100 iterations, the absolute improvement for successive deviances was less than 0.01.

In the Λ part of Table 4.5, the shading shows the locus correlations, $\nu_{st}/\sqrt{\nu_{ss}\nu_{tt}}$, while the above-diagonal part shows the locus covariances, ν_{st}, when $\tau_s = 0$ (see Section 4.5.2). Most of the loci were highly correlated. This indicates that evaluation of quantitative DNA evidence with the assumption of independence across loci is an extensive simplification.

The different signal intensities of the fluorescent dyes were also identifiable in the parameter estimates. The strong signals of the green dye band and the weaker signals of the yellow dye band (Butler, 2005) were reflected in the parameter estimates of α_s. In Table 4.5, we see that the magnitude of the αs of the yellow fluorescence was smaller than that of the blue fluorescence, which again was smaller than that of the green fluorescence (except for loci D16 and D21).

In addition to the parameter estimates and deviance, we also computed the asymptotic variances of the estimates by the normality approximation of the MLE with the inverse Fisher Information as covariance matrix. We found that the estimated standard deviation of both $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}^2$ indicated reasonably good estimates of these parameters. Large asymptotic standard deviations of Λ did, however, indicate the possibility of model reductions.

4.5 Discussion

4.5.1 Validity of the hypothesis of a two-person mixture

When analysing the STR results of a crime scene stain, we need to be able to determine whether the stain is likely to originate from a two-person mixture or not. In this section, we demonstrate how this is possible using our model for the quantitative STR DNA data. In order to verify the hypothesis of a given two-person mixture, we simulated 1,000 vectors of peak areas, $\boldsymbol{M}_1^*\ldots,\boldsymbol{M}_{1000}^*$, for each of the 71 cases from the controlled experiments.

Simulations of the peak areas were conditioned on the observed peak heights and true profiles of the mixture, and we used the parameter estimates from Table 4.5. This corresponds to simulating under a null hypothesis with the T-matrix, $H = (H^{(1)}, H^{(2)})$ and $\boldsymbol{h}$ known together with fixed parameters $\boldsymbol{\alpha}$, $\boldsymbol{\sigma}^2$ and Λ, i.e. assuming that the stain originates from a two-person mixture.

Table 4.5: Parameter estimates after 30,000 iterations of the EM-algorithm with $\boldsymbol{\tau} = \mathbf{0}$ (Section 4.5.2). The Λ-matrix shows the covariances ν_{st} and correlations $\nu_{st}/\sqrt{\nu_{ss}\nu_{tt}}$ (shaded).

	Yellow dye flourescence				Blue dye flourescence				Green dye flourescence				
	FGA	*TH0*	*D19*	*D2*	*vWA*	*D3*	*D16*	*D21*	*D8*	*D18*	
Λ	1151.54	773.91	1441.00	1492.01	1044.48	857.46	1305.36	1033.50	397.34	1461.59	*FGA*
	0.75	925.69	1042.03	1090.16	587.58	664.55	527.21	654.26	582.88	1085.37	*TH0*
	0.94	0.76	2052.83	2151.76	1319.49	1050.95	1716.85	1279.93	619.22	1964.68	*D19*
	0.88	0.72	0.95	2481.70	1438.36	1237.07	1848.14	1354.94	765.35	2077.81	*D2*
	0.94	0.59	0.89	0.88	1082.10	821.78	1339.64	975.91	340.35	1449.91	*vWA*
	0.86	0.74	0.79	0.85	0.85	864.07	954.08	776.64	536.56	1142.77	*D3*
	0.88	0.40	0.87	0.85	0.93	0.74	1915.78	1201.83	246.50	1653.91	*D16*
	0.99	0.70	0.92	0.88	0.96	0.86	0.89	952.27	380.72	1353.25	*D21*
	0.41	0.67	0.48	0.54	0.36	0.64	0.20	0.43	821.00	750.02	*D8*
	0.92	0.76	0.93	0.89	0.94	0.83	0.81	0.94	0.56	2196.65	*D18*
α	5.53	5.99	6.15	7.01	7.64	8.25	9.10	8.92	10.19	10.18	
σ^2	596.53	730.29	1002.93	1236.31	1146.78	1331.79	1821.04	1797.39	1854.94	3208.73	

For each of the simulated peak area vectors, $\boldsymbol{M}_i^*$, we found the pair of profiles maximising the likelihood, $\hat{\boldsymbol{G}}_i = (\hat{G}_{i1}, \hat{G}_{i2})$, using the approach of (Tvedebrink et al., 2010, Chapter 5 of this thesis) and computed T and H associated with $\hat{\boldsymbol{G}}_i$. Using these quantities, we can determine the Mahalanobis distance,

$$M_d(\boldsymbol{M}_i^*, \hat{\boldsymbol{G}}_i) = (\boldsymbol{M}_i^* - \hat{\boldsymbol{M}}_{\hat{\boldsymbol{G}}_i})^\top \mathrm{Var}(\boldsymbol{M}_{\hat{\boldsymbol{G}}_i})^{-1} (\boldsymbol{M}_i^* - \hat{\boldsymbol{M}}_{\hat{\boldsymbol{G}}_i}), \tag{4.2}$$

where $\hat{\boldsymbol{M}}_{\hat{\boldsymbol{G}}_i}$ and $\mathrm{Var}(\boldsymbol{M}_{\hat{\boldsymbol{G}}_i})$ are the expected peak areas and variance assuming a mixture of $\hat{\boldsymbol{G}}_i$ respectively. If $\hat{\boldsymbol{G}}_i$ were equal to the true profiles of the mixture, then M_d would follow a χ^2_n-distribution with n being the number of observations in the mixture. However, the true mixture profiles may not always be identical to the pair of profiles maximising the likelihood. This may be due to stochastic variations and systematic components, e.g. stutter and pull-up effects. The former is caused by artefacts in the polymerase chain reaction resulting in an increase of peak intensities typically in the allelic position before the true allele. Pull-up effects are manifested as an increase of the true peaks caused by overlap of the spectra of the light emitted from the various fluorochromes, which are detected by a CCD camera in the data generating process (Butler, 2005). Hence, on average we expect M_d for $\hat{\boldsymbol{G}}_i$ to be smaller than for the true profiles which implies fewer degrees of freedom in the χ^2-distribution. Fig. 4.6 shows a histogram of 1,000 simulated Mahalanobis distances for the data given in Table 4.2. The superimposed curves indicate that the expectation of fewer than n degrees of freedom for the χ^2-distribution is reasonable, where $n = 31$ in this example. The hypothesis that the Mahalanobis distance follows a χ^2_{29}-distribution is supported by a Kolmogorov-Smirnoff test (p-value of 0.2410), whereas both 30 and 31 degrees of freedom are rejected (p-values are 0.0307 and 1.966×10^{-8}, respectively).

In crime casework the DNA may be degraded or partly degraded, which implies that results only are obtained for short STR loci/alleles (loci/alleles with low base pair numbers), but not (or weak results) with longer STR loci/alleles (loci/alleles for high base pair numbers). This is a potential problem since this is not incorporated in the model due to the assumptions on inter-loci correlation.

However, the Mahalanobis distance M_d in (4.2) can be decomposed into two parts evaluating the quality of the sample, $M_d^{(q)}$ in (4.3), and the goodness of fit of a proposed mixture $\boldsymbol{G} = (G', G'')$ of two profiles, $M_d^{(m)}$ in (4.4). Let $\Sigma_{\boldsymbol{M}|\boldsymbol{M}_+} = \mathrm{Var}(\boldsymbol{M}_{\boldsymbol{G}}|\boldsymbol{M}_+)$ and $\Sigma_{\boldsymbol{M}_+} = \mathrm{Var}(\boldsymbol{M}_{+,\boldsymbol{G}})$, then

$$M_d^{(q)}(\boldsymbol{M}, \boldsymbol{G}) = (\boldsymbol{M}_+ - \boldsymbol{\mu}_{+,\boldsymbol{G}})^\top \Sigma_{\boldsymbol{M}_+}^{-1} (\boldsymbol{M}_+ - \boldsymbol{\mu}_{+,\boldsymbol{G}}), \tag{4.3}$$

$$M_d^{(m)}(\boldsymbol{M}, \boldsymbol{G}) = (\boldsymbol{M} - \boldsymbol{\mu}_{\boldsymbol{G}|+})^\top \Sigma_{\boldsymbol{M}|\boldsymbol{M}_+}^{-} (\boldsymbol{M} - \boldsymbol{\mu}_{\boldsymbol{G}|+}), \tag{4.4}$$

where $\boldsymbol{M}_+$ is the vector of loci peak area sums and $\boldsymbol{\mu}_{\boldsymbol{G}|+}$ ($\boldsymbol{\mu}_{+,\boldsymbol{G}}$) are the expected peak areas (sums) conditioned on the loci sums for profiles $\boldsymbol{G}$. The reason for this decomposition follows from the normality assumption, where $f(\boldsymbol{M}) = f_{\boldsymbol{M}|+}(\boldsymbol{M}|\boldsymbol{M}_+) f_+(\boldsymbol{M}_+)$, which in density functions yields

$$|\Sigma_{\boldsymbol{M}}|^{-\frac{1}{2}} e^{-\frac{1}{2} M_d(\boldsymbol{M}, \boldsymbol{G})} = |\Sigma_{\boldsymbol{M}|\boldsymbol{M}_+}|^{-\frac{1}{2}} |\Sigma_{\boldsymbol{M}_+}|^{-\frac{1}{2}} e^{-\frac{1}{2}\left\{M_d^{(m)}(\boldsymbol{M}, \boldsymbol{G}) + M_d^{(q)}(\boldsymbol{M}, \boldsymbol{G})\right\}},$$

where $\Sigma_{\boldsymbol{M}} = \mathrm{Var}(\boldsymbol{M}_{\boldsymbol{G}})$. We note that $\boldsymbol{M}|\boldsymbol{M}_+$ is a distribution restricted to the affine subspace with fixed peak area sums.

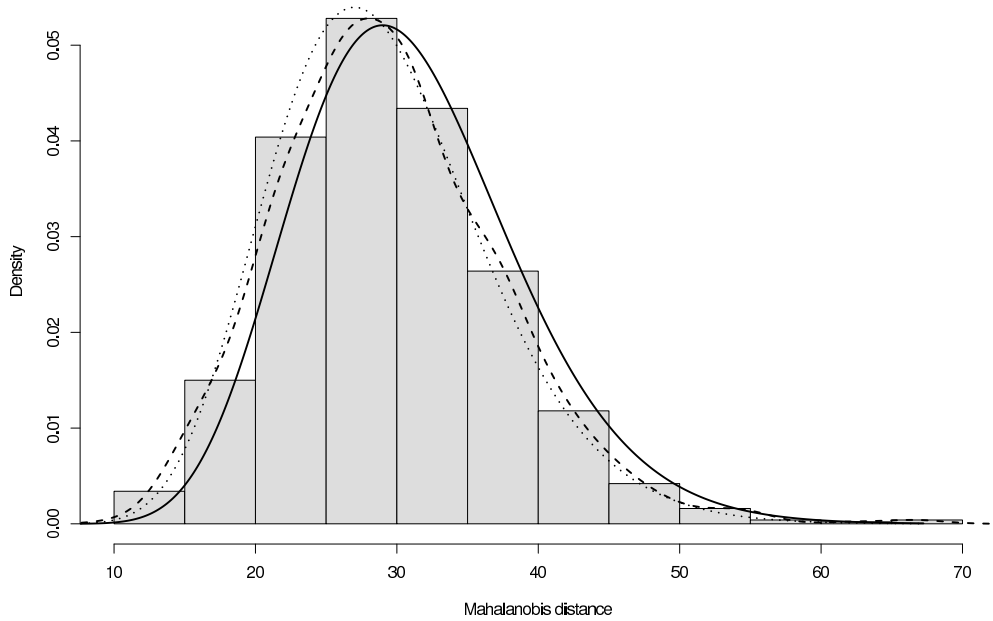

Figure 4.6: Histogram of Mahalanobis distances for simulations based on data from Table 4.2. Superimposed are a χ^2_{31}-distribution (solid), Gaussian based kernel density estimate (dashed) and a χ^2_{29}-distribution (dotted).

Since $|\Sigma_M|^{-\frac{1}{2}} = |\Sigma_{M|M_+}|^{-\frac{1}{2}}|\Sigma_{M_+}|^{-\frac{1}{2}}$, taking -2 log on both sides of the equation gives the decomposition of the Mahalanobis distance (4.2) into the two parts (4.3) and (4.4). Both Mahalanobis distances, $M_d^{(q)}(\boldsymbol{M},\boldsymbol{G})$ and $M_d^{(m)}(\boldsymbol{M},\boldsymbol{G})$, follow χ^2-distributions with S and $n-S$ degrees of freedom, respectively.

In Fig. 4.7, we have plotted histograms of the p-values for $M_d^{(m)}$ and $M_d^{(q)}$ for 66 real crime cases made available by The Section of Forensic Genetics, University of Copenhagen. In all cases the contributors are not known for certain. However, the circumstances of the crime cases made a victim and suspect profile available for each case. The two profiles matched and completely explained the mixed profile the stain.

The left panel shows the histogram of the p-values from $M_d^{(m)}$ assessing how well the proposed pair of profiles matched the mixture given the assumptions of the model. The histogram of the p-values indicated that the model is applicable to STR results in real crime cases, since large p-values, or equivalently small Mahalanobis-distances, imply that H_p is supported by the evidence. The right panel of Fig. 4.7 shows that more than half (35 cases) of the p-values from the test of the sample quality were less than 0.01. This indicates that most of the crime case samples had been subject to degradation of the DNA material. Degradation of the DNA is often complicating the interpretation of DNA mixtures. It is worth emphasising that imbalances caused by degraded DNA may imply that no pair of profiles has $M_d \leq \chi^2_{n,(1-\alpha)}$, where $\chi^2_{k,(1-\alpha)}$ is the critical value on significance level α (e.g. $\alpha = 0.01$) for a χ^2_k-distributed variable. However, conditioned on

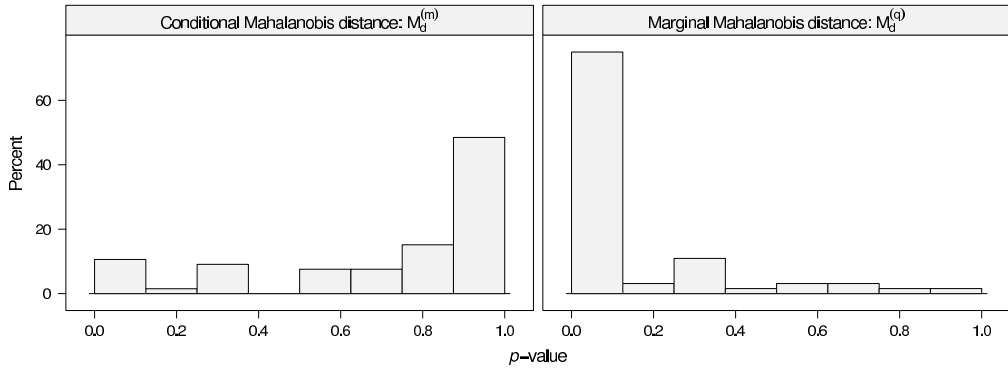

Figure 4.7: Histogram of p-values of the Mahalanobis distances of 66 crime cases in which we had found the pair of profiles maximising the likelihood. For these profiles, we have decomposed the overall Mahalanobis distance M_d into $M_d^{(m)}$ and $M_d^{(q)}$.

the loci sums, such imbalances do not affect the evaluation of a particular pair of profiles, i.e. $M_d^{(m)} \leq \chi^2_{n-S,(1-\alpha)}$ is possible.

In order to investigate whether an observed stain may originate from a two-person mixture, the evaluation of $M_d^{(m)}(\boldsymbol{M}, \hat{\boldsymbol{G}})$ needs to be less than $\chi^2_{n-S,(1-\alpha)}$. If this is not the case for the observed stain, it may be a mixture of more than two contributors or the results are strongly influenced by DNA degradation, drop-outs, stutters, pull-up effects, etc. With $M_d^{(m)}(\boldsymbol{M}, \hat{\boldsymbol{G}}) \leq \chi^2_{n-S,(1-\alpha)}$, it is plausible for the observed stain to be a mixture of two individuals since, for the pair of profiles maximising the likelihood, the conditional Mahalanobis distance is sufficiently small. Then the quality of the sample may be investigated by evaluating $M_d^{(q)}$ and observing if it falls above the critical value $\chi^2_{S,(1-\alpha)}$, e.g. $\alpha = 0.01$. If so, this indicates unexpected imbalances between loci, which may be due to e.g. degraded DNA, inhibitors affecting only certain loci or allelic drop-outs.

4.5.2 Model reductions

When fitting the parameters of the model, we find for our specific data set that the additional variance components, $\tau_s, s \in \mathcal{S}$, were infinitesimally small compared to the contributions of ν_{ss}. A χ^2-test indicated that the goodness of fit was not significantly improved by this parameter. Hence, the results reported in Table 4.5 corresponded to the model with $\tau_s = 0$ for all $s \in \mathcal{S}$. Investigations showed that further reduction of the covariance structure was not supported by the data (see Appendix 4.C for more details).

4.6 Conclusion

In the example of Section 4.3.1, the usual evaluation of the likelihood by considering $LR_{\mathcal{G}} = P(\mathcal{G}|H_p)/P(\mathcal{G}|H_d)$ gave a likelihood ratio supporting the H_p-hypothesis with a likelihood ratio larger than 10^{13}. However, when including the quantitative information, the weight of evidence was decreased to a likelihood ratio, $LR_{\mathcal{GQ}}$, less than one. This was true even with limits of ± 0.25 for the mixture proportion balances in the setup of Bill et al. (2005). The likelihood ratio without taking the quantitative information into account corresponded to the situation, where all combinations passing the guidelines of Bill et al. (2005) were given identical weights. Hence, excluding possible combinations from entering the likelihood ratio based on the quantitative information was not sufficient for an accurate estimate of the likelihood ratio based on quantitative information.

For cases where the qualitative results strongly support that the suspect contributed to a mixed stain, the inclusion of the quantitative information may further support the conclusion. Conversely, the likelihood ratio may decrease supporting the H_d-hypothesis. Both situations were demonstrated by the example of Section 4.3.1. Hence, the evaluation of the quantitative information using a statistical model is of great importance in order to assess the weight of evidence obtained from DNA mixtures.

The model derived in this paper incorporates both information on qualitative traits (STR alleles) and on quantitative aspects of the STR alleles (peak heights and areas). Graphical diagnostics (not included in this manuscript) indicate that the model is well suited for the evaluation of $P(\mathcal{Q}|\mathcal{G}, H)$. Furthermore, assuming independence of the peak areas of the various STR is a simplification that cannot be supported by the work carried out in this paper. Hence, inter-locus correlations or other means of correction need to be considered when assessing the weight of evidence from quantitative data in forensic DNA STR settings.

The concordance between the model properties and prior knowledge of differences in amplification efficiency of various STR loci and in emission intensities of various fluorescent dyes adds further support to the model.

The model described in the present paper is also applicable in other fields of science. A useful property is the handling of variable dimension of the observations while exploiting compound symmetries (Votaw, 1948). For example similar problems with modelling covariance structures may arise in animal breeding studies, where the litter size varies and offsprings may be related through the same breeding lines.

Appendices

4.A The model

In this section, we provide more mathematical details than given in Section 4.2.2. The model assumes proportionality of the mean and variance of $\boldsymbol{A} \sim \mathcal{N}_{4S}(\boldsymbol{\mu}, \Delta)$. The covariance, Δ, is a diagonal matrix with elements $\sigma_s^2 H^{(k)}$ and $\boldsymbol{\mu}$ is a vector partitioned in a similar way with the element $\alpha_s H^{(k)}$ for both peak areas associated with locus s and person k.

The observable peak area measurements, $\boldsymbol{M}$, were defined as a linear transformation, T, such that $\boldsymbol{M} = T\boldsymbol{A} + \boldsymbol{\varepsilon}$. In order to model the proportionality of the mean and variance of $\boldsymbol{M}$, we defined the scaled residuals $\tilde{\boldsymbol{\varepsilon}} = \left(\varepsilon_i / \sqrt{h_i}\right)_{i=1}^{n}$, where $n = \sum_{s\in\mathcal{S}} n_s$. For $\tilde{\boldsymbol{\varepsilon}}$, we assumed a compound symmetry covariance matrix $\tilde{\Omega}$ (Votaw, 1948). Since $\tilde{\boldsymbol{\varepsilon}} = \text{diag}(\boldsymbol{h})^{-1/2}\boldsymbol{\varepsilon}$, the covariance of $\boldsymbol{\varepsilon}$ is $\text{Cov}(\boldsymbol{\varepsilon}) = \Omega = \text{diag}(\boldsymbol{h})^{1/2}\tilde{\Omega}\text{diag}(\boldsymbol{h})^{1/2}$. We parametrised the covariance, $\tilde{\Omega}$, as an additive structure using $\Lambda = \{\nu_{st}\}_{s,t\in\mathcal{S}}$ and $\boldsymbol{\tau} = (\tau_s)_{s\in\mathcal{S}}$, such that $\text{Cov}(\tilde{\boldsymbol{\varepsilon}}_s, \tilde{\boldsymbol{\varepsilon}}_t) = \nu_{st}\mathbf{1}_{n_s}\mathbf{1}_{n_t} + \delta_{st}\tau_s I_{n_s}$, where $\tilde{\boldsymbol{\varepsilon}}_s$ are the scaled residuals of locus s, $\mathbf{1}_k$ is a k-dimensional vector of ones, and δ_{st} is the Kronecker delta. For implementation of the EM-algorithm, we need the conditional distribution of $\boldsymbol{A}|\boldsymbol{M}$. Using Lauritzen (1996, Proposition C.5), this is

$$\boldsymbol{A}|\boldsymbol{M} \sim \mathcal{N}_{4S}\left\{\boldsymbol{\mu} + \Delta T^\top \left(T\Delta T^\top + \Omega\right)^{-1}(\boldsymbol{M} - T\boldsymbol{\mu}), \Delta - \Delta T^\top \left(T\Delta T^\top + \Omega\right)^{-1} T\Delta\right\}. \tag{4.5}$$

The model for $\boldsymbol{M}$ corresponds to a linear mixed effects model:

$$\boldsymbol{M} = X\boldsymbol{\alpha} + Z(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)^\top, \quad \text{where } \boldsymbol{\xi}_1 \sim \mathcal{N}(\mathbf{0}, \text{diag}(\mathbf{1}_4\sigma_s^2)_{s\in\mathcal{S}}) \text{ and } \boldsymbol{\xi}_2 \sim \mathcal{N}(\mathbf{0}, \Omega) \tag{4.6}$$

for some case specific design matrices X and Z. However, estimation of the variance components are complicated due to the varying dimensions of $\boldsymbol{M}$ and $\boldsymbol{M}_s$, $s \in \mathcal{S}$ from case to case.

4.B EM-estimators

In order to handle the complete structure of $\boldsymbol{A}$ that includes the missing data problem, we used the EM-algorithm to impute the unobservable data. However, since the dimensions of $\boldsymbol{M}$ and sub-vectors hereof varied from case to case, we obtained a likelihood that was not very well suited for implementation of the EM-algorithm. This was due to the dependence on n_s in the covariance of the locus-wise average of the scaled residuals $\bar{\tilde{\boldsymbol{\varepsilon}}} = (\bar{\tilde{\boldsymbol{\varepsilon}}}_1, \ldots, \bar{\tilde{\boldsymbol{\varepsilon}}}_S)$,

$$\text{Cov}(\bar{\tilde{\boldsymbol{\varepsilon}}}) = \text{diag}(\tau_s/n_s)_{s\in\mathcal{S}} + \Lambda = \text{diag}(\boldsymbol{\tau}/\boldsymbol{n}) + \Lambda,$$

where $\boldsymbol{n} = (n_s)_{s\in\mathcal{S}}$ and the vector division is done component-wise, $\boldsymbol{x}/\boldsymbol{y} = (x_i/y_i)_{i=1}^n$.

The problem was solved using appropriate auxiliary variables $\boldsymbol{v}$ and $\boldsymbol{u}$, which we assumed to be independent and zero-mean normal distributed variables with covariances Λ and $\text{diag}(\boldsymbol{\tau}/\boldsymbol{n})$, respectively. By introducing $\boldsymbol{v}$ and $\boldsymbol{u}$, we obtained a likelihood of a full exponential family, where the estimation of $\boldsymbol{\tau}$ and Λ may be done separately. The use of auxiliary variables is

equivalent to adding constraints on the diagonal elements of Λ. By assuming $\mathrm{Cov}(\boldsymbol{v}) = \Lambda$, we get the constraint that $v_{ss} > 0$, $s \in \mathcal{S}$. In (4.6), this corresponds to splitting $\boldsymbol{\xi}_2$ into two independent parts $\boldsymbol{\xi}_{21}$ and $\boldsymbol{\xi}_{22}$, $\boldsymbol{\xi}_2 = \boldsymbol{\xi}_{21} + \boldsymbol{\xi}_{22}$, where $\boldsymbol{\xi}_{21} \sim \mathcal{N}(\mathbf{0}, Q_c \Lambda Q_c^\top)$ and $\boldsymbol{\xi}_{22} \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\tau_s \mathbf{1}_{n_s})_{s\in\mathcal{S}})$ with Q_c defined in (4.7).

Hence, the E-step consisted of imputing $\boldsymbol{A}$, $\boldsymbol{u}$ and $\boldsymbol{v}$ given the observations $\boldsymbol{M}$. In the M-step, we used that the full likelihood factorises into two terms modelling the biological part of the data given the measurement noise, $(\boldsymbol{A}, \boldsymbol{M})|(\boldsymbol{u}, \boldsymbol{v}, \tilde{\varepsilon})$, and the noise, $(\boldsymbol{u}, \boldsymbol{v}, \tilde{\varepsilon})$, respectively:

$$f(\boldsymbol{A}, \boldsymbol{M}, \boldsymbol{u}, \boldsymbol{v}, \tilde{\varepsilon}; \Delta, \boldsymbol{\mu}, \boldsymbol{\tau}, \Lambda | \boldsymbol{H}, \boldsymbol{h}) = g(\boldsymbol{A}, \boldsymbol{M}; \Delta, \boldsymbol{\mu} | \boldsymbol{u}, \boldsymbol{v}, \tilde{\varepsilon}, \boldsymbol{H}, \boldsymbol{h}) h(\boldsymbol{u}, \boldsymbol{v}, \tilde{\varepsilon}; \boldsymbol{\tau}, \Lambda | \boldsymbol{H}, \boldsymbol{h})$$

with g and h being the density functions of the two multivariate normal distributions below:

$$g: \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{M} \end{pmatrix} \Big| \varepsilon \sim \mathcal{N}\left\{ \begin{pmatrix} \boldsymbol{\mu} \\ T\boldsymbol{\mu} + \varepsilon \end{pmatrix}, \begin{bmatrix} \Delta & \Delta T^\top \\ T\Delta & T\Delta T^\top \end{bmatrix} \right\}$$

$$h: \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \\ \tilde{\varepsilon} \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \mathrm{diag}(\boldsymbol{\tau}/\boldsymbol{n}) & O & \mathrm{diag}(\boldsymbol{\tau}/\boldsymbol{n}) Q_c^\top \\ O & \Lambda & \Lambda Q_c^\top \\ Q_c \mathrm{diag}(\boldsymbol{\tau}/\boldsymbol{n}) & Q_c \Lambda & \tilde{\Omega} \end{bmatrix} \right\},$$

where Q_c is defined in (4.7). In order to derive the estimators of the parameters entering the functions g and h, we defined two matrices Q and Q_c,

$$Q = \begin{bmatrix} \mathbf{1}_4 & \dots & O \\ \vdots & \ddots & \vdots \\ O & \dots & \mathbf{1}_4 \end{bmatrix} \quad \text{and} \quad Q_c = \begin{bmatrix} \mathbf{1}_{n_{1c}} & \dots & O \\ \vdots & \ddots & \vdots \\ O & \dots & \mathbf{1}_{n_{Sc}} \end{bmatrix}, \tag{4.7}$$

where subscript c refers to case c, $c = 1, \dots, C$. Furthermore, the DNA proxy $H = (H^{(1)}, H^{(2)})$ is expanded to a $4S$-dimensional vector, $\boldsymbol{H} = (\boldsymbol{H}_s)_{s\in\mathcal{S}}$, where the components $\boldsymbol{H}_s$ are fixed for all loci, $\boldsymbol{H}_s = (H^{(1)}, H^{(1)}, H^{(2)}, H^{(2)})$. Note, that the compound symmetry structure of the covariance of $\tilde{\varepsilon}$ with $\boldsymbol{\tau} = \mathbf{0}$ can be written as $\tilde{\Omega} = Q_c \Lambda Q_c^\top$. The estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}^2$ can be found as

$$\hat{\boldsymbol{\alpha}} = \frac{\sum_c Q^\top \mathrm{E}(\boldsymbol{A}_c | \boldsymbol{M}_c)}{\sum_c Q^\top \boldsymbol{H}_c}$$

$$\hat{\boldsymbol{\sigma}}^2 = (4C - 1)^{-1} \sum_c Q^\top \left[\frac{\{\mathrm{E}(\boldsymbol{A}_c | \boldsymbol{M}_c) - \boldsymbol{\mu}_c\}^2 + \mathrm{diag}\{\mathrm{Cov}(\boldsymbol{A}_c | \boldsymbol{M}_c)\}}{\boldsymbol{H}_c} \right],$$

where the squaring of a vector is done component-wise, $\boldsymbol{x}^2 = (x_i^2)_{i=1}^n$ and $\mathrm{diag}\{B\}$ extract the diagonal vector of B, $\mathrm{diag}\{B\} = (B_{ii})_{i=1}^n$. Furthermore, the moments of $\boldsymbol{A}_c | \boldsymbol{M}_c$ are given in (4.5). The estimators of $\boldsymbol{\tau} = (\tau_s)_{s\in\mathcal{S}}$ and Λ are,

$$\begin{aligned} \hat{\tau}_s &= n_{s+}^{-1} \sum_c \left\{ \mathrm{E}(u_{sc}^2 | \boldsymbol{M}_c) n_{sc} + \mathrm{E}(\tilde{\varepsilon}_s^\top \tilde{\varepsilon}_s - n_s \tilde{\varepsilon}_s^2 | \boldsymbol{M}_c) \right\} \\ \hat{\Lambda} &= C^{-1} \sum_c \left\{ \mathrm{E}(\boldsymbol{v}_c | \boldsymbol{M}_c) \mathrm{E}(\boldsymbol{v}_c | \boldsymbol{M}_c)^\top + \mathrm{Cov}(\boldsymbol{v}_c | \boldsymbol{M}_c) \right\}. \end{aligned}$$

For both $\boldsymbol{v}$ and $\boldsymbol{u}$, the covariance with $\boldsymbol{M}$ is expressed as $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{M}) = \mathrm{Cov}(\boldsymbol{x}) Q_c^\top \mathrm{diag}(\boldsymbol{h})^{1/2}$, for $\boldsymbol{x}$ replaced by $\boldsymbol{v}$ or $\boldsymbol{u}$. The conditional moments entering the estimation

equations may be found using the formulae for computing conditional moments in the multivariate normal distribution, $\mathrm{E}(\boldsymbol{X}|\boldsymbol{Y}) = \boldsymbol{\mu_X} + \Theta_{12}\Theta_{22}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu_Y})$ and $\mathrm{Cov}(\boldsymbol{X}|\boldsymbol{Y}) = \Theta_{11} - \Theta_{12}\Theta_{22}^{-1}\Theta_{21}$ for $(\boldsymbol{X}, \boldsymbol{Y})^\top \sim \mathcal{N}((\boldsymbol{\mu_X}, \boldsymbol{\mu_Y})^\top, \Theta)$ with $\Theta = \left[\begin{smallmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{smallmatrix}\right]$ (Lauritzen, 1996, Proposition C.5).

4.C Model reduction

As mentioned in Section 4.4, the large asymptotic standard deviations indicated that the covariance structure of $\tilde{\Omega}$ could be simplified. The estimated $\boldsymbol{\tau}$ parameters for nearly all loci were negligible compared to ν_{ss}. Let $\mathrm{Diag}(A_i)_{i=1}^n$ be a block-diagonal matrix with matrices A_i, $i = 1, \ldots, n$ as elements and the square root of a vector defined as $\sqrt{\boldsymbol{x}} = (\sqrt{x_i})_{i=1}^n$. Then, we may write the covariance matrix of $\boldsymbol{M}$, $\Sigma_{\boldsymbol{M}}$, as:

$$\Sigma_{\boldsymbol{M}} = T\Delta T^\top + \Omega = \mathrm{Diag}\left(\sigma_s^2 T_s \mathrm{diag}(\boldsymbol{H}_s) T_s^\top + \tau_s \sqrt{\boldsymbol{h}_s}\sqrt{\boldsymbol{h}_s}^\top\right)_{s\in\mathcal{S}} + \left[\nu_{st}\sqrt{\boldsymbol{h}_s}\sqrt{\boldsymbol{h}_t}^\top\right]_{s,t\in\mathcal{S}}.$$

From the equation above, we see that setting $\boldsymbol{\tau} = \boldsymbol{0}$ does not introduce any singularities in $\Sigma_{\boldsymbol{M}}$. Hence, the asymptotic theory is not violated. In order to test whether $\boldsymbol{\tau}$ was statistically significant, we used an approximately χ^2-distributed test-statistic with the difference in parameters as degrees of freedom (Cox and Hinkley, 1974). In the full model, there were $S(S+3)/2$ parameters. By restricting $\boldsymbol{\tau} = \boldsymbol{0}$, we removed S parameters and the χ_S^2-test yielded a p-value of 0.9999 supporting the hypothesis of $\boldsymbol{\tau} = \boldsymbol{0}$. The reported parameter estimates in Table 4.5 were based on this restricted model.

Data exploration and the estimated parameters of Λ from Table 4.5 suggest that further model reductions may be feasible. Possible parametrisations of $\tilde{\Omega}$ may be,

$$\mathrm{Cov}(\tilde{\varepsilon}_s, \tilde{\varepsilon}_t) = \nu_{d(s),d(t)}\mathbf{1}_{n_s}\mathbf{1}_{n_t}^\top + \delta_{st}\tau_s I_{n_s} \tag{4.8}$$

$$\mathrm{Cov}(\tilde{\varepsilon}_s, \tilde{\varepsilon}_t) = \nu_{d(s),d(t)}\mathbf{1}_{n_s}\mathbf{1}_{n_t}^\top + \delta_{d(s)d(t)}\tau_{d(s)} I_{n_s} \tag{4.9}$$

$$\mathrm{Cov}(\tilde{\varepsilon}_s, \tilde{\varepsilon}_t) = \nu\mathbf{1}_{n_s}\mathbf{1}_{n_t}^\top + \delta_{st}\tau_s I_{n_s}, \tag{4.10}$$

where d maps locus to fluorescence dye colour, e.g. $d(\mathrm{FGA}) = \mathrm{Yellow}$. The covariance structures in (4.8)-(4.10) all use fewer parameters in $\tilde{\Omega}$ than the restricted model with $D(D+1)/2+S$, $D(D+3)/2$ and $1+S$ parameters, respectively, where D is the number of dye colours. In our data $D = 3$ and $S = 10$ and thus we removed 39, 46 and 44 parameters, respectively. The three tests indicated that there were significant differences between the full model and any of the reduced models, all with p-values < 0.0001. Hence, the model with the best fit included locus dependent parameters for the between and within covariance on the measurement errors. Inspection of the correlation matrix in Table 4.5 indicated that locus D8 was the only locus with an average between-locus-correlation less than 0.5. This may well cause the dye covariance models to have a poor fit.

However, one has to bear in mind that the parameter estimates were based on a limited training set. Hence, the rejections of the hypotheses of simpler models may be biased towards the four profiles included in the training set. In order to fully verify the model we need to increase the proportion of alleles from each locus and also the number of homozygous profiles. This will

reduce the possible individual specific effect that may exist in the training set. Such work is in progress.

A more detailed description of the model and the implementation of the EM-algorithm with full R-source code are available on line at http://people.math.aau.dk/~tvede/dna. The programs can also be obtained from http://www.blackwellpublishing.com/rss.

Acknowledgements

The authors would like to thank Prof. Bruce S. Weir, University of Washington, for some clarifying comments on an earlier version of the manuscript. We also thank Ms. Catharina Steentoft for collecting the DNA profiles from the crime case work used in Section 4.5.1, and Ms. Lisbeth Grubbe Nielsen for thorough review of language and grammar. Furthermore, very helpful comments were made by the journal's editors and anonymous reviewers. The 22nd ISFG Congress-proceedings (Tvedebrink et al., 2008) has a brief model description.

Bibliography

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

Bill, M. et al. (2005). PENDULUM - a guideline-based approach to the interpretation of STR mixtures. *Forensic Science International 148*, 181–189.

Butler, J. M. (2005). *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* (2 ed.). Burlington, MA: Elsevier Academic Press Inc., U.S.

Cowell, R. G. (2009). Validation of an STR peak area model. *Forensic Science International: Genetics 3*(3), 193–199.

Cowell, R. G., S. L. Lauritzen, and J. Mortera (2007a). A gamma model for DNA mixture analyses. *Bayesian Analysis 2*(2), 333–348.

Cowell, R. G., S. L. Lauritzen, and J. Mortera (2007b). Identification and separation of DNA mixtures using peak area information. *Forensic Science International 166*, 28–34.

Cowell, R. G., S. L. Lauritzen, and J. Mortera (2010). Probabilistic expert systems for handling artifacts in complex DNA mixtures. *Forensic Science International: Genetics*. In Press.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. Chapman and Hall Ltd.

Curran, J. M. (2008). A MCMC method for resolving two person mixtures. *Science & Justice 48*, 168–177.

Evett, I. W. and B. S. Weir (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sunderland, MA: Sinauer Associates.

Gill, P. D. et al. (1998). Interpreting simple STR mixtures using allele peak areas. *Forensic Science International 91*(1), 41–53.

Gill, P. D. et al. (2006). DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International 160*(2-3), 90–101.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

Little, R. and D. Rubin (2002). *Statistical Analysis with missing data* (2 ed.). Wiley.

Perlin, M. W. and B. Szabady (2001). Linear mixture analysis: A mathematical approach to resolving mixed DNA samples. *Journal of Forensic Science 46*(6), 1372–1378.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2008). Amplification of DNA mixtures - Missing data approach. *Forensic Science International: Genetics Supplement Series 1*, 664–666.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2009). Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics 3*(4), 222–226.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2010). Identifying contributors of DNA mixtures by of quantitative information of STR typing. *Journal of Computational Biology*. Accepted for publication.

Votaw, D. F. (1948). Testing compound symmetry in a normal multivariate distribution. *Annals of Mathematical Statistics 19*(4), 447–473.

Wang, T., N. Xue, and J. D. Birdwell (2006). Least-square deconvolution: A framework for interpreting short tandem repeat mixtures. *Journal of Forensic Science 51*(6), 1284–1297.

4.7 Supplementary remarks

As briefly mentioned at page 79, the model presented above is a case of the larger class of linear mixed effects models. However, what distinguishes the model from other types of linear mixed effects models, is the property of handling varying dimensions of the observation matrix and subvectors hereof under the assumed mean and covariance structure. Typically an experimental design is set up such that n_s and n (as defined above) are constant over the various factors of the experiment. In order to construct interesting and realistic experiment useful to forensic genetics it is not possible to fulfil such restrictions. However, by restricting the intra-locus correlations to be positive, the EM-algorithm may be used to fit the model to data where the subvectors of the response vary across samples.

The model extends the *LR* by including the quantitative information in the evidence calculations. By evaluating $L(\boldsymbol{M}|\boldsymbol{G})$ for a given pair of DNA profiles, $\boldsymbol{G}$, it is possible to assess the goodness-of-fit for a proposed pair of DNA profiles versus the observed peak intensities. However, since the model presented above assumes intra-locus correlations, it is very time consuming and computational intense to search for a pair of best matching profiles $\hat{\boldsymbol{G}} = \max_{\boldsymbol{G}} L(\boldsymbol{M}|\boldsymbol{G})$, since the configuration on the various loci affect each other through the non-zero correlations.

Hence, in order to perform such a task, we need to relax some of the assumptions for fast computation and evaluation. In the following chapter we present a statistical model and an efficient algorithm for finding a pair of best matching profiles. The basic assumptions are similar to those discussed above, with the difference that the peak intensities within each locus is assumed conditionally independent. That is, by conditioning on an ancillary statistic (for the mixture ratio) we assume that the configuration of the DNA profiles in locus s is independent of configurations in locus t for all $t \neq s$.

The methodology differs from previous approaches since it is frequentistic and based on a statistical model taking the present proportionality of mean and variance of the peak intensities into account. There are several Bayesian methods for modelling and separating DNA mixtures, e.g. Cowell et al. (2007a,b, 2010); Cowell (2009) discussed the use of probabilistic expert systems to model DNA mixtures using first a normal distribution (2007a-paper) and later a gamma distribution, and also Curran (2008) took a Bayesian approach and modelled the peak intensities using a multivariate normal distribution. However, Curran (2008) did not include the proportionality of the mean and variance, which is a intrinsic feature of the gamma models of Cowell et al.

Earlier Perlin and Szabady (2001) and Wang et al. (2006) used linear models to model the peak intensities of DNA mixtures using a frequentistic approach. However, their models did not take the mentioned proportionalities of the first two moments into account, and their methods did not allow for efficient and consistent modelling of all loci simultaneously. For example, Wang et al. (2006) did not incorporate a common mixture ratio across loci even though there are strong biological and biochemical arguments for this assumption. Furthermore, did the method of Wang et al. (2006) call for a reasonably large amount of manual labour in order to use the output from their method.

CHAPTER 5

Identifying contributors of DNA mixtures by means of quantitative information of STR typing

Publication details

Co-authors: Poul Svante Eriksen[*], Helle Smidt Mogensen[†] and Niels Morling[†]

[*] *Department of Mathematical Sciences*
Aalborg University
[†] *Section of Forensic Genetics, Department of Forensic Medicine*
Faculty of Health Science, University of Copenhagen

Journal: Journal of Computational Biology (Accepted for publication)

Abstract:
Estimating the weight of evidence in forensic genetics is often done in terms of a likelihood ratio, *LR*. The *LR* evaluates the probability of the observed evidence under competing hypotheses. Most often probabilities used in the *LR* only consider the evidence from the genomic variation identified using polymorphic genetic markers. However, modern typing techniques supply additional quantitative data, which contain very important information about the observed evidence. This is particularly true for cases of DNA mixtures, where more than one individual has contributed to the observed biological stain.

This paper presents a method for including the quantitative information of STR DNA mixtures in the *LR*. Also, an efficient algorithmic method for finding the best matching combination of DNA mixture profiles is derived and implemented in an on-line tool for two- and three-person DNA mixtures.

Finally, we demonstrate for two-person mixtures, how this best matching pair of profiles can be used in estimating the likelihood ratio using importance sampling. The reason for using importance sampling for estimating the likelihood ratio is the often vast number of combinations of profiles needed for the evaluation of the weight of evidence.

Keywords:
Forensic genetics; STR DNA; DNA mixture; Greedy algorithm; Finding best pair of matching profiles; Importance sampling.

5.1 Introduction

When a crime has been committed, biological traces are often found at the scene of crime. In many cases, more than one individual have contributed to the stain, which is then determined a DNA mixture. The evaluation of DNA mixtures is often complex and laborious taking experienced case workers lots of time and effort to analyse.

Most modern DNA typing techniques are based upon polymerase chain reaction (PCR) producing millions of copies of the DNA string. The amount of DNA in the PCR vessel pre-PCR is reflected in the concentration of target molecules post-PCR. The targets used in forensic genetics are selected such that they are highly polymorphic (large number of possible alleles) which gives a high power of discrimination. Furthermore, the genetic markers used for forensic purposes are non-coding and should ideally be neutral with respect to selection.

The prevalent technology used in forensic genetics to perform genetic identification uses short tandem repeat (STR) polymorphisms. This method relies on variability in the length of certain repeat motifs in the genome. The STR DNA profile is observed via a so called electropherogram (EPG), where the alleles are identified as signal peaks above a signal to noise threshold (shaded cones of Figure 5.2). For a single person DNA profile one can observe either one or two peaks referring to the situation, where the DNA profile is either homozygous (identical alleles on both chromosomes) or heterozygous (different alleles on each chromosome). The commercial kits used for identification purposes typically contain between 10 to 15 genetic markers (also called loci: plural for locus). Within each locus the number of alleles varies from 5 to 20. For the

kit (SGM Plus kit, Applied Biosystems, AB) depicted in Figure 5.2, the labels “D3”, “vWA”, . . . , “FGA” refer to locus names and the integer values above the locus name corresponds to the observed allele types for that particular locus.

It is possible only to observe the cumulative peaks in the EPG. That is, the peak heights are expected to be twice the height for homozygous loci relative to the heterozygous loci, since the two identical alleles doubles the amount of pre-PCR product for the homozygous peaks. This is also true for DNA mixtures where alleles shared by two or more contributors will reflect the contribution from more donors as higher peaks. Hence, for DNA mixtures with two contributors the number of observable peaks ranges from one to four alleles depending on the particular profiles in the mixture.

The kit used for STR typing comprises a set of loci, $\mathcal{S}$, used for discrimination. For an arbitrary two-person mixture the number of possible combinations are given by $1^{S_1}7^{S_2}12^{S_3}6^{S_4}$, where S_i is the number of loci with i observations and $S = \sum_{i=1}^{4} S_i$, is the total number of loci used for discrimination, i.e. S is the size of $\mathcal{S}$. The numbers 1, 7, 12 and 6 comes from the number of possible combinations (see Table 5.1) when observing 1, 2, 3 and 4 alleles, respectively.

Table 5.1: Possible combinations in a two-person mixture with one to four alleles.

Alleles	Possible combinations						
a	(aa, aa)						
a, b	(aa, ab)	(aa, bb)	(ab, aa)	(ab, ab)	(ab, bb)	(bb, ab)	(bb, aa)
a, b, c	(aa, bc)	(ab, ac)	(ab, bc)	(ab, cc)	(ac, ab)	(ac, bb)	
	(ac, bc)	(bb, ac)	(bc, ab)	(bc, ac)	(bc, aa)	(cc, ab)	
a, b, c, d	(ab, cd)	(ac, bd)	(ad, bc)	(bc, ad)	(bd, ac)	(cd, ab)	

In most cases, this leads to an intractable number of combinations. However, using the quantitative STR data (peak heights and peak areas), the number of plausible combinations often decreases substantially. In this paper, we develop a statistical model for STR DNA mixtures. The statistical model is intended to measure the agreement between the expected peak intensities for a proposed combination of DNA profiles and the actual observed peak intensities. Hence, we use an objective criterion to discriminate among the possible combinations in Table 5.1.

In order to incorporate the peak intensities in the likelihood ratio (LR), we first demonstrate how to find a best matching pair of profiles for a given two-person mixture using an efficient algorithmic approach. This algorithm iteratively builds up a best matching combination of profiles using the statistical model for the peak intensities. The algorithm has been implemented in a free on-line tool available at the first author’s web-site. The statistical model and algorithmic construction are different from previously proposed methods for DNA mixture separation (e.g. Perlin and Szabady, 2001; Bill et al., 2005; Wang et al., 2006; Cowell et al., 2007a,b; Curran, 2008).

The inclusion of the quantitative information in the *LR* is done by assigning a weight to each combination of DNA profiles consistent with the observed STR types. The weight reflects the probability of observing the observed peak intensities given a specific combination of profiles. The denominator in the *LR* will in most cases yield a sum over an intractable number of combination. By sampling "close" to the best matching combination returned by the algorithm, we show how importance sampling may be used to estimate the *LR*.

5.2 Data

The model is based on exploration of controlled experiments of two-person mixtures conducted at The Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark. From the data exploration it is evident that the mean and covariance structure of the peak areas must satisfy proportionality of:

- peak areas and peak heights,
- peak area and amount of DNA in the mixture,
- the mean and variance of the peak areas.

These assumptions is supported by Figures 1 and 2 in Tvedebrink et al. (2010). The experiments consisted of pairwise two-person mixtures in various mixture ratios of the four profiles in Table 5.2. The data were prepared as described in Tvedebrink et al. (2009).

Table 5.2: The four STR profiles used in the controlled pairwise two-person mixture experiments.

	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
A	14,18	17,19	12,14	20,24	10,13	30.2,32.2	13,13	12,13	8,9	20,22
B	15,16	14,16	10,12	17,25	13,16	30,30	13,13	14,15	6,9	19,23
C	15,16	15,17	11,11	19,25	8,12	29,31	15,17	13,13	6,8	23,24
D	16,19	15,17	10,12	23,25	13,13	28,30	12,16	13,15	6,7	20,23

5.3 Modelling peak areas of a two-person mixture

For the search of a pair of best matching profiles to be feasible, we assume the peak areas of the various loci to be conditionally independent given the loci area sums, $\boldsymbol{A}_+$. Performing the inference conditioned on $\boldsymbol{A}_+$ satisfies the reasoning of Cox (1958) as $\boldsymbol{A}_+$ is an ancillary statistic for the mixture ratio, i.e. $\boldsymbol{A}_+$ is fixed for all values of the mixture ratio. Furthermore, we assume that the peak areas are multivariate normal distributed with conditional mean vector, $\mathbb{E}(\boldsymbol{A}_s|A_{s,+})$,

and covariance matrix, $\mathbb{C}\text{ov}(\boldsymbol{A}_s|A_{s,+})$, defined as

$$\begin{aligned}\mathbb{E}(\boldsymbol{A}_s|A_{s,+}) &= [\alpha \boldsymbol{P}_{s,1} + (1-\alpha)\boldsymbol{P}_{s,2}]\frac{A_{s,+}}{2} \qquad \text{and} \\ \mathbb{C}\text{ov}(\boldsymbol{A}_s|A_{s,+}) &= \tau^2 C_s \text{diag}(\boldsymbol{h}_s) C_s^\top, \end{aligned} \tag{5.1}$$

where α denotes the proportion with which person 1 contributes to the mixture, and $C_s = I_{n_s} - n_s^{-1}\mathbf{1}_{n_s}\mathbf{1}_{n_s}^\top$ with n_s, $1 \leq n_s \leq 4$, being the number of observed peaks at locus s. Note that α is supposed to be common to all loci. The definition of the covariance matrix is close to the ordinary covariance when conditioning on the vector sum. However, as the variance of the peak area is assumed proportional to the mean, we use the diagonal matrix $\text{diag}(\boldsymbol{h}_s)$, where $\boldsymbol{h}_s$ is the associated peak heights on locus s, to obtain weighted observations that stabilise the variance. Furthermore, τ^2 is a common variance parameter for all loci, $s \in \mathcal{S}$.

The $\boldsymbol{P}_{s,k}$-vector is a vector of indicators taking values 0, 1 or 2 referring to the number of copies that person k has of each allele in the mixture on locus s. E.g., if the two individuals contributing to the mixture have genotypes (10, 12) and (14, 14), respectively, we will have $\boldsymbol{P}_{s,1} = (1,1,0)^\top$ and $\boldsymbol{P}_{s,2} = (0,0,2)^\top$. Assuming no chromosomal anomalies, each individual carries two alleles at each locus which implies the sum of $\boldsymbol{P}_{s,k}$ to be 2 for all k.

The model presented here is different from e.g. the ones of Cowell et al. (2007a,b) and Curran (2008) who both takes a Bayesian approach. The model of Cowell et al. (2007a) assumes the peak heights to be gamma-distributed to ensure proportionality of the mean and variance, whereas Cowell et al. (2007b) assumes normality of the peak heights with parameters chosen to ensure proportionality of mean and variance. As mentioned in Curran (2008), the model of Cowell et al. (2007a) makes a crude adjustment for a repeat number effect, which is no longer relevant. In Curran (2008) the peak heights are assumed multivariate normal, but here no attempt is done in order to ensure proportionality of the mean and variance. Furthermore, by conditioning on the peak area sums within each locus we acknowledge the strong inter-locus correlation. In addition to the methods based on statistical models, there are several methods that rely on heuristics and guidelines (Gill et al., 1998; Clayton et al., 1998; Perlin and Szabady, 2001; Bill et al., 2005; Wang et al., 2006; Gill et al., 2006). Cowell et al. (2007b) gives a nice review of most of these methods in their introductory section.

5.4 Finding best matching pair of profiles

In order to find the most likely pair of profiles matching the observed mixture under the assumptions made by the model, one can decrease the number of possibilities using the following arguments. Let the observed peak areas within each locus, s, be sorted such that $A_{s,(1)} < \cdots < A_{s,(n_s)}$, and assume that $DNA_1 < DNA_2$, where DNA_k is the amount of DNA contributed by person k.

Then, for a locus with four observed peaks ($n_s = 4$), the only likely pair of profiles given the model relate the alleles with peak areas $(A_{s,(1)}, A_{s,(2)})$ and $(A_{s,(3)}, A_{s,(4)})$ to person 1 and person 2, respectively. For loci with one observation ($n_s = 1$), the two individuals need both to be homozygous for the observed allele, while for two ($n_s = 2$) or three observations ($n_s = 3$), the possible profiles are listed in Table 5.3 (the notation $\mathcal{J}_2$ and $\mathcal{J}_3$ is used in Section 5.4.1).

Table 5.3: Possible profiles for loci with two and three observations.

$\mathcal{J}_2$:	$\boldsymbol{P}_{s,1}\boldsymbol{P}_{s,2}$		$\boldsymbol{P}_{s,1}\boldsymbol{P}_{s,2}$		$\boldsymbol{P}_{s,1}\boldsymbol{P}_{s,2}$		$\boldsymbol{P}_{s,1}\boldsymbol{P}_{s,2}$	
$A_{s,(1)}$	1	1	2	0	1	0	0	1
$A_{s,(2)}$	1	1	0	2	1	2	2	1

$\mathcal{J}_3$:	$\boldsymbol{P}_{s,1}\boldsymbol{P}_{s,2}$		$\boldsymbol{P}_{s,1}\boldsymbol{P}_{s,2}$		$\boldsymbol{P}_{s,1}\boldsymbol{P}_{s,2}$		$\boldsymbol{P}_{s,1}\boldsymbol{P}_{s,2}$	
$A_{s,(1)}$	2	0	1	0	1	0	0	1
$A_{s,(2)}$	0	1	1	0	0	1	0	1
$A_{s,(3)}$	0	1	0	2	1	1	2	0

In Table 5.3, $\boldsymbol{P}_{s,1}$ and $\boldsymbol{P}_{s,2}$ refers to the profiles of person 1 and person 2 on the particular locus s, respectively, and the cell values to the number of alleles associated with the profiles. The reason for not considering the three and eight other combinations for loci with two and three observations (Table 5.1), respectively, is that, for any of these combinations, one of the four combinations listed in Table 5.3 will be more likely under the model assumptions, i.e. have a better fit to the observed data. E.g. would $\boldsymbol{P}_{s,1} = (0,2)^\top$ and $\boldsymbol{P}_{s,2} = (2,0)^\top$ be unlikely as we assumed person 1 to have the lowest contribution and the second area to be the larger.

The numbers of possible pairs of profiles for loci with two, three and four observations are respectively 7, 12 and 6, when discarding the information from peak areas and only using combinatorics. Thus, using the assumptions of the model, we decrease the number of profiles which needs to be examined in order to find the most likely profiles forming the observed mixture.

We assume the peak areas to be normally distributed with conditional means and covariances as specified in (5.1). Due to the conditional independence of the loci, the overall estimates of α and τ^2 are found as sums over the loci. Let $W_s = C_s\text{diag}(\boldsymbol{h}_s)C_s^\top$, then we can write the conditional distribution as $\boldsymbol{A}_s|A_{s,+} \sim \mathcal{N}_{n_s}(\alpha\boldsymbol{x}_0^s - \boldsymbol{x}_1^s, \tau^2 W_s)$, where $\boldsymbol{x}_0^s = (\boldsymbol{P}_{s,1} - \boldsymbol{P}_{s,2})A_{s,+}/2$ and $\boldsymbol{x}_1^s = \boldsymbol{P}_{s,2}A_{s,+}/2$ are the terms of the mean, linear and constant in α, respectively. Solving the likelihood equation with respect to α and τ^2 yield the unbiased estimators

$$\begin{aligned} \hat{\alpha} &= \frac{\sum_{s\in\mathcal{S}} \boldsymbol{x}_0^{s\top} W_s^- (\boldsymbol{A}_s - \boldsymbol{x}_1^s)}{\sum_{s\in\mathcal{S}} \boldsymbol{x}_0^{s\top} W_s^- \boldsymbol{x}_0^s} \qquad \text{and} \\ \hat{\tau}^2 &= N^{-1} \sum_{s\in\mathcal{S}} (\boldsymbol{A}_s - \hat{\alpha}\boldsymbol{x}_0^s - \boldsymbol{x}_1^s)^\top W_s^- (\boldsymbol{A}_s - \hat{\alpha}\boldsymbol{x}_0^s - \boldsymbol{x}_1^s), \end{aligned} \tag{5.2}$$

where $N = n_+ - S - 1 = \sum_{s\in\mathcal{S}}(n_s - 1) - 1$ and W_s^- is the generalised inverse of W_s. We have to use the generalised inverse of W_s as W_s has the rank $n_s - 1$. An approximation to this model assumes that the precision matrix, $\tau^{-2}W_s^{-1}$, is given by $\tau^{-2}C_s\text{diag}(\boldsymbol{h}_s)^{-1}C_s^\top$. Hence, we have a closed form expression for the inverse covariance matrix yielding simple expressions for the estimators of α and τ^2,

$$\begin{aligned} \tilde{\alpha} &= \frac{\sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} x_{0,i}^s (A_{s,i} - x_{1,i}^s) h_{s,i}^{-1}}{\sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} {x_{0,i}^s}^2 h_{s,i}^{-1}} \qquad \text{and} \\ \tilde{\tau}^2 &= N^{-1} \sum_{s\in\mathcal{S}} \sum_{i=1}^{n_s} (A_{s,i} - \tilde{\alpha}x_{0,i}^s - x_{1,i}^s)^2 h_{s,i}^{-1}, \end{aligned}$$

where $A_{s,i}$, $h_{s,i}$, $x_{0,i}^s$ and $x_{1,i}^s$ are the ith components of the respective bold faced vectors. We denote

the unbiased maximum likelihood estimates for the two models as $(\hat{\alpha}, \hat{\tau})$ and $(\tilde{\alpha}, \tilde{\tau})$, respectively. The latter version is what is implemented in an on-line tool as discussed in Section 5.4.2.

In addition to the estimate of α, we are also interested in determining a confidence interval for α. The conditional variance of $\hat{\alpha}$ given $\boldsymbol{A}_+$ is found using the covariance operator on both sides of (5.2),

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\hat{\alpha}|\boldsymbol{A}_+) &= \frac{\mathbb{C}\mathrm{ov}\left(\sum_{s\in\mathcal{S}} \boldsymbol{x}_0^{s\top} W_s^- (\boldsymbol{A}_s - \boldsymbol{x}_1^s)\middle|\boldsymbol{A}_+\right)}{\left(\sum_{s\in\mathcal{S}} \boldsymbol{x}_0^{s\top} W_s^- \boldsymbol{x}_0^s\right)^2} \\
&= \frac{\sum_{s\in\mathcal{S}} \boldsymbol{x}_0^{s\top} W_s^- \mathbb{C}\mathrm{ov}(\boldsymbol{A}_s|\boldsymbol{A}_+) W_s^- \boldsymbol{x}_0^s}{\left(\sum_{s\in\mathcal{S}} \boldsymbol{x}_0^{s\top} W_s^- \boldsymbol{x}_0^s\right)^2} \\
&= \tau^2 \left(\sum_{s\in\mathcal{S}} \boldsymbol{x}_0^{s\top} W_s^- \boldsymbol{x}_0^s\right)^{-1}, \qquad (5.3)
\end{aligned}$$

where we from the first to second equality used the conditional independence of $\boldsymbol{A}_s$ and $\boldsymbol{A}_t$ given $\boldsymbol{A}_+$, and second to third properties of the covariance together with the expression of $\mathbb{C}\mathrm{ov}(\boldsymbol{A}_s|\boldsymbol{A}_+)$ in (5.1). The confidence interval of α given $\boldsymbol{A}_+$ is then given by

$$\mathrm{CI}_\beta(\alpha) = \hat{\alpha} \pm t_{1-\beta/2,N} \frac{\hat{\tau}}{\sqrt{\sum_{s\in\mathcal{S}} \boldsymbol{x}_0^{s\top} W_s^- \boldsymbol{x}_0^s}},$$

where $t_{1-\beta/2,N}$ is the critical value on significance level β for a t-distribution with $N = n_+ - S - 1$ degrees of freedom. A similar confidence interval using the $(\tilde{\alpha}, \tilde{\tau})$-estimates is obtained by inserting the $(\tilde{\alpha}, \tilde{\tau})$-estimates instead of $(\hat{\alpha}, \hat{\tau})$ and replacing W^- with W^{-1}. From the expression of $\mathrm{CI}_\beta(\alpha)$, it is obvious that a small τ-estimate decreases the width of the confidence interval and thus increases the trust in the estimated mixture proportion.

5.4.1 Greedy algorithm

This model was used in an algorithm for finding the most likely pair of profiles contributing to an observed mixture where the STR profiles of both individuals were assumed unknown. First, define the set $\mathcal{J} = \{\mathcal{J}_1, \ldots, \mathcal{J}_4\}$, where $\mathcal{J}_i$ is the set of plausible profiles for loci with $n_s = i$. These sets were defined in Section 5.4 (Table 5.3). The pseudo code for a greedy algorithm finding a pair of profiles (locally) maximising the likelihood of the model specified by (5.1) is given in Figure 5.1. A greedy algorithm is any algorithm that solves a problem by making the locally optimum choice at each stage with the hope of finding the global optimum. A graphical representation of the algorithm is given in Figure 5.7 for a general number of contributors, m. The algorithm works with both $(\hat{\alpha}, \hat{\tau})$ or $(\tilde{\alpha}, \tilde{\tau})$ as estimates of (α, τ).

The greedy algorithm initiates by estimating α based on a locus s with four present alleles. The loci of $\mathcal{S}_4$ contain full information on the mixture ratio, α, and are thus used for assessing this quantity. In succession, the loci with three and two ($\mathcal{S}_3$ and $\mathcal{S}_2$, respectively) observations are analysed and the combination with the smallest contribution to τ and best concordance to the

Algorithm: Find best matching pair of STR profiles.

Let $\mathcal{T} = \emptyset$, $\hat{\alpha} = 0$ and $\hat{\tau}^2 = \infty$.
While $\hat{\tau}^2$ decreases or $\mathcal{T} \neq \mathcal{S}$
  For $i \in \{4, 3, 2\}$
    For $s \in \mathcal{S}_i = \{s : s \in \mathcal{S} \text{ and } n_s = i\}$
      Choose combination $j \in \mathcal{J}_i$ minimising $\hat{\tau}^2$
      Set $\mathcal{T} = \{\mathcal{T} \setminus (s, \cdot)\} \cup (s, j)$ and compute $\hat{\alpha}$
Return $\hat{\alpha}$, $\hat{\tau}$ and $\mathcal{T}$.

Figure 5.1: Greedy algorithm for finding a pair of profiles (locally) maximising the likelihood of (5.1).

previously determined mixture proportion is chosen. The set $\mathcal{T}$ contains a list of the optimal combinations on previously visited loci and is updated after each iteration. On termination, the greedy algorithm returns the best matching pair of profiles together with the estimates of α and τ. The algorithm is designed to perform calculations and decisions similar to those of a forensic geneticist when analysing a two-person mixture.

The optimisation problem is complicated since the inputs of the function that we are interested in minimising depend on each other, $f(\alpha,(\boldsymbol{P}_{s,1}, \boldsymbol{P}_{s,2})_{s\in\mathcal{S}}) = \sum_{s\in\mathcal{S}} D_s$, where $D_s = (\boldsymbol{A}_s - \alpha\boldsymbol{x}_0^s - \boldsymbol{x}_1^s)^\top W_s^-(\boldsymbol{A}_s - \alpha\boldsymbol{x}_0^s - \boldsymbol{x}_1^s)$. Here, f denotes the object function and $(\boldsymbol{P}_{s,1}, \boldsymbol{P}_{s,2})_{s\in\mathcal{S}}$ the set of possible combinations for all loci, $s \in \mathcal{S}$. It is easy to see that, for a fixed α, we can minimise D_s for each locus s by choosing the combination yielding the smallest square distance. Similarly, fixing the combinations for all loci, α is estimated using (5.2). However, from the construction of the greedy algorithm, the algorithm chooses the combination that minimises τ^2 for locus s given α and the configurations on loci previously visited loci, $t \in \{\mathcal{T} \setminus s\}$. This ensures locally optimal solutions, and for most practical purposes, the algorithm returns a global maximum. One should note that when the algorithm recovers the best matching pair of profiles, we still need to consider all profiles *close* to these profiles consistent with the evidence for likelihood ratio evaluation (see Section 5.5 for further details).

5.4.2 On-line implementation

The greedy algorithm of Figure 5.1 together with the methods for evaluating the goodness of fit for a given pair of profiles are implemented in an on-line application. The on-line implementation applies the $(\tilde{\alpha}, \tilde{\tau})$-estimates when finding the best matching pair of profiles. The two-person (and three-person) mixture separator is available on-line at the first author's website (http://people.math.aau.dk/~tvede/dna/). The script can plot the expected and observed peak areas for visual inspection of the fit (see Figure 5.2).

The script allows for user uploads of csv-files containing information about loci, alleles, peak heights and peak areas. The loci implemented are those contained in the SGM Plus and Identifiler kits (AB) excluding amelogenin.

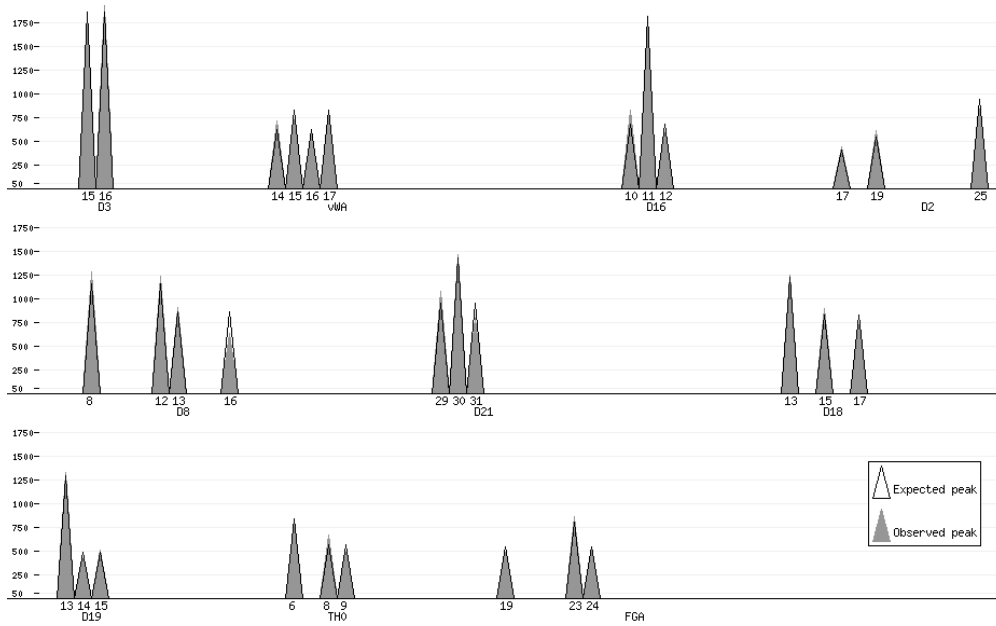

Figure 5.2: Plot produced by the on-line implementation of the algorithm (http://people.math.aau.dk/~tvede/dna/ - sample data file "Paper case"). The observed peaks, ▲, are based on data from Table 5.4, and the expected peaks, △, assuming a mixture of the best matching pair of STR profiles (Table 5.5). The observed and expected peaks coincide for nearly all peaks.

Apart from finding the best matching pair of unknown profiles, the user can specify a suspect profile, and the script finds the best matching unknown profile for two-person mixtures.

Example of a two-person mixture separation in an 1:1 mixture ratio

We demonstrate the algorithm and implementation on data from a controlled experiment conducted at the Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark. The data are presented in Table 5.4 together with information on the true profiles of the mixture (denoted by ∘ and •).

The algorithm found that the two profiles of Table 5.5 are the best matching pair of profiles. The profiles are consistent with the true profiles of the mixture except for loci TH0 and FGA. In Figure 5.2, we have plotted the data from Table 5.4 (solid cones, ▲) together with the best matching pair of profiles as listed in Table 5.5.

In Figure 5.3, the traces of the parameter estimates of α (dashed) and τ^2 (solid) are plotted for each successive iteration with the final parameter estimates being $\tilde{\alpha} = 0.43$ (95%-CI: [0.40 ; 0.45]) and $\tilde{\tau}^2 = 1134.04$. Evaluating the mixture of the true profiles (marked by ∘ and • in

Table 5.4: Data used in demonstrating the algorithm. The ○ and ● represents profile 1 and 2, respectively.

Locus	Allele		Height	Area
D3	15	○ ●	1802	15410
D3	16	○ ●	1939	16282
vWA	14	○	712	6128
vWA	15	●	725	6620
vWA	16	○	626	5637
vWA	17	●	830	7362
D16	10	○	824	7910
D16	11	●	1772	17231
D16	12	○	586	6101
D2	17	○	434	4558
D2	19	●	612	6563
D2	25	○ ●	843	9257
D8	8	●	1284	10782
D8	12	●	1232	10359
D8	13	○	903	7891
D8	16	○	638	5291

Locus	Allele		Height	Area
D21	29	●	1073	9454
D21	30	○	1469	12828
D21	31	●	798	6992
D18	13	○	1247	12302
D18	15	●	899	9104
D18	17	●	726	7549
D19	13	●	1332	10534
D19	14	○	416	3478
D19	15	○	504	3968
TH0	6	○ ●	820	6739
TH0	8	●	668	5573
TH0	9	○	486	4004
FGA	19	○	490	4415
FGA	23	○ ●	865	7968
FGA	24	●	527	5036

Table 5.5: Best matching pair of profiles for the data in Table 5.4. This pair of profiles is pictured in Figure 5.2 as the expected peaks.

Locus	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
Minor	15,16	14,16	10,12	17,25	13,16	30,30	13,13	14,15	6,6	23,23
Major	15,16	15,17	11,11	19,25	8,12	29,31	15,17	13,13	8,9	19,24

Table 5.4), the α estimate is almost unchanged ($\tilde{\alpha} = 0.42$), but with an increase in $\tilde{\tau}^2$ to 1266.34 indicating a slightly worse fit.

The fact that a combination different from the true one has a better fit, indicates that there are multiple explanations of the trace since it is a 1:1-mixture (α close to 0.5). However, the difference in τ^2-estimates for the two combinations will only have a minor influence in the evaluation of the evidence.

Example of a two-person mixture separation in an 1:2 mixture ratio

Wang et al. (2006, Table 10) presented data from a two-person DNA mixture with known minor (victim) and major (suspect) profiles. Curran (2008) and others have analysed these data in order to demonstrate their models for separating two-person DNA mixtures. Using the on-line implementation we obtained the true profiles with $\alpha = 0.30$ (95%-CI: [0.28 ; 0.32]) and $\tau^2 = 124.87$.

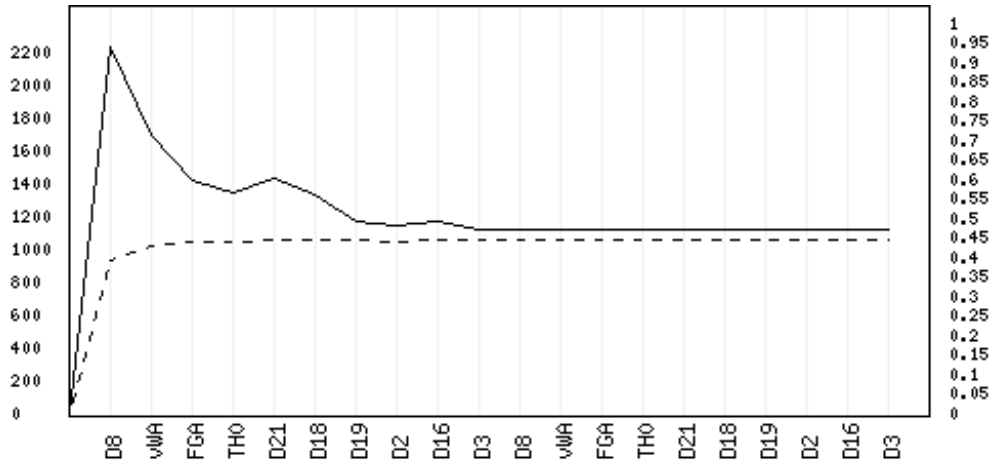

Figure 5.3: Trace of the parameter estimates of α (dashed/right ordinate labels) and τ^2 (solid/left ordinate labels). The plot is produced by the on-line tool available at http://people.math.aau.dk/~tvede/dna/.

5.4.3 Dropping non-fitting loci

In some cases, the stain may be contaminated, and it may be subject to drop-in or drop-out. Drop-ins are allelic peaks present in the DNA profile not belonging to the true profiles. Drop-ins may occur at random (contamination) or by more systematic mechanisms such as stuttering or pull-up effects. Stutters are caused by artefacts in the polymerase chain reaction resulting in an increase of peak intensities typically in the allelic position before the true peaks. Pull-up effects are manifested as an increase of true peaks caused by overlap of the spectra of the light emitted from the various fluorochromes, which are detected by a CCD camera in the data generating process (Butler, 2005). Drop-outs are allelic peaks of the true profiles that are absent in the DNA profile due to, e.g. low amount of DNA or degradation of the DNA. In such cases, the observed peak heights and peak areas no longer originates solely from a two-person mixture. Hence, the proportionalities of Section 5.1 need no longer to be satisfied and the mean structure of (5.1) may not explain the observed peak heights and peak areas in all loci.

We use an F-test approach to evaluate whether any of the included loci $s \in \mathcal{S}$ has significant unexpected balances due to e.g. stutters, degradation or contamination. The purpose is to return a list of loci in which the hypothesis of a two-person mixture can be supported.

For each locus, the contribution to τ^2 is computed by D_s, which we assume to follow a $\chi^2_{n_s-1}$-distribution. Hence, to test whether any locus contributes significantly to the overall variance, τ^2, we evaluate for each locus $s \in \mathcal{S}$ the ratio

$$\frac{(n_s-1)^{-1}D_s}{(n_+-S-n_s-1)^{-1}\sum_{t\in\{\mathcal{S}\setminus s\}} D_t} \sim F_{(n_s-1),(n_+-S-n_s-1)},$$

where $F_{(\nu_1),(\nu_2)}$ is an F-distribution with ν_1 numerator and ν_2 denominator degrees of freedom. Since we perform this test for all loci, we make a Bonferroni-correction to compensate for multiple testing. We apply this procedure successively and drop the most significant locus (if any) until no locus has a significant test-value. This facility is also available in the on-line implementation.

If the variance contribution from multiple loci is large, the test-value will not indicate any significant locus as the overall noise of the sample is large or may be a mixture of more than two individuals. This will result in large values for the overall τ^2.

5.5 Likelihood ratio

Let $\mathcal{G}$ be the DNA profile of the crime stain, and G_S and G_{U_i} the profiles of the suspect and unknown contributor i, respectively. Furthermore, the evidence, $\mathcal{E}$, consists of both quantitative information (peak heights and areas), $\mathcal{Q}$, and the genetic crime stain (allelic information), $\mathcal{G}$. The probability $P(\mathcal{E}|H)$ factories as $P(\mathcal{Q},\mathcal{G}|H) = P(\mathcal{Q}|\mathcal{G},H)P(\mathcal{G}|H)$ using the definition of conditional probabilities. Since $\mathcal{Q}$ is a continuous stochastic variable, we use the likelihood of our model, $L(\boldsymbol{A}|G',G'') = \prod_{s\in\mathbb{S}}\{|W_s|^{-1/2}\exp(-\frac{1}{2}D_s)\}$, to evaluate $P(\mathcal{Q}|\mathcal{G},H)$, where the hypothesis H involves profiles G' and G''.

Let $\mathcal{C}_p = \{G_U : (G_S,G_U) \equiv \mathcal{G}\}$ be the set of unknown profiles that together with G_S are consistent with $\mathcal{G}$, then $P(\mathcal{G}|G_S,G_U) = 1$ for $G_U \in \mathcal{C}_p$ and 0 otherwise, i.e. $\mathcal{C}_p$ is the set of possible unknowns under H_p. Similarly, let $\mathcal{C}_d = \{(G_{U_1},G_{U_2}) : (G_{U_1},G_{U_2}) \equiv \mathcal{G}\}$ be the set of two unknown profiles consistent with $\mathcal{G}$, i.e. possible pairs of profiles under H_d. This partitioning of the set of profiles is equivalent to Assumption 2 in Evett et al. (1998), where the authors argue that the only genotype configurations of interest are those profiles (G',G'') inducing the observation of allelic peaks in $\mathcal{G}$, i.e. $P(\mathcal{G}|G',G'') = 1$ and 0 otherwise. The $LR = P(\mathcal{E}|H_p)/P(\mathcal{E}|H_d)$ can be formed as:

$$LR = \frac{\sum\limits_{G_U\in\mathcal{C}_p} L(\boldsymbol{A}|G_S,G_U)P(G_U)}{\sum\limits_{(G_{U_1},G_{U_2})\in\mathcal{C}_d} L(\boldsymbol{A}|G_{U_1},G_{U_2})P(G_{U_1},G_{U_2})}. \tag{5.4}$$

The $P(G)$ is the profile probability as applied in the regular likelihood ratio (Evett and Weir, 1998), where $P(G)$ may be computed using the θ-correction (Nichols and Balding, 1991; Buckleton et al., 2005). The expression in (5.4) is similar to equations (5) and (6) of Evett et al. (1998) who made a Bayesian formulation of the LR for DNA mixtures.

If a case includes a victim with profile G_V, the set $\mathcal{C}_p = \{(G_S,G_V) \equiv \mathcal{G}\}$ only contain one element, (G_S,G_V). Hence, the likelihood ratio simplifies further

$$LR = \frac{L(\boldsymbol{A}|G_S,G_V)}{\sum\limits_{G_U\in\mathcal{C}_d} L(\boldsymbol{A}|G_V,G_U)P(G_U)},$$

where for this simpler case $\mathcal{C}_d = \{G_U : (G_V,G_U) \equiv \mathcal{G}\}$.

Table 5.6: Expected peak areas for a two-person mixture (expressed in term of α). The list is minimal such that equivalent combinations up to numeration of alleles are avoided. The expected peak areas are ordered by lexicographic order of the allele designation.

n_s	Observed alleles	Combinations	Expected peak areas in terms of α
1	a^4	(aa, aa)	$(2) \times A_{s,+}/2$
2	a^3b	(aa, ab)	$(1+\alpha, 1-\alpha) \times A_{s,+}/2$
		(ab, aa)	$(2-\alpha, \alpha) \times A_{s,+}/2$
	a^2b^2	(aa, bb)	$(2\alpha, 2(1-\alpha)) \times A_{s,+}/2$
		(ab, ab)	$(1, 1) \times A_{s,+}/2$
3	a^2bc	(aa, bc)	$(2\alpha, 1-\alpha, 1-\alpha) \times A_{s,+}/2$
		(ab, ac)	$(1, \alpha, 1-\alpha) \times A_{s,+}/2$
		(bc, aa)	$(2(1-\alpha), \alpha, \alpha) \times A_{s,+}/2$
4	$abcd$	(ab, cd)	$(\alpha, \alpha, 1-\alpha, 1-\alpha) \times A_{s,+}/2$
		(ac, bd)	$(\alpha, 1-\alpha, \alpha, 1-\alpha) \times A_{s,+}/2$

In some cases, the value of $L(\boldsymbol{A}|G_S, G_V)$ may be very much lower than the likelihood value for the pair of best matching profiles. This indicates that it is inappropriate to assume that the evidence is a mixture of G_S and G_V - even though the profiles (G_S, G_V) are consistent with $\mathcal{G}$.

The sums involved in the evaluation of the likelihood ratio will often involve an intractable number of terms depending on the number of loci and number of observed peaks in each locus. As the inclusion of all possible combinations is infeasible, we need at least to include combinations with a numerical impact on the likelihood ratio for the approximation of the true likelihood ratio to be satisfactory for forensic use.

The best matching pair of profiles will provide an estimate, of the mixture proportion α. The expected peak areas in Table 5.6 (expressed in terms of α) indicate that alternative combinations need to have an α-estimate close to the estimate of the best matching pair in order to have a reasonable fit. We exploit this result when defining our proposal distribution in the section on importance sampling.

5.6 Importance sampling of the likelihood ratio

An exact assessment of the weight of evidence comprises evaluation of every term of the numerator and denominator of (5.4). However, this is infeasible and other methods of evaluating the evidence need to be considered. In this section, we show how importance sampling can be used, for estimation of the weight of evidence by assigning weights to the individual combinations. Maimon (2010) also considered importance sampling in a Bayesian context for modelling DNA mixtures.

Let $\mathcal{C}_d = \{(G_{U_1}, G_{U_2}) \equiv \mathcal{G}\}$, and $\boldsymbol{G} = (G', G'')$ refer to a pair of profiles (G', G''). The expression of $P(\mathcal{E}|H_d)$ can be interpreted as a expectation of $\mathcal{Q}$ with respect to the probability measure P on $\mathcal{G}$:

$$P(\mathcal{E}|H_d) = \sum_{\boldsymbol{G}\in\mathcal{C}_d} L(\boldsymbol{A}|\boldsymbol{G})P(\boldsymbol{G}) = \mathbb{E}(h(\mathcal{E}); P). \tag{5.5}$$

Hence, simulating combinations $\boldsymbol{G}$ from $\mathcal{G}$ with respect to P may be used to estimate $P(\mathcal{E}|H_d)$. However, simulation with respect to P does not take the quantitative evidence, $\mathcal{Q}$, into account and will thus yield a poor estimate of $P(\mathcal{E}|H_d)$ due to the possible larger numerical impact from $L(\boldsymbol{A}|\boldsymbol{G})$ compared to $P(\boldsymbol{G})$ in (5.4). To handle this, we use importance sampling based on the "marginal" likelihood values of each combination.

Let $q(\boldsymbol{G}) = \prod_{s\in\mathcal{S}} q_s(\boldsymbol{G}_s)$, where $\boldsymbol{G}_s = (G'_s, G''_s)$ is the profiles on locus s and

$$q_s(\boldsymbol{G}_s) = \frac{L(\boldsymbol{A}|\boldsymbol{G}_s, \hat{\boldsymbol{G}}_{-s})P(\boldsymbol{G}_s)}{\sum_{i=1}^{N_s} L(\boldsymbol{A}|\boldsymbol{G}_{s,i}, \hat{\boldsymbol{G}}_{-s})P(\boldsymbol{G}_{s,i})}, \tag{5.6}$$

where N_s is the number of combinations for the observed number of alleles, $(\boldsymbol{G}_s, \hat{\boldsymbol{G}}_{-s})$ is the particular combination on locus s merged with the best matching combination, $\hat{\boldsymbol{G}}$, in the remaining loci, $t \in \{\mathcal{S} \setminus s\}$, and the sum in the denominator is over all possible combinations, N_s, in locus s merge with the best matching combination in the remaining loci. Hence, $L(\boldsymbol{A}|\boldsymbol{G}_s, \hat{\boldsymbol{G}}_{-s})$ is called the "marginal" likelihood as it gives the likelihood for the particular combination on locus s with the combinations on the remaining loci identical to the best matching pair of profiles. Furthermore, the denominator of (5.6) is a constant, B_s, for each locus. Using this proposal distribution, $P(\mathcal{E}|H_d)$ may be expressed as an expectation with respect to q,

$$P(\mathcal{E}|H_d) = \sum_{\boldsymbol{G}\in\mathcal{C}_d} L(\boldsymbol{A}|\boldsymbol{G})\frac{P(\boldsymbol{G})}{q(\boldsymbol{G})}q(\boldsymbol{G}) = \mathbb{E}(h(\mathcal{E})W(\mathcal{E}); q),$$

where $W(\mathcal{E}) = P(\boldsymbol{G})/q(\boldsymbol{G})$ is the importance weight. Since $P(\boldsymbol{G}) = \prod_{s\in\mathcal{S}} P(\boldsymbol{G}_s)$ and $B = \prod_{s\in\mathcal{S}} B_s$, the ratio of $L(\boldsymbol{A}|\boldsymbol{G})P(\boldsymbol{G})/q(\boldsymbol{G})$ is nearly constant:

$$\frac{L(\boldsymbol{A}|\boldsymbol{G})P(\boldsymbol{G})}{\frac{\prod_{s\in\mathcal{S}}\{L(\boldsymbol{A}|\boldsymbol{G}_s,\hat{\boldsymbol{G}}_{-s})P(\boldsymbol{G}_s)\}}{\prod_{s\in\mathcal{S}} B_s}} = \frac{L(\boldsymbol{A}|\boldsymbol{G})B}{\prod\limits_{s\in\mathcal{S}} L(\boldsymbol{A}|\boldsymbol{G}_s, \hat{\boldsymbol{G}}_{-s})},$$

where the product in the denominator in many cases is a good approximation to $L(\boldsymbol{A}|\boldsymbol{G})$. This constantness of $h(\mathcal{E})W(\mathcal{E})$ improves the performance of importance sampling and reduces the number of samples needed for results with low variance (Robert and Casella, 2004).

In order to estimate $P(\mathcal{E}|H_d)$, we draw combinations $\boldsymbol{G}_i, i = 1, \ldots, M$, from $q(\boldsymbol{G})$ and compute the Monte Carlo estimate,

$$\hat{P}(\mathcal{E}|H_d) = \frac{1}{M}\sum_{i=1}^{M} L(\boldsymbol{A}|\boldsymbol{G}_i)W(\boldsymbol{G}_i), \qquad \boldsymbol{G}_i \sim q(\boldsymbol{G}),$$

where $W(\boldsymbol{G}_i) = P(\boldsymbol{G}_i)/q(\boldsymbol{G}_i)$ are the importance weights.

The estimate is unbiased as the terms are independently simulated from $q(\boldsymbol{G})$ and all have expectation $\mathbb{E}(h(\mathcal{E})W(\mathcal{E});q) = P(\mathcal{E}|H_d)$. For the variance of $\hat{P}(\mathcal{E}|H_d)$, we compute

$$\mathbb{V}\text{ar}(\hat{P}(\mathcal{E}|H_d))=\frac{1}{M-1}\sum_{i=1}^{M}\left\{[L(\boldsymbol{A}|\boldsymbol{G}_i)w(\boldsymbol{G}_i)]^2-\hat{P}(\mathcal{E}|H_d)^2\right\}.$$

The numerator of *LR*, $P(\mathcal{E}|H_d)$, can be handled similarly taking into consideration that we are summing over a restricted set of combinations, $\mathcal{C}_p$, all including the suspect's profile, $\mathcal{C}_p = \{G_U : (G_S, G_U) \equiv \mathcal{G}\}$. The greedy algorithm of Figure 5.1 is also applicable when specifying a suspect. We only need another ordering of the observations and a different set of $\mathcal{J}$-matrices using the extra information of the suspect's profile. This implies that there exists a best matching combination, $\hat{\boldsymbol{G}}^{(S)}$, in $\mathcal{C}_p$ having the same properties as $\hat{\boldsymbol{G}}$ for the unrestricted set, $\mathcal{C}_d$. Hence, importance sampling may also be used in estimating $P(\mathcal{E}|H_p)$ with similar formulae as those for estimating $P(\mathcal{E}|H_d)$.

5.6.1 Example of estimating *LR* using importance sampling

The best matching pair of profiles for the data in Table 5.4 was found in Table 5.5 and were used for estimating $q(\boldsymbol{G})$ and the constant B. In the computations, we assumed uniform distributions of the allele probabilities. Table 5.7, lists the profile of a fictive suspect, G_S, together with the unknown profile maximising the likelihood with G_S fixed. This pair plays the role of $\hat{\boldsymbol{G}}^{(S)}$ in this example. In Figure 5.4 the observed, ▲, and expected peak heights, △, assuming a mixture of these profiles are plotted.

Table 5.7: Suspect's STR profile together with best matching STR profile of an unknown person.

Locus	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
Suspect	16,16	15,17	11,11	25,25	13,16	30,30	15,17	15,15	8,9	24,24
Unknown	15,15	14,16	10,12	17,19	8,12	29,31	13,13	13,14	6,6	19,23

In order to verify the validity of our methodology and implementation of the importance sampler, we limited our data to include only loci on the blue fluorescent dye band (D3, vWA, D16 and D2). The total number of possible combinations for the blue loci is $7^1 12^2 6^1 = 6{,}048$ and it is therefore possible to compute the correct value of $P(\mathcal{E}|H_d) = 0.481335 \times 10^{-10}$. For the suspect's profile specified in Table 5.7, locus D3 is the only blue locus for which it is possible to alter the unknown profile and still have consistency with $\mathcal{G}$. Hence, there are only two terms in the $P(\mathcal{E}|H_p)$ when restricting the analysis to the blue dye band. The value of $P(\mathcal{E}|H_p) = 0.225730 \times 10^{-13}$ indicating that the suspect is not likely to be a true contributor of the DNA mixture since $P(\mathcal{E}|H_p) < P(\mathcal{E}|H_d)$.

In order to evaluate the performance of the importance sampler, we computed 1,000 estimates of $P(\mathcal{E}|H_d)$ each based on 10,000 samples. The estimates are plotted together with the correct value

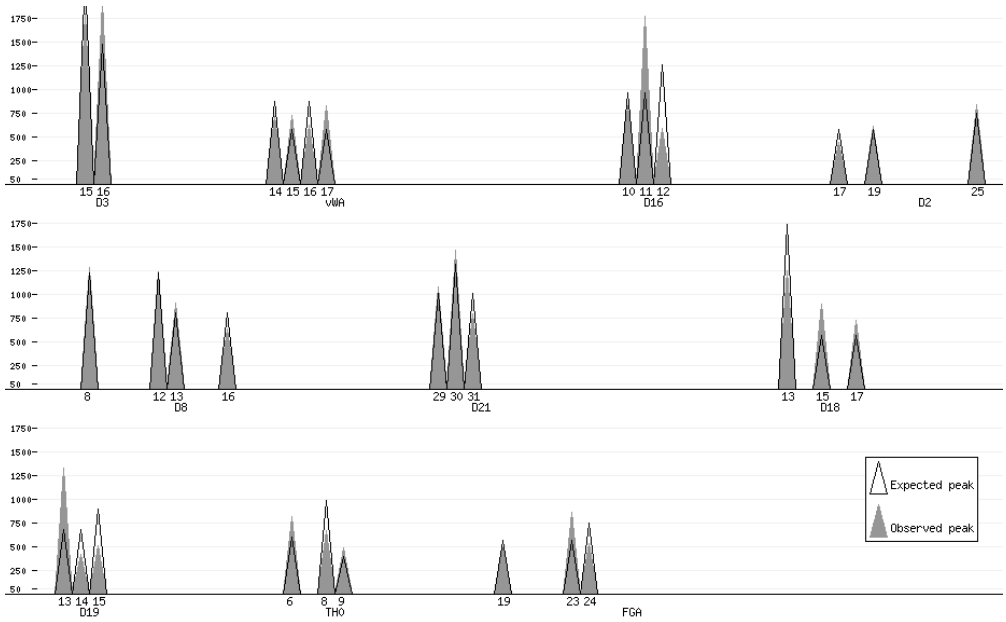

Figure 5.4: Plot of the observed peaks, ▲, and the expected peaks, △, assuming a mixture of the suspect and best matching unknown (STR profiles of Table 5.7).

in the histogram of Figure 5.5. The distribution of the estimates tends to be skew for this particular example, but with most of the estimates close to the true value of $P(\mathcal{E}|H_d)$. The mean of the estimates, $\overline{P}(\mathcal{E}|H_d)$, is 0.483731×10^{-10} with a standard deviation of 0.184432×10^{-11}. From the central limit theorem we may approximate the (positive) distribution of $\overline{P}(\mathcal{E}|H_d)$ with a normal distribution and compute an approximative 95%-confidence interval: $[0.122243, 0.8452178] \times 10^{-10}$. In forensic genetics it is common practice to evaluate the evidence anti-conservative, meaning that the estimates and approximations are favourable to the suspect/defendant (Balding, 2005). For a conservative *LR* the estimate of the numerator should be larger than the true value, $\hat{P}(\mathcal{E}|H_d) > P(\mathcal{E}|H_d)$. However, 66% of the importance sample estimates are smaller than the true value for this particular example. A likely explanation for this is that the sampling scheme places to much of the probability mass close to the best matching pair of profiles. Hence, the (very) large set of less likely combinations are not included in the estimate.

5.7 Results

The algorithm was tested on data from 71 controlled two-person mixtures with known profiles. Hence, it was possible to validate the suggested profiles returned by the separation algorithm. Table 5.8 summarises the comparisons with the best matching pair of profiles and the true mixture profiles.

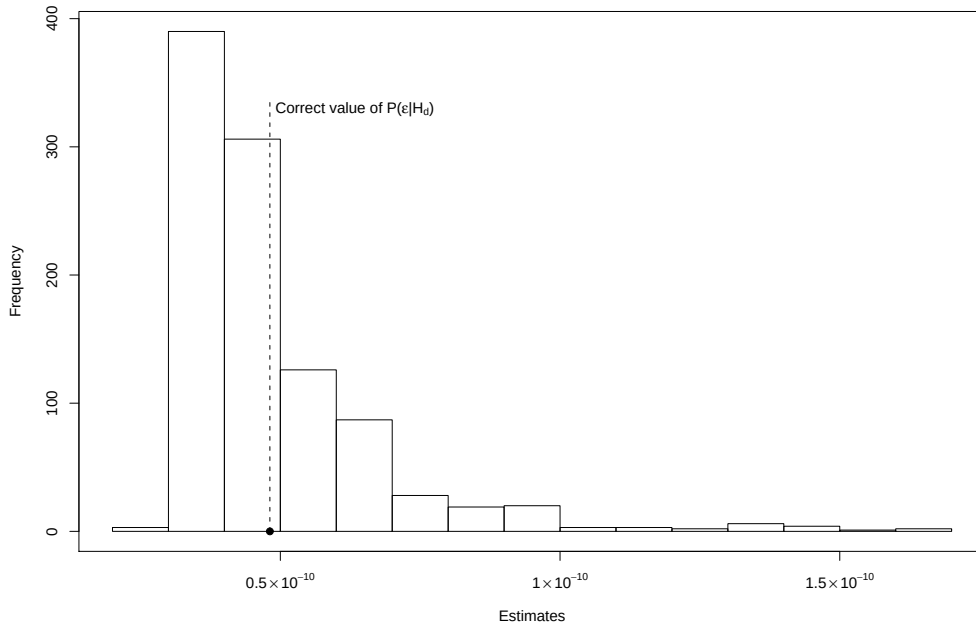

Figure 5.5: Histogram of 1,000 estimates of $P(\mathcal{E}|H_d)$ each based on 10,000 samples.

Table 5.8: Detailed summary table with the number of correctly separated loci, *x*, stratified by mixture ratio.

	Cases with both profiles correct in *x* of 10 loci								Cases with major profile correct in *x* of 10 loci								Cases with minor profile correct in *x* of 10 loci							
Ratio	3	4	5	6	7	8	9	10	3	4	5	6	7	8	9	10	3	4	5	6	7	8	9	10
1:1	1	3	0	2	2	2	1	0	1	2	1	2	2	2	1	0	1	3	0	2	2	2	1	0
1:2	0	0	0	0	2	5	7	8	0	0	0	0	0	2	8	12	0	0	0	0	2	4	8	8
1:4	0	0	0	1	3	2	6	7	0	0	0	0	0	0	0	19	0	0	0	1	3	2	6	7
1:8	0	0	0	2	5	4	0	4	0	0	0	0	0	0	0	15	0	0	0	2	5	4	0	4
1:16	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	4	0	0	1	0	0	0	0	3
Total	1	3	1	5	12	13	14	22	1	2	1	2	2	4	9	50	1	3	1	5	12	12	15	22

From the bottom row of Table 5.8, we see that the separation algorithm returned the true mixture profiles as the best matching combination 22 times. The number of cases where one (14 cases), two (13 cases) or three (12 cases) loci were wrongly separated were almost the same. In five cases, half or less of the loci were correctly separated.

In 50 cases, the true major profile were correctly identified and in another 13 there were inconsistency in at most two loci between the major profile of the best matching pair and the true major profile. Furthermore, Table 5.8 shows that the eight remaining cases with incorrect identification of the major profile had mixture ratio 1:1. Hence, in these cases, there were no obvious major

profiles as the amounts of DNA contributed were (almost) equal. Furthermore, for 1:1-mixtures, there are many pairs of profiles yielding similar goodness of fit to the observed peak intensities, which previously was exemplified in Section 5.4.2. The algorithm is less successful in identification of the minor profile. However, in most cases, the minor profile was separated correctly in seven or more loci.

In addition to the 71 DNA mixtures from controlled experiments, the separation algorithm was used to separate 64 two-person DNA mixtures from real crime cases. For each of the 64 crime cases the laboratory had two reference samples that were consistent with the observed stain. Three experienced forensic geneticists tried to identify both the major and minor profiles of the mixture without knowing the true profiles of the mixture for each mixture (blinded experiment). In Table 5.9, the results from the separation using the separation algorithm is compared to those of the forensic geneticists.

Table 5.9: Comparison of the performance of the separation algorithm and forensic geneticists. The counts show the number of loci with the minor and major profiles correctly identified.

	Geneticists		Algorithm	
Correct loci	Minor	Major	Minor	Major
10	8	31	16	36
9	16	8	16	9
8	13	8	14	5
7	13	4	10	6
6	6	7	1	2
5	3	2	3	2
4	2	2	2	2
3	2	1	2	2
2	0	0	0	0
1	1	1	0	0

The total number of correctly separated mixtures was 16 for the separation algorithm and 8 for the forensic geneticists. The samples where the minor contributor were correctly identified in all loci also had the major component correct (see Table 5.9). As for the controlled experiments, the success rate was dependent by the mixture ratio, with number of correctly separated loci decreasing as α increased towards 0.5.

Furthermore, it should be noted that the forensic geneticists were forced to call some pairs of profiles resulting in some inconclusive statements. That is, the forensic geneticists were forced to deduce major and minor profiles in cases where the regular protocol of the laboratory would not support the separation of profiles.

5.8 Discussion

Using the quantitative information from STR DNA analysis in terms of peak intensities is presently the only way to separate STR mixture results. Based on a statistical model, we developed a simple greedy algorithm for finding the best matching pair of profiles.

Our model is based on few assumptions that are widely accepted among forensic geneticists. The statistical model made it possible to make objective comparisons of various combinations by evaluating the likelihood values. From the normal distribution assumption, this value is computed by τ^{-N}, which implies that the lower τ estimate, the better concordance between observed and expected peaks.

Importance sampling was used in order to estimate the likelihood ratio since this becomes computationally difficult when $7^{S_2}12^{S_3}6^{S_4}$ terms need to be evaluated in the numerator of the *LR* with H_d:(G_{U_1}, G_{U_2}). The method showed to be efficient, and future work will consist of implementation of sampling schemes that explore more of the sample space. This implementation would ideally result in fewer estimates that are less than the true value.

5.9 Conclusion

By using the greedy algorithm of Section 5.4.1, we demonstrated that it is possible to automate the separation of DNA mixtures. However, due to the assumption of no occurrence of drop-out or stutters, the model may be too simple for more complicated cases. Hence, this methodology is applicable to cases where the analysis today is standard but time-consuming. This allows the forensic geneticists to focus on more complex crime cases.

Future work comprises the development of a methodology for handling drop-outs and stutters. Since stutters are profile independent (stutters from parental peaks are constant for all alleged combinations of profiles), it is possible to remove stutters from the data prior to separation and interpretation. Allowing for drop-outs while finding a best matching pair of profiles is also possible. Using the methodology of Tvedebrink et al. (2009), the probability of drop-out, $P(D|\hat{H})$, is assessed conditioned on a given profile.

Acknowledgements

The authors would like to thank Dr. Jakob Larsen and Dr. Frederik Torp Petersen (both from The Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark) for their assistance in manually analysing the DNA mixtures from real crime cases.

Appendix

5.A The general case with m contributors

In the general case with m contributors to the mixed stain, our method can be generalised by assuming the mixture proportions $\alpha_1, \ldots, \alpha_m$ to be strictly increasing,

$$\alpha_1 < \cdots < \alpha_{m-1} < \alpha_m = 1 - \alpha_+, \quad \alpha_+ = \sum_{i=1}^{m-1} \alpha_i. \tag{5.7}$$

The conditional covariance structure is the same as specified in (5.1), where the conditional mean is:

$$\mathbb{E}(\boldsymbol{A}_s | A_{s,+}) = \frac{A_{s,+}}{2}\left[\sum_{i=1}^{m-1} \alpha_i \boldsymbol{P}_{s,i} + \boldsymbol{P}_{s,m}\left(1 - \sum_{i=1}^{m-1} \alpha_i\right)\right] = X_{s,m} + \sum_{i=1}^{m-1} \alpha_i X_{s,i}, \tag{5.8}$$

where $X_{s,i} = (\boldsymbol{P}_{s,i} - \boldsymbol{P}_{s,m})A_{s,+}/2$ for $i = 1, \ldots, m-1$ and $X_{s,m} = \boldsymbol{P}_{s,m}A_{s,+}/2$. In order to find the MLE of $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^{m-1}$, we solve the likelihood equations for $\ell(\boldsymbol{\alpha}, \tau^2; (\boldsymbol{A}_s)_{s\in\mathcal{S}})$ with respect to $\boldsymbol{\alpha}$. This implies that the MLE of $\boldsymbol{\alpha}$ is:

$$\hat{\boldsymbol{\alpha}} = \left(\sum_{s\in\mathcal{S}} X_s^\top W_s^- X_s\right)^{-1}\left(\sum_{s\in\mathcal{S}} X_s^\top W_s^- (\boldsymbol{A}_s - X_{s,m})\right).$$

Furthermore, the estimate of τ^2 in the general setting is

$$\hat{\tau}^2 = N^{-1}\sum_{s\in\mathcal{S}} (\boldsymbol{A}_s - X_s\hat{\boldsymbol{\alpha}} - X_{s,m})^\top W_s^- (\boldsymbol{A}_s - X_s\hat{\boldsymbol{\alpha}} - X_{s,m}),$$

where $N = n_+ - S - m + 1 = \sum_{s\in\mathcal{S}}(n_s - 1) - (m-1)$.

5.A.1 Greedy algorithm

The greedy algorithm of Figure 5.1 needs only a few modifications to be applicable to the general case. Most important is the specification of the number of contributors, m. This needs to be decided before running the algorithm. For the algorithm to be successful, there should preferably be at least one locus with $2m$ peaks as this increases the confidence in the estimate $\hat{\boldsymbol{\alpha}}$. The modified greedy algorithm for m contributors to a DNA mixture is given in Figure 5.6.

Furthermore, it is necessary to check if the $\hat{\boldsymbol{\alpha}}$-estimate satisfies the inequalities of (5.7) for each combination. In Table 5.10, we list fictive data together with two combinations both implying a perfect fit. Both matrices are valid as the orders in the α-sum columns satisfy the condition (5.7). However, for Combination 1, the estimate $\hat{\boldsymbol{\alpha}}_1 = (0.2, 0.45)$ does not satisfies (5.7) while the estimate for Combination 2 $\hat{\boldsymbol{\alpha}}_2 = (0.2, 0.35)$ does. Hence, Combination 2 is chosen over Combination 1.

Algorithm: Find best matching set of m profiles

Specify the number m of contributors.
Let $\mathcal{T} = \emptyset$, $\hat{\boldsymbol{\alpha}} = \mathbf{0}$ and $\hat{\tau}^2 = \infty$.
While $\hat{\tau}^2$ decreases or $\mathcal{T} \neq \mathcal{S}$
  For $i \in \{2m, \ldots, 2\}$
    For $s \in \mathcal{S}_i = \{s : s \in \mathcal{S} \text{ and } n_s = i\}$
      Choose combination $j \in \mathcal{J}_i$ minimising $\hat{\tau}^2$ and satisfying restrictions of (5.7)
      Set $\mathcal{T} = \{\mathcal{T} \setminus (s, \cdot)\} \cup (s, j)$ and compute $\hat{\boldsymbol{\alpha}}$
Return $\hat{\boldsymbol{\alpha}}$, $\hat{\tau}$ and $\mathcal{T}$.

Figure 5.6: Greedy algorithm for finding a set of profiles (locally) maximising the likelihood of (5.8).

Table 5.10: Fictive data showing the importance of ensuring (5.7) is satisfied.

	Combination 1				Combination 2			
Area	$\boldsymbol{P}_1$	$\boldsymbol{P}_2$	$\boldsymbol{P}_3$	α-sum	$\boldsymbol{P}_1$	$\boldsymbol{P}_2$	$\boldsymbol{P}_3$	α-sum
200	1	0	0	α_1	1	0	0	α_1
450	0	1	0	α_2	0	0	1	$1-\alpha_+$
550	1	0	1	$1-\alpha_2$	1	1	0	$\alpha_1+\alpha_2$
800	0	1	1	$1-\alpha_1$	0	1	1	$1-\alpha_1$

In Figure 5.7, the greedy algorithm of Figure 5.6 is described by a diagram emphasising the various steps in the procedure of finding the best matching combination.

In step A, the parameters $\boldsymbol{\alpha}$ and τ are estimated using only the loci with $2m$ observed peaks. Step B determines the profile combination (see step D) on the current locus that minimises τ given the combinations on the already visited loci. The algorithm visits the blocks of loci with equal numbers of observed alleles in decreasing order: $2m-1, \ldots, 2$. If any of the blocks is empty, the algorithms skips forward to the next nonempty block. The order within each block of loci with $2m-i$ observed peaks is arbitrary. When reaching the last locus, the combination and estimates of $\boldsymbol{\alpha}$ and τ are saved.

In step C, the algorithm visits each locus searching for a combination that might decrease τ with all remaining loci combinations fixed. If τ is non-changed the algorithm stops. Otherwise step C is looped until a fixed τ-value is obtained. On termination the algorithm returns the combination and estimates of $\boldsymbol{\alpha}$ and τ.

Step D pictures that, for each locus with less than $2m$ peaks, there are several combinations of profiles that need to be investigated. In the figure, each ∘ depicts a combination and • symbolises the current optimal configuration. The black arrow shows which combination is currently tested. When all the combinations are tested the one with smallest τ is returned.

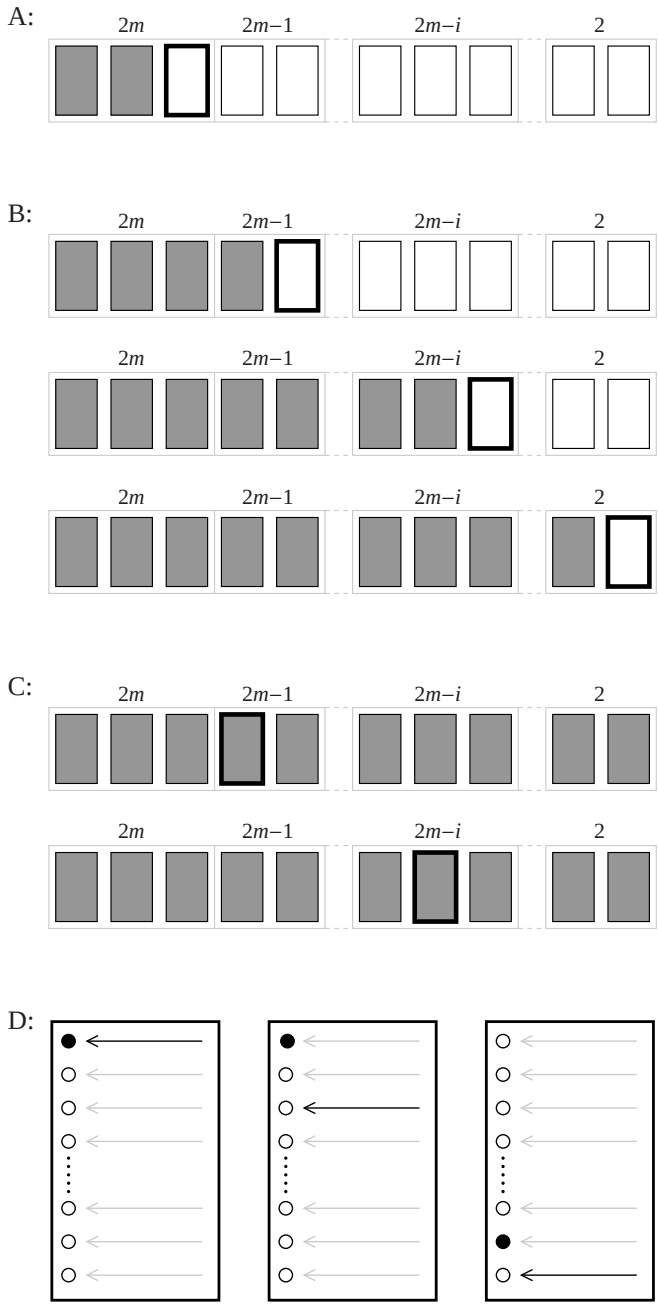

Figure 5.7: Diagram describing the greedy algorithm for resolving DNA mixtures. The shaded boxes show the loci previously visited by the algorithm. The bold lined box shows the current locus under investigation.

Bibliography

Balding, D. J. (2005). *Weight-of-evidence for Forensic DNA Profiles*. Chichester, West Sussex: John Wiley & Sons, Ltd.

Bill, M. et al. (2005). PENDULUM - a guideline-based approach to the interpretation of STR mixtures. *Forensic Science International 148*, 181–189.

Buckleton, J. S., C. M. Triggs, and S. J. Walsh (2005). *Forensic DNA evidence interpretation*, pp. 217–274. Boca Raton, FL: CRC Press.

Butler, J. M. (2005). *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* (2 ed.). Burlington, MA: Elsevier Academic Press Inc., U.S.

Clayton, T. M., J. P. Whitaker, R. Sparkes, and P. D. Gill (1998). Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International 91*, 55–70.

Cowell, R. G., S. L. Lauritzen, and J. Mortera (2007a). A gamma model for DNA mixture analyses. *Bayesian Analysis 2*(2), 333–348.

Cowell, R. G., S. L. Lauritzen, and J. Mortera (2007b). Identification and separation of DNA mixtures using peak area information. *Forensic Science International 166*, 28–34.

Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics 29*(2), 357–372.

Curran, J. M. (2008). A MCMC method for resolving two person mixtures. *Science & Justice 48*, 168–177.

Evett, I. W., P. D. Gill, and J. A. Lambert (1998). Taking account of peak areas when interpreting mixed DNA profiles. *Journal of Forensic Sciences 43*(1), 62–69.

Evett, I. W. and B. S. Weir (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sunderland, MA: Sinauer Associates.

Gill, P. D. et al. (1998). Interpreting simple STR mixtures using allele peak areas. *Forensic Science International 91*(1), 41–53.

Gill, P. D. et al. (2006). DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International 160*(2-3), 90–101.

Maimon, G. (2010). *A Bayesian approach to the statistical interpretation of DNA evidence*. Ph. D. thesis, Department of Mathematics and Statistics, McGill University, Montreal, Canada.

Nichols, R. A. and D. J. Balding (1991). Effects of population structure on DNA fingerprint analysis in forensic science. *Heredity 66*, 297–302.

Perlin, M. W. and B. Szabady (2001). Linear mixture analysis: A mathematical approach to resolving mixed DNA samples. *Journal of Forensic Science 46*(6), 1372–1378.

Robert, C. P. and G. Casella (2004). *Monte Carlo Statistical Methods* (2 ed.). Springer.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2009). Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics 3*(4), 222–226.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2010). Evaluating the weight of evidence using quantitative STR data in DNA mixtures. *Journal of the Royal Statistical Society. Series C, Applied statistics.* In Press.

Wang, T., N. Xue, and J. D. Birdwell (2006). Least-square deconvolution: A framework for interpreting short tandem repeat mixtures. *Journal of Forensic Science 51*(6), 1284–1297.

5.10 Supplementary remarks

In the above manuscript only two-person mixtures were analysed in practice, but the appendix demonstrated how to extend the model and algorithm to handle *m*-person mixtures. The Section of Forensic Genetics, University of Copenhagen, also prepared three-person mixtures. Five different DNA profiles were mixed in trios in the mixture ratios: 1:2:4. The five DNA profiles are listed in Table 5.11. There is $\binom{5}{3} = 10$ different triple-wise combinations and each triple is analysed in six different mixture ratios (permutations of the three profiles). This gives 120 samples since each case is analysed in duplicates. However, 17 samples were discarded due to pipette and amplification errors leaving 103 samples to be analysed.

Table 5.11: The five DNA profiles used in the three-person mixtures.

	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
A	14,18	17,19	12,14	20,24	10,13	30.2,32.2	13,13	12,13	8,9	20,22
B	17,18	14,17	9,9	17,23	13,15	28,28	14,19	14,15.2	8,9.3	20,24
C	16,18	16,19	10,13	16,23	11,14	31,32.2	15,19	12,15	9,9.3	20,24
D	15,18	16,18	9,11	20,21	12,13	29,32.2	13,14	13,14	6,7	22,22
E	15,19	15,17	12,13	16,19	12,13	27,30	13,15	13,14	9,9.3	19,25

The on-line implementation is programmed such that it handles both the analysis of single source stains, two- and three-person mixtures. In Figure 5.8 the peak intensities for a mixture of profiles B, D and C (see Table 5.11) is plotted together with the expected values for the best matching combination.

Since we know the true profiles, we are able to compare the best matching combination with the true profiles as for the two-person mixtures. In Table 5.12 the three inferred profiles are listed. The major profile coincides with profile B while the mid profile differs from profile D by one allele in locus D19. The minor profile has five correct and 4 partially-correct loci compared to profile C.

Table 5.12: The estimated profiles from the separation of the three-person mixture of Figure 5.8. The major profile coincides with profile B in all loci, the mid profile differs by one allele from profile D in locus D19, and the minor is correctly identified in five loci (compared to profile C).

Locus	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
Minor profile	16,17	17,19	10,13	16,17	11,14	28,31	14,14	12,15	9,9.3	20,24
Mid profile	15,18	16,18	9,11	20,21	12,13	29,32.2	13,15	13,14	6,7	22,22
Major profile	17,18	14,17	9,9	17,23	13,15	28,28	14,19	14,15.2	8,9.3	20,24

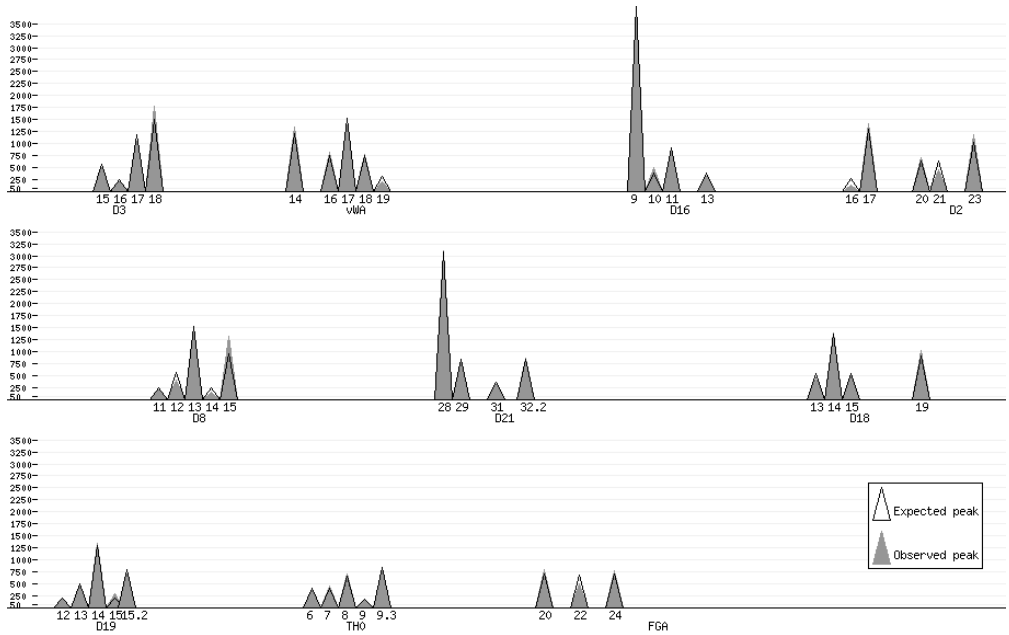

Figure 5.8: Three-person DNA mixture of profiles B, D and C in mixture ratio 4:2:1 (see Table 5.11).

The performance of the mixture separator for the three-person DNA mixtures is summarised in Table 5.13. For each three-person mixture the number of correctly (both alleles correctly identified) and semi-correctly (exactly one alleles correctly identified) loci are computed. This is done separately for the major, mid and minor profile where the median of the corresponding amounts of DNA for these classifications are 335 pg, 168 pg and 84 pg. The first count in each cell refers to the major profile, the second to the mid profile and lastly the minor component.

In 76 cases (73.8%) the major profile was correctly identified on at least eight loci (and partially correct on the remaining ones), while 52 cases (50.5%) had the mid profile correct on at least six loci. The success rate for the minor component was unsatisfactory low. However, the low amounts of DNA compared to the other two components implies that the contributions from the minor profile are within the limits of variation one would expect for the larger peak intensities. That is, the unbalances induced by adding the fraction from the minor component to the peaks of the mid and major profiles is masked by the variability of these peaks.

The authors have in collaboration with Aalborg University and University of Copenhagen applied for a patent for the intellectual rights of the mixture separating algorithm presented above:
Name of invention: A Computer-Assisted Method of Analyzing a DNA Mixture.
Application details: U.S. Provisional Application 61/148221 filed Jan. 29, 2009.

Table 5.13: Summary table of the separation of three-person mixtures. Each cell corresponds to the number of "major/mid/minor" counts for the number of correctly identified loci stratified on matches and partial-matches. Row number is full matches and columns partial matches.

	0	1	2	3	4	5	6	7	8	9	10
0	0/0/0	0/0/0	0/ 0/0	0/0/0	0/0/0	0/0/0	0/0/ 0	0/0/0	0/1/0	0/1/1	0/0/0
1	0/0/0	0/0/0	0/ 0/0	0/0/0	0/0/0	0/1/1	0/0/ 1	0/1/3	0/0/1	0/0/1	
2	0/0/0	0/0/0	0/ 0/0	0/0/0	0/0/0	0/1/2	0/2/ 6	0/5/3	0/2/5		
3	0/0/0	0/0/0	0/ 0/0	0/0/0	0/2/5	0/2/6	0/2/10	0/5/4			
4	0/0/0	0/0/0	0/ 0/1	0/1/1	0/4/5	0/2/7	0/4/ 6				
5	0/0/0	0/0/0	0/ 2/1	1/1/0	0/5/8	3/7/6					
6	0/0/0	0/0/0	0/ 1/0	2/6/8	8/9/4						
7	0/0/0	0/2/0	0/ 6/2	13/8/2							
8	0/0/0	0/2/0	26/11/3								
9	0/0/0	25/6/0									
10	25/1/0										

CHAPTER 6

Estimating the probability of allelic drop-out of STR alleles in forensic genetics

Publication details

Co-authors: Poul Svante Eriksen[*], Helle Smidt Mogensen[†] and Niels Morling[†]

[*] *Department of Mathematical Sciences*
Aalborg University
[†] *Section of Forensic Genetics, Department of Forensic Medicine*
Faculty of Health Science, University of Copenhagen

Journal: Forensic Science International: Genetics 3 (2009) 222-226

DOI: doi:10.1016/j.fsigen.2009.02.002

Abstract:
In crime cases with available DNA evidence, the amount of DNA is often sparse due to the setting of the crime. In such cases, allelic drop-out of one or more true alleles in STR typing is possible. We present a statistical model for estimating the per locus and overall probability of allelic drop-out using the results of all STR loci in the case sample as reference. The methodology of logistic regression is appropriate for this analysis, and we demonstrate how to incorporate this in a forensic genetic framework.

Keywords:
Drop-out probability; forensic genetics; logistic regression; STR.

6.1 Introduction

When assessing the weight of the evidence of STR typing in forensic genetics, the arguments depend on the observable alleles in the crime stain. However, due to technical and biochemical issues, it is possible that a true allele in the sample is not detected by the genetic typing method, i.e. allelic drop-out (Gill et al., 2006). The probability of this event will affect the weight of evidence with a decrease in the power of discrimination as the drop-out probability increases since less individuals can be excluded as possible contributors.

It is well-known that in samples of high quality, i.e. high amount of DNA (for all contributors if it is a mixture) and no contamination or degradation, the probability of observing a drop-out is practically zero. Using logistic regression, we formalised this intuition by using the results of all STR loci in the sample as an indicator of the amount of DNA. The statistical analysis showed that the drop-out probability is locus dependent.

The DNA commission of the ISFG stressed the importance of considering allelic drop-out in the recommendation on mixture interpretation (Gill et al., 2006, recommendation 7). In recommendation 7, the intuition of the logistic model was explained, but how to assess $P(D)$ was not formalised. The estimation of $P(D)$ is important because it influences the estimation of the weight of the evidence in the calculation of the likelihood ratio (LR).

6.2 Material and methods

6.2.1 Data

The analysis was based on 175 controlled experiments conducted at The Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark. The experiments consisted of pairwise mixtures of four profiles and samples with only one contributor diluted in water.

Genomic DNA from blood-samples from two males and two females was extracted by a standard phenol-chloroform extraction method. DNA was quantified in triplicates using the Quantifiler® Human DNA Quantification kit (Applied Biosystems) with Human Genomic DNA Male

(Promega) as the quantification-standard on a ABIPrism® 7000. The median DNA concentrations were used. Each sample was diluted in water to 500 pg DNA/μl. The DNA concentrations in the diluted samples were measured again in triplicates and the median DNA concentration was used.

Six two-person mixtures of DNA (w/v) in proportions 16:1, 8:1, 4:1, 2:1, 1:1, 1:2, 1:4; 1:8 and 1:16 were made of DNA from each of the four persons. The amount of DNA from each person in the mixtures was calculated based on the DNA concentration in each sample. Each of the four samples were serially diluted with water in the proportions 16:1, 8:1, 4:1, 2:1 and 1:1.

The amount of DNA in each mixture ranged from 328 to 528 pg DNA, and from 24.6 to 410 pg DNA in the diluted samples and were amplified twice with the AmpFℓSTR® SGM Plus®-kit (Applied Biosystems) as recommended by the manufacturer in an ABI GeneAmp® 9700 PCR thermocycler.

One ul of the amplificates in 15 ul HiDi® Formamide (Applied Biosystems) was analysed on an ABI Prism® 3100 Genetic Analyzer using POP4 as the polymer and 5 kV injection voltage for 6 seconds. DNA fragments were detected and fragment sizes were estimated with GeneScan 3.7 with a detection threshold of 50 rfu. Genotypes were assigned using GenoTyper 3.7 with the Kazam macro (Applied Biosystems) with no stutter filter applied.

We excluded all alleles in stutter positions of true alleles to avoid complications of masked drop-outs due to stutter effects. Table 6.1 presents the number of observed alleles, dropouts and the proportion of drop-outs for each locus.

Table 6.1: Observed drop-outs in the data set stratified by locus. All drop-outs were single contributor alleles.

	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
Observed	306	356	322	398	362	375	315	220	258	313
Drop-outs	10	11	11	14	11	7	10	10	17	18
Proportion	0.03	0.03	0.03	0.04	0.03	0.02	0.03	0.05	0.07	0.06

There was a tendency for the high molecular loci to have more drop-outs than the remaining ones within each fluorescent dye colour. This indicates a locus dependence of the probability of drop-out.

6.2.2 Logistic regression model

Let D be the event "The contributor's allele has dropped out", and $\bar{D}$ when no drop-out occurs, implying that $P(\bar{D}) = 1-P(D)$. For evidence evaluation, we are interested in quantifying the probability of allelic drop-out $P(D)$. As mentioned in Section 6.1, we wish to model this probability conditioned on the observed stain.

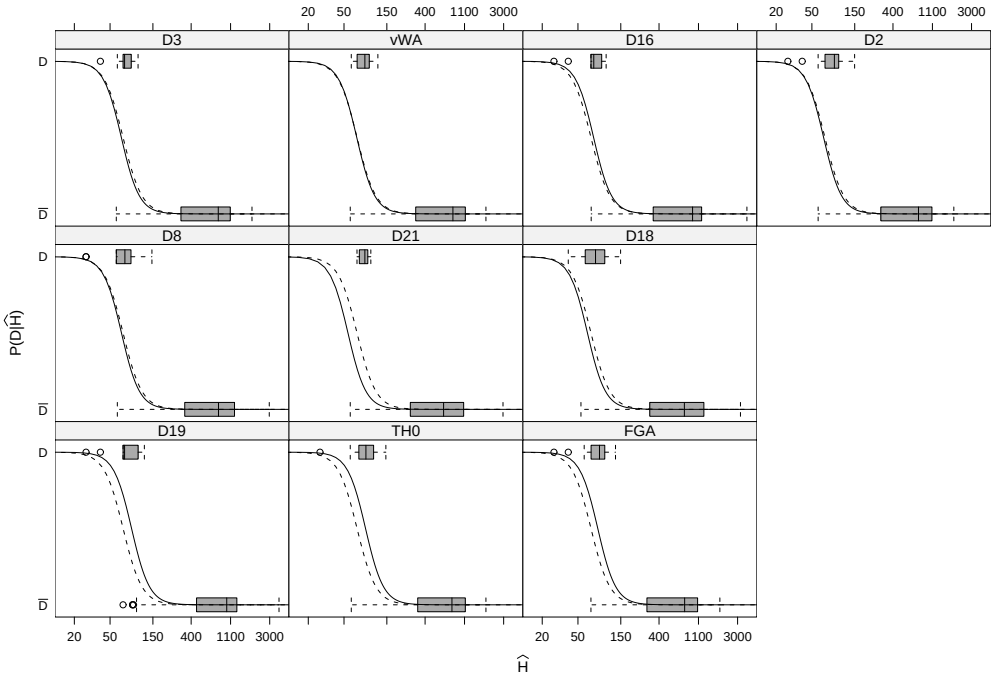

Figure 6.1: Locus specific logistic curves (solid) together with an overall estimate (dashed). The plot is on log-scale ensuring $P(D|\widehat{H}=0)=1$. At each panel box-plots are added, summarising the empirical distribution of $\widehat{H}$ for D and $\overline{D}$.

We define H as the sum of *observed* peak heights divided by a sum of indicators with value two for homozygous alleles and one for heterozygous, i.e. for h_i being the ith height measurement $H=(n_{\text{het}}+2n_{\text{hom}})^{-1}\sum_{i=1}^{n}h_i$, where $n=n_{\text{het}}+n_{\text{hom}}$ is the number of heterozygous and homozygous alleles in the profile. This was previously demonstrated to be a good proxy for the amount of DNA contributed to a stain (Tvedebrink et al., 2010). If the stain is a mixture assumed to have K contributors, we only use the alleles where person $k, k=1,\ldots,K$, is a single contributor for estimating $H^{(k)}$. We use $\widehat{H}$ as a summary statistics for the observed stain in our analysis and use logistic regression to model $P(D|\widehat{H})$, where $\widehat{H}$ is found from H as (for $K=2$),

$$P(D|\widehat{H})=\begin{cases}P(D|H), & \text{Non-shared het allele}\\ P(D|2H), & \text{Non-shared hom allele}\\ P(D|H^{(1)}+H^{(2)}), & \text{Shared het allele,}\end{cases}$$

where $H^{(1)}$ and $H^{(2)}$ may be weighted by 2 if the contributors of the shared alleles are homozygous.

Logistic regression is a standard way to estimate the probabilities for a dichotomous response stochastic variable when explanatory variables are assumed to change the probability of the event (McCullagh and Nelder, 1989). The logistic model is particularly simple in this case since we only have one explanatory variable, $\widehat{H}$,

$$P(D|\widehat{H}) = \frac{\exp(\beta_0 + \beta_1 \log \widehat{H})}{1 + \exp(\beta_0 + \beta_1 \log \widehat{H})},$$

where β_1 showed to be negative such that $P(D|\widehat{H})$ decreases as $\widehat{H}$ increases, and the use of $\log \widehat{H}$ rather than $\widehat{H}$, ensures that with β_1 being negative $P(D|\widehat{H} = 0) = 1$. When we condition on $\widehat{H}$, we assume the event of two allelic drop-outs of the same contributor are independent, which is also an underlying assumption of the logistic regression. That is, $P(D_1, D_2|\widehat{H}) = P(D_1|\widehat{H})P(D_2|\widehat{H})$, where D_i: "Allele i of the contributor with DNA proxy $\widehat{H}$ has dropped out".

6.3 Results and discussion

The analysis showed that the intercept parameter, β_0, varied between loci with a p-value of 0.01 indicating a significant difference between loci (Venables and Ripley, 2002). A similar test for the slope parameter, β_1, indicated that this parameter did not vary significantly across loci (p-value of 0.49). In addition, there was no significant change of the drop-out probability caused by the allelic number indicating that larger alleles within the same locus has the same drop-out probability as smaller alleles. However, in the data set, the largest allelic difference was eight repeat units. This variability may be to small to demonstrate that a possible allelic effect is significant.

The parameters for locus s are thus $\beta_{0,s}$ and β_1 for computing $P(D|\widehat{H})$, where we use the same $\widehat{H}$ for all loci. The parameter estimate of β_1 is -4.35 and the estimates of $\beta_{0,s}$ are given in Table 6.2.

Table 6.2: Estimates of $\beta_{0,s}$ and β_1 based on the experiments of Section 6.2.1.

Locus	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
$\beta_{0,s}$	18.26	18.43	18.75	18.31	18.28	17.45	18.07	19.40	19.40	19.21

Note, that $\beta_{0,s}$ are larger for the loci of the yellow fluorescent dye band indicating their larger drop-out probability as observed in Table 6.1. The corresponding logistic curves for the parameters of Table 6.2 are plotted in Figure 6.1 together with an overall estimate not stratifying on loci. The parameters for the overall curve are $\beta_0 = 17.56$ and $\beta_1 = -4.14$.

In Figure 6.1, the box-plot added to each panel shows the DNA proxy $\widehat{H}$ for the drop-outs (D) and observed alleles ($\overline{D}$). The boxes indicate the inter-quartile range (middle fifty percent of

the data) of the observations and the whiskers extend to the most extreme data points within 1.5 times the lengths of the boxes. Remaining points are marked by dots.

It is clear from Figure 6.1 that there is an overlap of the whiskers in the box-plots. This implies that the classification of drop-outs is associated with uncertainty as one would expect. In particular, it is true for D21 where all drop-outs observed had a mean height, $\widehat{H}$, above 70. This may be due to the specific alleles in our data set (for D21, these were 28, 29, 30, 30.2, 31 and 32.2) and possible individual specific effects from having only four different profiles in the data.

We used the estimated parameters of Table 6.2 in order to create a table of the mean peak heights that correspond to the specific drop-out probabilities. For the ten different loci included in our data set, these mean heights are presented in Table 6.3.

Table 6.3: Mean peak heights (rfu) for various drop-out probabilities for ten STR loci.

$P(D\|\widehat{H})$	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA	Overall
0.0001	556	577	622	562	558	461	531	722	723	692	648
0.0005	384	399	430	388	385	318	367	499	499	478	439
0.0010	327	340	366	331	328	271	313	425	426	407	371
0.0050	226	235	253	228	226	187	216	293	294	281	251
0.0100	192	200	215	194	193	159	184	250	250	239	212
0.0500	132	137	147	133	132	109	126	171	171	164	142
0.1000	111	115	124	112	111	92	106	144	144	138	119
0.2000	92	95	103	93	92	76	88	119	120	114	98
0.3000	81	84	91	82	81	67	78	105	106	101	86
0.4000	73	76	82	74	74	61	70	95	95	91	77
0.5000	67	69	75	68	67	55	64	87	87	83	70
0.6000	61	63	68	62	61	50	58	79	79	76	63
0.7000	55	57	62	56	55	46	53	71	71	68	57
0.8000	49	50	54	49	49	40	46	63	63	60	50
0.9000	40	42	45	41	40	33	39	52	52	50	41
0.9500	34	35	38	34	34	28	32	44	44	42	34
0.9900	23	24	26	23	23	19	22	30	30	29	23

Computing the Brier Score (Brier, 1950) for the estimated locus specific model, we find that the Brier Score $= n^{-1}\sum_{i=1}^{n}(D_i - P(D|\widehat{H}_i))^2 = 0.02$, where D_i is indicator for dropout of the allele of the data and $\widehat{H}_i$ is the associated proxy for the amount of DNA. A Brier Score close to zero

indicates that the model is adequate. A simulated p-value of 0.156 indicates a satisfying fit of the model. Furthermore, we tried to improve the model by using linear splines (Harrell Jr., 2001) with knots at log(75) and log(100), but these model extensions were not supported by the data.

The use of the logit function implies that the interpretation is made in terms of log odds. The log odds of the drop-out probability conditioned on $\widehat{H}$ is linear in $\log \widehat{H}$,

$$\text{logit}P(D|\widehat{H}) = \log \frac{P(D|\widehat{H})}{P(\bar{D}|\widehat{H})} = \beta_{0,s} + \beta_{1,s} \log \widehat{H}.$$

Using H as the explanatory variable implies lower variability on the DNA proxy than if only using a single peak height observation, e.g. the peak height on the same locus of a heterozygous allele that has not dropped-out. Furthermore, in real crime cases such an allele might not be observed, since both alleles of a heterozygous might have dropped-out or the other allele may be shared with an other contributor if the stain is a mixture.

Gill et al. (2000) discussed the importance of addressing the risk of allelic drop-out and how to incorporate this into the likelihood ratio. Combining our approach for estimating $P(D|\widehat{H})$ with the methodology of Gill et al. (2000) may be a feasible approach for better assessment of the weight of evidence when the level of the peak heights indicates the possibility of drop-outs.

6.4 Conclusion

We have demonstrated a simple and applicable way of assessing the drop-out probabilities of STR alleles in forensic genetics. The drop-out probabilities computed using the model concur with the prior knowledge of the drop-out behaviour varying with the observed peak heights.

Future work consists of testing the model on a larger data set including more alleles. With a larger data set, it may also be possible to test whether alleles or fragment length has a significant effect on the drop-out probability as the individual specific effect decreases with the number of different profiles.

It is worth emphasising that the drop-out probabilities may vary between laboratories, machinery within the same laboratory and typing kits used for profiling. This is due to differences in e.g. the ability to amplify the DNA in the PCR and in the potential to measure the light intensities for the electropherogram. Hence, before applying this methodology in the likelihood ratio for evidence calculations, the laboratory needs to perform experiments with known profiles in order to estimate the parameters in the logistic regression model.

Appendix

6.A Examples

In forensic genetics it is common to use the likelihood ratio $LR = P(E|H_p)/P(E|H_d)$ as mean to assess the weight of evidence. Here $P(E|H)$ is the probability of observing the evidence E given the hypothesis H. The prosecutors hypothesis, H_p, often include more profiles from identified individuals than under the defence hypothesis H_d. Having a single contributor stain H_p may state "The suspect is the only contributor to the crime stain", whereas H_d: "An unknown individual unrelated to the suspect is the only contributor to the crime stain".

In the situation where the hypotheses induces that an allelic drop-out has occurred one needs to specify the profiles that constitute the observed stain in order to compute the profile specific drop-out probability for both H_p and H_d.

6.A.1 Example with data from a controlled experiment

We used the data in Table 6.4 to demonstrate the technique of computing the drop-out probability of a given allele. The data originated from a mixture of a controlled experiment with the two profiles A and B denoted in Table 6.4 by ∘ and •, respectively, where A contributed with 31.4 pg/ul and B with 424.6 pg/ul.

Table 6.4: Data used in the example of the Appendix 6.A. The sample was a mixture of the two profiles A and B (denoted by ∘ and •) contributing 31.4 pg/ul and 424.6 pg/ul, respectively.

Locus	Allele			Height	Area
D3	15		•	766	7264
D3	16	∘	•	991	9165
D3	19	∘		–	–
vWA	15	∘	•	788	7631
vWA	17	∘	•	710	6678
D16	10	∘		117	1201
D16	11		•	1765	18858
D16	12	∘		–	–
D2	19		•	746	8816
D2	23	∘		–	–
D2	25	∘	•	696	8432
D8	8		•	967	9145
D8	12		•	895	8350
D8	13	∘		–	–

Locus	Allele			Height	Area
D21	28	∘		70	660
D21	29		•	767	7169
D21	30	∘		102	1024
D21	31		•	889	8283
D18	12	∘		70	736
D18	15		•	766	8501
D18	16	∘		127	1341
D18	17		•	687	7856
D19	13	∘	•	1525	12862
D19	15	∘		–	–
TH0	6	∘	•	836	7333
TH0	7	∘		82	736
TH0	8		•	595	5249
FGA	20	∘		–	–
FGA	23	∘	•	638	6507
FGA	24		•	549	5542

Under the assumption that the data in Table 6.4 originated from a two-person mixture, we need to specify a possible pair of profiles explaining the observed alleles. We compute the individual DNA proxies $H^{(A)}$ and $H^{(B)}$ as defined in Section 6.2.2 for the two profiles A and B of Table 6.4,

$$H^{(A)} = \frac{117 + 70 + 102 + 70 + 127 + 82}{6} = 94.67$$
$$H^{(B)} = \frac{766 + 1765 + 746 + 967 + 895 + 767 + 889 + 766 + 687 + 595 + 549}{10 + (2 \times 1)} = 782.67.$$

Let allele 19 in locus D3 be denoted by $D3_{19}$, then from Table 6.4 we found that the homozygous allele $D8_{13}$ and the following non-shared heterozygous alleles of profile A had dropped out: $D3_{19}$, $D16_{12}$, $D2_{23}$, $D19_{15}$, and FGA_{20}.

The DNA proxy was the same for all the heterozygous drop-outs, $\widehat{H} = H^{(A)}$, and for the homozygous allele $\widehat{H} = 2H^{(A)}$. The parameter estimates of Table 6.2 were then used in order to compute the locus specific drop-out probabilities. Below, we demonstrate how to compute the drop-out probabilities for $D3_{19}$, $D19_{15}$ and $D8_{13}$:

$$\begin{aligned} P(D_{\mathrm{D3}_{19}}|\widehat{H}) &= \frac{\exp(18.26 - 4.35\log(94.67))}{1 + \exp(18.26 - 4.35\log(94.67))} = 0.177, \\ P(D_{\mathrm{D19}_{15}}|\widehat{H}) &= \frac{\exp(19.40 - 4.35\log(94.67))}{1 + \exp(19.40 - 4.35\log(94.67))} = 0.403, \\ P(D_{\mathrm{D8}_{13}}|\widehat{H}) &= \frac{\exp(18.28 - 4.35\log(189.33))}{1 + \exp(18.28 - 4.35\log(189.33))} = 0.011. \end{aligned}$$

Suppose we only had information on profile B, e.g. B being the victim of a crime, and that the profile of the suspect S only gave a partial match. For simplicity, we use the same mean height estimate for the suspect as for A, i.e. $H^{(S)} = H^{(A)}$. In locus D19, only allele 13 was observed and a shared allele may have dropped out. Assuming suspect S is homozygous for allele 11 and profile B is heterozygous with alleles 11 and 13, the DNA proxy is $\widehat{H} = 2H^{(S)} + H^{(B)} = 189.33 + 782.67 = 972$ and the drop-out probability is

$$P(D_{\mathrm{D19}_{11}}|\widehat{H}) = \frac{\exp(19.40 - 4.35\log(972))}{1 + \exp(19.40 - 4.35\log(972))} = 2.69 \times 10^{-5}.$$

6.A.2 Example in the recommendation of the ISFG Commission

Following the idea of Example 1 given in (Gill et al., 2006, Appendix B.2), we compute the likelihood ratio using our model for assessing the drop-out probabilities.

Assume that the genetic stain $\mathcal{G} = (a, c, d)$ and that the prosecutors hypothesis claims that the suspect, $G_S = (a, b)$ is a contributor to the stain. For this hypothesis to be true, the b allele must

have dropped out. In this example, we only consider data from one locus as in Table 3 of the ISFG recommendations. We re-use the data from TH0 in Table 6.4 in order to exemplify how to evaluate the *LR*. For consistency with the example of Gill et al. (2006), denote allele 7 by *a* and let *c* and *d* be allele 6 and 8, respectively.

From Table 6.4, we compute the following estimates of $\widehat{H}$ and the associated $P(D|\widehat{H})$ for every combination of the alleles assuming a two-person mixture. Suppose that a contributor has non shared alleles *mn* and that the DNA proxy for this combination is H_{mn}. Then $P(D_{mn}) = P(D|H_{mn})$ is the drop-out probability of either *m* or *n*. Alternatively in the actual case there may be one shared allele, *m*, and in this case $P(D_{m,m}) = P(D|\widehat{H}_{m,m}) = P(D|H_{mn}+H_{mo})$ is the drop-out probability for allele *m* when shared by two individuals with the combinations *mn* and *mo*. The probability $P(\mathcal{G}|H_p)$ is

$$P(\mathcal{G}|H_p) = 2P(cd)P(\bar{D}_{cd})^2P(D_{ab})P(\bar{D}_{ab}),$$

since allele *b* is assumed to have dropped out.

Assume that an allele, *Q*, has dropped out implying that the two profiles are heterozygous not sharing any allele. That is, *Q* is any allele of $\mathcal{A}_{\text{TH0}} \setminus \{a, c, d\}$ with allele probability $P(Q) = 1-[P(a)+P(c)+P(d)]$, where $\mathcal{A}_{\text{TH0}}$ is the set of alleles for locus TH0. All of the observed alleles must be paired with the missing allele in order to compute the specific drop-out probabilities as these differ due to the different peak heights. From Table 6.5, it is clear that $P(D_{aQ})$ is the largest of the three as expected since the peak height of *a* is only 82 rfu. When paired with any of *c* or *d*, the drop-out probabilities are practically zero as one would require. Hence, the terms $P(D_{cQ})$ and $P(D_{dQ})$ are also indicators of a poor agreement with the heterozygote balance when pairing *ad* and *ac*, respectively.

Table 6.5: DNA proxies and drop-out probabilities for various profiles

Profile(s)	Notation	H	$\hat{H}$	$P(D\|\hat{H})$
aQ	$P(D_{aQ})$	82.0	82.0	5.57×10^{-1}
ac	$P(D_{ac})$	459.0	459.0	7.02×10^{-4}
ad	$P(D_{ad})$	338.5	338.5	2.63×10^{-3}
cQ	$P(D_{cQ})$	836.0	836.0	5.17×10^{-5}
dQ	$P(D_{dQ})$	595.0	595.0	2.27×10^{-4}
cd	$P(D_{cd})$	715.5	715.5	1.02×10^{-4}
aa	$P(D_{aa})$	82.0	164.0	5.82×10^{-2}
cc	$P(D_{cc})$	836.0	1672.0	2.54×10^{-6}
dd	$P(D_{dd})$	595.0	1190.0	1.11×10^{-5}
ac, ad	$P(D_{a,a})$	459.0, 338.5	797.5	6.35×10^{-5}
ac, cd	$P(D_{c,c})$	459.0, 715.5	1174.5	1.18×10^{-5}
ad, cd	$P(D_{d,d})$	338.5, 715.5	1054.0	1.89×10^{-5}

The probability of the evidence given the defence hypothesis and that one allele has dropped out is given as

$$P_1(\mathcal{G}|H_d) = 8P(acdQ)[P(\bar{D}_{ac})^2P(D_{dQ})P(\bar{D}_{dQ}) + P(\bar{D}_{ad})^2P(D_{cQ})P(\bar{D}_{cQ}) + P(\bar{D}_{cd})^2P(D_{aQ})P(\bar{D}_{aQ})],$$

where the multiplication by 8 is due to the number of pairwise combinations of the alleles, e.g. pairing the alleles ac and dQ may be done as $(ac)(dQ)$, $(ac)(Qd)$, $(ca)(dQ)$ and $(ca)(Qd)$; interchanging the profiles yields the eight combinations.

The defence hypothesis, H_d, also comprises the scenario where no alleles has dropped out. This implies that either an allele is shared or one contributor is homozygous. The probability of $P_0(\mathcal{G}|H_d)$ is:

$$\begin{aligned} P_0(\mathcal{G}|H_d) = P(acd)[&P(a)\left\{4P(\bar{D}_{aa})P(\bar{D}_{cd})^2+8P(\bar{D}_{a,a})P(\bar{D}_{ac})P(\bar{D}_{ad})\right\} \\ &+ P(c)\left\{4P(\bar{D}_{cc})P(\bar{D}_{ad})^2+8P(\bar{D}_{c,c})P(\bar{D}_{ac})P(\bar{D}_{cd})\right\} \\ &+ P(d)\left\{4P(\bar{D}_{dd})P(\bar{D}_{ac})^2+8P(\bar{D}_{d,d})P(\bar{D}_{ad})P(\bar{D}_{cd})\right\}]. \end{aligned}$$

It is worth noting that the probabilities $P(D_{ac})$ and $P(D_{ad})$ are misleading as the combination of a together with c or d causes substantially imbalances in the profile's peak heights.

In order to compute the likelihood ratio, LR, we need only to compute the ratio of $P(\mathcal{G}|H_p)$ to $P_1(\mathcal{G}|H_d) + P_0(\mathcal{G}|H_d)$. As in Gill et al. (2006), we assume uniform allele probabilities of 0.1 for the observed alleles implying that $P(Q) = 0.7$, yielding a LR of

$$LR = \frac{P(\mathcal{G}|H_p)}{P_1(\mathcal{G}|H_d)+P_0(\mathcal{G}|H_d)} = \frac{0.0049}{0.0014+0.0035} = 1.0007.$$

For the same scenario, Gill et al. (2006) considered uniform probabilities of 0.02 of the observed alleles. Using our model, this implies a LR of 9.6111.

Bibliography

Balding, D. J. and J. S. Buckleton (2009). Interpreting low template DNA profiles. *Forensic Science International: Genetics 4*(1), 1–10.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*, 1–3.

Gill, P. D. et al. (2006). DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International 160*(2-3), 90–101.

Gill, P. D. and J. S. Buckleton (2010a). A universal strategy to interpret DNA profiles that does not require a definition of low-copy-number. *Forensic Science International: Genetics 4*(4), 221–227.

Gill, P. D. and J. S. Buckleton (2010b). Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. *Journal of Forensic Sciences 55*(1), 265–268.

Gill, P. D., J. M. Curran, and K. Elliot (2005). A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. *Nucleic Acids Research 33*(2), 632–643.

Gill, P. D., J. Whitaker, C. Flaxman, N. Brown, and J. S. Buckleton (2000). An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International 112*(1), 17–40.

Harrell Jr., F. E. (2001). *Regression Modeling Strategies*. Springer.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. Chapman and Hall.

Petricevic, S. et al. (2009). Validation and development of interpretation guidelines for low copy number (LCN) DNA profiling in New Zealand using the AmpFℓSTR SGM Plus(TM) multiplex. *Forensic Science International: Genetics In Press, Corrected Proof*.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2009). Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics 3*(4), 222–226.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2010). Evaluating the weight of evidence using quantitative STR data in DNA mixtures. *Journal of the Royal Statistical Society. Series C, Applied statistics*. In Press.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (4 ed.). Springer.

6.5 Supplementary remarks

Several authors and commentators in forensic genetics have already accepted the model above as a mean to estimate the probability of allelic drop-out (Balding and Buckleton, 2009; Petricevic et al., 2009; Gill and Buckleton, 2010a,b). However, as with any piece of science and each model criticism has also been put forward. Balding and Buckleton (2009) argue that the drop-out probability of a homozygous allele, $P(D_2)$, should satisfy the property that $P(D_2) < P(D)^2$, where $P(D)$ is the drop-out probability of a heterozygous allele for the same DNA profile. Their argument is based on the fact that the superposition of two low intensity peaks should have smaller drop-out probabilities than when peaks are considered separately. That is, allelic drop-out may occur due to absence of molecules associated with a particular allele, but may also be due to the insufficient amount of molecules to trigger the observation of an allele. In the latter case, the amount of DNA might imply that heterozygous alleles yield peak height observations close to 50 rfu while homozygous alleles are closer to 80 rfu.

Balding and Buckleton (2009) suggested that $P(D_2) = \alpha P(D)^2$ for some value of $\alpha < 1$, and was chosen since it satisfy the their requirement. Based on a survey from some forensic laboratories Balding and Buckleton (2009) suggest $\alpha = 0.5$. However, there are at least two problems with the α-approach. First, there is not a solid model behind the suggestion, and second, how do one choose the correct value for α? The model fitted to the experimental data in Tvedebrink et al. (2009) has $P(D_2; H) > P(D; H)^2$ for $H > 136$ rfu. However, the differences are in the fourth decimal place and has no practical implications. D. J. Balding (personal communication, 2010) suggested to use $\sqrt{H}$ rather than $\log(H)$. This transformation yields a slightly better fit to the data and postpone the issue of $P(D_2; H) > P(D; H)^2$ to H-values > 201 rfu.

Gill et al. (2005) demonstrated how to simulate DNA mixtures by mimicking the procedure carried out by a forensic laboratory: DNA extraction, aliquot sampling, PCR efficiency and measurement variability. A similar approach is listed below:

(1) Assume that there are N chromosomes extracted for typing.
(2) Of these do $n_{(0)}$ carry the specific allele of interest, where $n_{(0)} = \text{bin}(N, x/46)$ where $x = 1$ for heterozygous and $x = 2$ for homozygous alleles, respectively.
(3) The PCR process is assumed to be a binomial process: $n_{(c)} = n_{(c-1)} + \text{bin}(n_{(c-1)}, \pi_{\text{PCR}})$, for $c = 1, \ldots, C$, cycles, where π_{PCR} is the PCR efficiency for each cycle in the PCR process.
(4) If $n_{(C)}$ measured with noise gives reason to peak heights lower than a given threshold we declare a drop-out.

By running (1)-(4) several times with varying initial values N we get an simulated distribution of $P(D)$. In Figure 6.2 simulations for heterozygous and homozygous alleles are simulated for varying amounts of DNA. In these simulations $\pi_{\text{PCR}} = 0.85$, $C = 28$ and each point is based on 5,000 simulations. The solid curve is fitted to the heterozygous data points (open points) by $\text{logit}\, P(D; H) = \beta_0 + \beta_1 \log(H)$ and demonstrates that the model fits the data well over the whole range of the response. Dashed curves show the same regression with $\sqrt{(}H)$ as covariate, and the probit approach is discussed below. The fitted parameters $(\hat{\beta}_0, \hat{\beta}_1)$ were used to draw the curves for the homozygous simulations (closed points) with $\log(2H)$ as covariate. The plot shows good agreement between the simulation homozygous data points and the model predictions. The

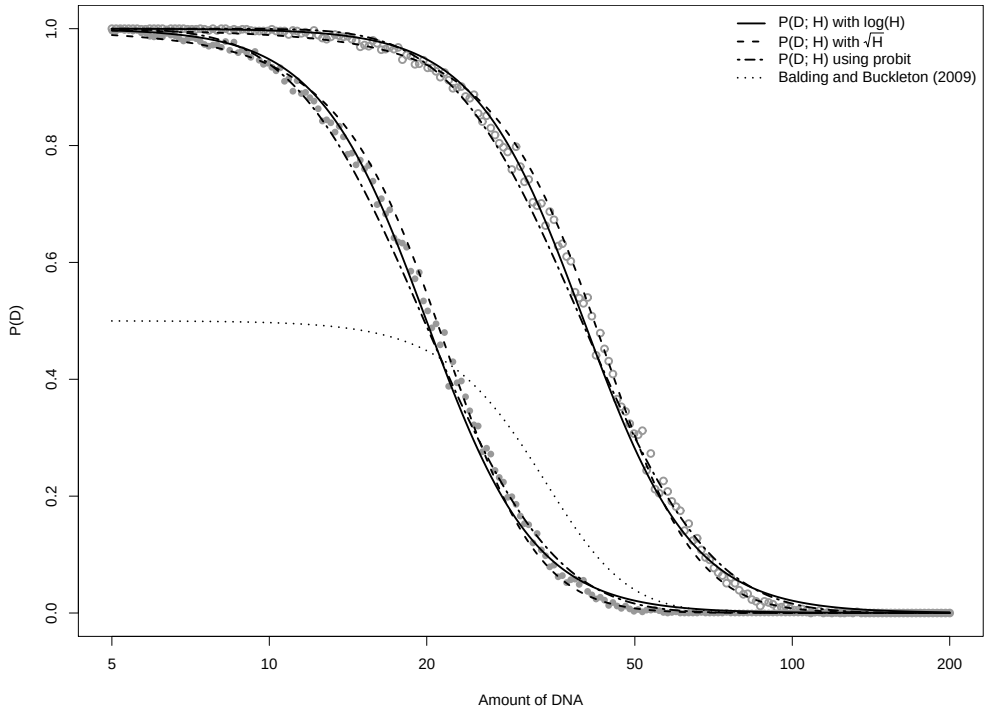

Figure 6.2: Simulations using (1)-(4) for varying amounts of DNA. Open points are heterozygous simulations, and closed points homozygous. The curves are explained by the legend.

dotted curve represents the $\alpha P(D;H)^2$-model of Balding and Buckleton (2009) with $\alpha = 0.5$. The impression is quite different from the logistic regression fitted to the data.

Another way to model the probability of allelic drop-out may be derived taking a slightly different approach than above. Let X denote the number of molecules in a aliquot sampled for PCR. We assume that if X is less than some threshold M the signal will not be sufficiently strong to trigger the CCD camera and thus the signal will be undetected implying allelic drop-out.

Assume that the aliquot is sampled from a total number of molecules N in the extract. Furthermore, the spacial pattern is Poisson distributed with intensity λ (the λ parameter reflects the concentration of molecules), which implies the position of the molecules are independent of each other. The aliquot proportion p of molecules for PCR processing is sampled from this extract, which again is Poisson distributed with intensity $p\lambda$.

Now, X is Poisson distributed with some unknown intensity $\beta = p\lambda$, since p is also unknown. We assume that the average peak height, H, is proportional to the number of sampled molecules, X, such that $H \approx kX$, $X \sim \text{Poisson}(\beta)$, which implies that $\mathbb{E}(X) = \beta \approx k^{-1}H = cH$. Rather than assuming $P(D) = P(X = 0)$, i.e. drop-out only happens when no molecules are samples, we

allow a positive number of molecules to be sampled:

$$P(D;H) = P(X \leq M) = P\left(\frac{X - cH}{\sqrt{cH}} \leq \frac{M - cH}{\sqrt{cH}}\right) \approx \Phi\left(\frac{M - cH}{\sqrt{cH}}\right),$$

where the approximation of a Poisson distribution by the normal distribution is satisfied by the large value of cH. This implies that $\text{probit}[P(D;H)] = \Phi^{-1}[P(D;H)] = \beta_1 H^{-1/2} + \beta_2 H^{1/2}$. Fitting this model using the same dataset as used in Tvedebrink et al. (2009) yields a similar fit as that of the original article. Similarly, did this approach indicate good agreement with the simulations discussed above (as shown in Figure 6.2). Hence, this method which is more closely related to the biochemistry than the logistic assumption adds further support to the logistic regression approach through the similarity in results.

The Section of Forensic Genetics, University of Copenhagen, conducted after the publication of Tvedebrink et al. (2009) more experiments with dilutions of DNA profiles. These experiments investigated the applicability of the drop-out model to different DNA genotyping kits and varying number of cycles in the PCR process (see summary of the results in Table 6.6).

Table 6.6: Summary of the experiments with diluted samples using the SEfiler kit (Applied Biosystems). Samples from identical aliquots were used in order to compare the effect of increasing number of PCR cycles.

Cycles	Classification	D3	vWA	D16	D2	D8	SE33	D19	TH0	FGA	D21	D18
28	Observed	151	152	116	139	153	127	134	148	115	130	108
	Drop-outs	17	16	11	29	15	19	15	20	10	16	17
29	Observed	165	156	125	162	165	141	144	160	120	141	122
	Drop-outs	7	16	5	10	7	9	8	12	8	9	6
30	Observed	170	168	127	168	168	148	151	167	126	148	124
	Drop-outs	2	4	3	4	4	2	1	5	2	2	4

The overall properties of the model did not change, only an extra term caused by the cycle-factor was included:

$$\text{logit}P(D;H,C) = (\beta_{0,s} + \gamma_{0,C}) + (\beta_1 + \gamma_{1,C}) \log(H)$$

Hence, the overall interpretation of the model is the same. It is worth mentioning that $\hat{\gamma}_{0,30} < \hat{\gamma}_{0,29} < \hat{\gamma}_{0,28} = 0$ and $0 = \hat{\gamma}_{1,28} < \hat{\gamma}_{1,29} < \hat{\gamma}_{1,30}$. This implies that for the same locus and $H_0 > 60$ rfu fixed:

$$P(D;\hat{H} = H_0, C = 28) < P(D;\hat{H} = H_0, C = 29) < P(D;\hat{H} = H_0, C = 30).$$

This seems counter-intuitive since more PCR cycles implies higher peaks. However, with $\hat{H} = H_0$ for all three levels of C, it is more likely to have drop-out for $C = 30$ than $C = 28$ since one would expect that peaks with 30 cycles on average are higher than peaks from a 28 cycle PCR

process. In Figure 6.3 a plot similar to Figure 6.1 summaries the estimated model. Each panel shows a box plot of the drop-out events for the associated $\hat{H}$-estimate with the fitted logistic curves superimposed.

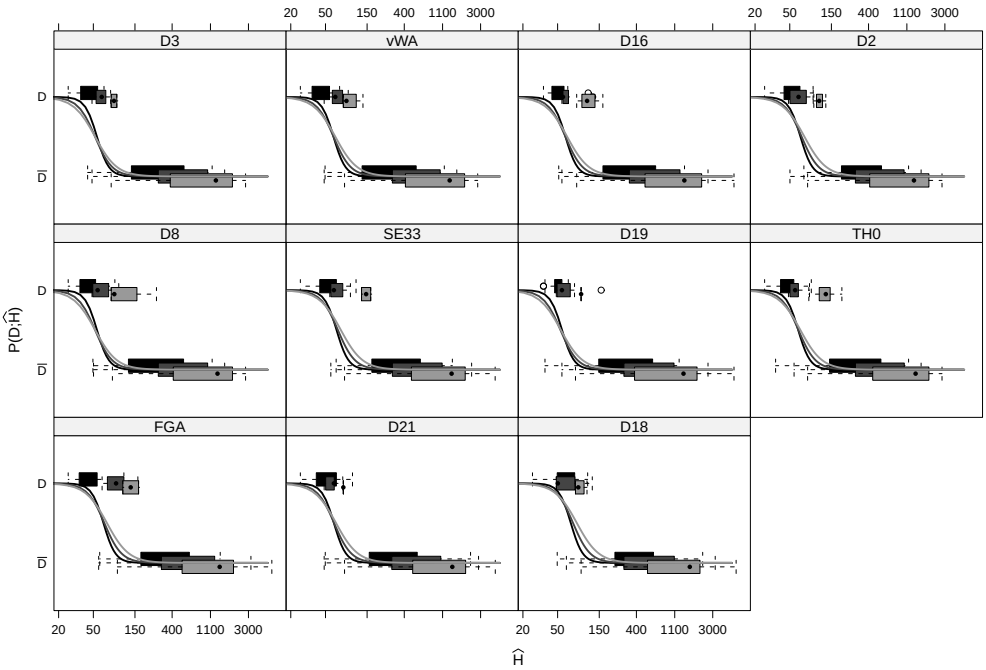

Figure 6.3: Box-plots of the $\hat{H}$-estimates stratified by drop-out event. The boxes are vertically shifted for visual comprehension. Black: 28 cycles, dark gray: 29 cycles and light gray: 30 cycles.

For low amounts of DNA there might be a potential bias when estimating H. This is due to the fact that for a profile with many drop-outs, the peaks with heights above the detection threshold, are "outliers" with respect to the peak height distribution. Hence, estimates based on these observations tends to systematically overestimate the amount of DNA through biased estimates of H. However, for a moderate number of drop-outs the bias is of minor concern.

6.5.1 Mixture separation allowing for allelic drop-out

The models for DNA mixtures presented in Chapters 4 and 5 do not allow for allelic drop-out in their original formulation. However, they are both expendable for handling this sort of issues, in particular the mixture separating model which is discussed in details below.

For two-person mixtures, the possibility of allelic drop-out implies that for loci with three and two observed alleles, the possible list of contributing DNA profiles should be extended with "wild cards". That is, observing alleles a, b and c, genotypes involving an additional allele different from the three needs to be considered. In practice, this implies that $\mathcal{J}_2$ and $\mathcal{J}_3$ from Table 5.3 should be extended as shown in Table 6.7 (denoted $\mathcal{J}_2'$ and $\mathcal{J}_3'$). Note that the columns in $\mathcal{J}_2'$ assuming allelic drop-out are identical to those of $\mathcal{J}_3$ and $\mathcal{J}_4$ with the first row(s) (lowest peak heights/smallest peak areas) removed. Similarly the additional column in $\mathcal{J}_3'$ relative to $\mathcal{J}_3$ refers to the lowest peak intensity of the minor contributor has dropped out (first row of $\mathcal{J}_4$). The number of expected alleles, N_s, is given over each block, where the number of drop-outs equals $N_s - n_s$.

Table 6.7: Extension of $\mathcal{J}_2$ and $\mathcal{J}_3$ of Table 5.3 allowing for drop-outs. The number of drop-outs is equal to $N_s - n_s$, where N_s is the expected number of alleles.

Expected number of alleles:	N_s=2								N_s=3								N_s=4	
Number of drop-outs:	0								1								2	
$\mathcal{J}_2'$:	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
$A_{s,(1)}$	1	1	2	0	1	0	0	1	0	1	1	0	0	1	0	1	0	1
$A_{s,(2)}$	1	1	0	2	1	2	2	1	0	1	0	2	1	1	2	0	0	1

Expected number of alleles:	N_s=3								N_s=4	
Number of drop-outs:	0								1	
$\mathcal{J}_3'$:	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
$A_{s,(1)}$	2	0	1	0	1	0	0	1	1	0
$A_{s,(2)}$	0	1	1	0	0	1	0	1	0	1
$A_{s,(3)}$	0	1	0	2	1	1	2	0	0	1

The statistical formulation of mean and variance are identical, drop-out allowed or not. However, due to the missing data problem induced by the assumption of allelic drop-out we must impute the missing data. An ad-hoc way to do this has been implemented and showed reasonably good results:

$N_s = 4$ and $n_s = 3$: The missing data is imputed by repeating the $A_{s,(1)}$-data row.

$N_s = 4$ and $n_s = 2$: The algorithm will always choose the leftmost configuration of $\mathcal{J}_2'$ since D_s will be similar to that of the rightmost configuration, i.e. the difference between the observed and expected peak areas is similar when conditioned on the locus sum, $A_{s,+}$. However, the ratio of the likelihood values will approximately be $P(D_s|H^{(1)})^2$ due to the two drop-outs for $N_s = 4$.

$N_s = 3$ and $n_s = 2$: There are four different configurations that need to be considered (numbers refer to order in $\mathcal{J}_2'$ where $N_s = 3$):

(1) If a homozygous allele drops-out the situation is the same as above for $N_s = 4$ and $n_s = 2$.

(2) The missing data is imputed by repeating the $A_{s,(1)}$-row.

(3) The missing data is imputed as the difference of the $A_{s,(3)}$- and $A_{s,(2)}$-row.
(4) The missing data is imputed by repeating the $A_{s,(1)}$-row.

A more rigorous approach would be to use the EM-algorithm, however, for practical purposes it is believed, that there would be no substantial difference.

Furthermore, the likelihood now includes an extra term $P(D;H)$, where H is calculated as described above. Assume that only alleles of the minor contributor have dropped out, then allowing for drop-outs in mixture separation, implies that the selection criterion needs to evaluate $P(D;\hat{H}{=}H^{(1)})^{n_{D_1}}P(D;\hat{H}{=}2H^{(1)})^{n_{D_2}}\tau^{-N}$, where n_{D_1} and n_{D_2} respectively are the number of heterozygous and homozygous drop-outs.

Example

Table 6.8 lists the STR data of a two-person mixture where several allelic drop-out has occurred. In fact only three of the minor profile's (marked by ○ in Table 6.8) alleles not shared by the major profile (●) had peak heights above the 50 rfu limit of detection.

Table 6.8: Observed STR data of a two-person DNA mixture. The estimated H-values were respectively 60.3 rfu (Profile ○) and 765.8 rfu (●). The profiles denoted by squares and triangles are respectively iditified without drop-out allowed and taking drop-out into consideration.

Locus	Allele	Profiles	Height	Area
D3	15	● ■ □ ▲ △	884	7787
D3	16	● ○ ■ □ ▲	816	7140
D3	19	○ △	-	-
vWA	15	● ○ ■ □ ▲ △	519	5067
vWA	17	● ○ ■ □ ▲ △	530	4928
D16	10	○	-	-
D16	11	● ■ □ ▲ △	1373	14302
D16	12	○	-	-
D2	19	● ■ □ ▲ △	565	6635
D2	23	○ △	-	-
D2	25	● ○ ■ □ ▲	518	6120
D8	8	● ■ □ ▲ △	993	8720
D8	12	● ■ ▲	807	7320
D8	13	○ □ △	78	891

Locus	Allele	Profiles	Height	Area
D21	28	○	-	-
D21	29	● ■ □ ▲ △	773	6867
D21	30	○	-	-
D21	31	● ■ ▲ △	637	5867
D18	12	○	-	-
D18	15	● ■ □ ▲ △	762	8449
D18	16	○ □ △	52	663
D18	17	● ■ ▲	644	7316
D19	13	● ○ ■ □ ▲ △	1163	9550
D19	15	○ □ △	51	631
TH0	6	● ○ ■ □ ▲ △	553	4691
TH0	7	○	-	-
TH0	8	● ■ □ ▲ △	572	4936
FGA	20	○ △	-	-
FGA	23	● ○ ■ □ ▲ △	403	4024
FGA	24	● ■ □ ▲	363	3651

The drop-out probabilities of one of minor profile's alleles for the various loci are listed in Table 6.9. From this table we see that it is likely that one or more of the minor profile's alleles has dropped out.

Table 6.9: Drop-out probabilities for the minor contributor (see Table 6.8). The probabilities were computed using $(\hat{\beta}_{s,0}, \hat{\beta}_1)$ from Tvedebrink et al. (2009).

Locus	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
$P(D\|H^{(1)}$=60.3)	0.61	0.65	0.72	0.62	0.61	0.41	0.56	0.83	0.83	0.80

We may analyse the data of Table 6.8 using the mixture separating algorithm. We first analyse the data assuming no allelic drop-out and next allowing for allelic drop-out. The likelihood value when not allowing for allelic drop-outs is 1.908×10^{-10} which corresponds to $\hat{\tau} = 7.65$. The identified profiles when drop-outs are neglected are denoted by square-symbols in Table 6.8. Similarly, when allowing for drop-outs $\hat{\tau} = 5.82$ and the likelihood value is 1.178×10^{-9}. Since the best matching profiles when allowing for drop-outs has three drop-outs (see the identified profiles in Table 6.8 marked by triangles) the likelihood value is computed as $\hat{\tau}^{-N} P(D_{\text{D3}}|H^{(1)})P(D_{\text{D2}}|H^{(1)})P(D_{\text{FGA}}|H^{(1)})$, with the locus specific drop-out probabilities listed in Table 6.9.

Even though the algorithm allowing for drop-out correctly identified three allelic drop-outs, the number of correctly identified loci for the two methods are almost the same. This is due to the peak height imbalances of the peaks of the major profile. Often the shared allele of the two true contributors is smaller than the one where the major profile is the only contributor (e.g. loci D3, D2 and TH0 in Table 6.8). Hence, for D3 and D2 the larger of the two observed alleles is assumed to be a shared allele.

In addition to the example above, we also simulated data mimicking a two-person DNA mixture. The minor contributor had a fixed mean peak height of 60 rfu, while the major component had peak heights of 2000, 1500, 1000, 750, 500, 250, 150, 100 and 75 rfu. The reason for decreasing the peak height of the major contributor is that since the variance is proportional to the mean, the contribution from the minor component is *masked* in the variability for large peak intensities. That is, it is not possible to detect whether the minor component has dropped out or if it shares alleles with the major profile. Hence, the smaller the peak intensities of the major, the easier it should become to detect allelic drop-out. For each mean value of the major peak height, the standard deviation, τ, takes integer values from 0 to 10. The locus where an allele of a four-allele locus has dropped out, the methods detects most of the drop-outs for small values of τ and the mean values of the major profile. For three-allele loci the method is less successful even for moderate values of τ and low major peak heights.

Both experimental data and the simulations indicate that it is difficult to identify allelic drop-out of a contributor to a DNA mixture. The problem with allelic drop-out is not only due to limited amount of DNA in the sample. Lowering the limit of detection naturally decreases the number of drop-outs. However, this might come with the cost of increased drop-in peaks. In the following paper we discuss how the background noise can be used to determine a limit of detection using the sample itself as reference.

CHAPTER 7

Sample and investigation specific filtering of quantitative data from STR DNA analysis

Publication details

Co-authors: Poul Svante Eriksen[*], Helle Smidt Mogensen[†] and Niels Morling[†]

[*] *Department of Mathematical Sciences*
Aalborg University
[†] *Section of Forensic Genetics, Department of Forensic Medicine*
Faculty of Health Science, University of Copenhagen

Journal: International Journal of Legal Medicine (Under preparation)

Abstract:
The discrimination between positive and negative results in forensic genetic STR DNA analyses is of outmost importance. We present a method for identification of STR alleles that is based on (1) discrimination between positive and negative STR results that are specific for the sample and each STR locus, (2) correction of stutter effects and (3) correction of pull-up effects. The sample and STR locus specific discrimination was based on a floating threshold that was estimated by means of distribution analysis of the true negative data elements, i.e. the noise component. The correction of stutter effects and pull-ups was based on regression analysis. The method was developed on the basis of STR data of serial dilutions of DNA from four persons in amounts ranging from 24.6 to 410 pg DNA. The method was tested on two types of data: (1) controlled experiments with two-person mixtures of DNA in proportions 16:1, 8:1, 4:1, 2:1, 1:1, 1:2, 1:4; 1:8 and 1:16 with a total of 328 to 528 pg DNA in the two-person mixtures, and (2) data from fingernail swabs from real crime cases.

The method yielded a 16% increase in allele assignment compared to that of a conventional assignment of STR alleles for the two-person DNA mixtures and 24% increase for the fingernail data. A further gain from the method was a more precise identification of the STR types of contaminated or otherwise compromised DNA samples.

Keywords:
STR typing; Allele assignment; Investigation specific floating threshold; Stutters; Pull-up effects.

7.1 Introduction

DNA typing with Short Tandem Repeat (STR) alleles is typically based on multiplex Polymerase Chain Reaction (PCR) amplification of the relevant STR DNA, capillary electrophoresis and fluorescence detection of the resulting PCR products. In European DNA crime case laboratories, the AmpFℓSTR SGM Plus kit (Applied Biosystems - AB) is widely used. The discrimination between positive and negative STR results may rely on the individual judgement of the scientist responsible for the STR typing or on fixed criteria like a cut-off of 50 relative fluorescent units (rfu) between positive and negative responses as recommended by the supplier of the kit. A fixed cut-off level may be very useful for practical routine work, but fixed cut-off values may ignore specific circumstances of the investigation being carried out and introduce errors in the interpretation of the results.

A fixed cut-off of 50 rfu is used in many laboratories for the analysis of routine results although other methods based on e.g. the signal-to-noise ratio may be used (Gilder et al., 2007). A number of factors influence the general magnitude of the fluorescent signal, e.g. (1) the amount of amplifiable DNA in the PCR, (2) the amount of fluorochrome molecules bound to the oligo-DNA molecules acting as primers in the PCR, (3) the number of PCR cycles, (4) the amount of detectable, amplified PCR products injected into the electrophoresis capillary typically controlled by the injection voltage and injection time, (5) the sensitivity of the fluorescent detections system and (6) other factors. The level of irrelevant signals, the noise component, is determined by factors like impurities of the fluorochrome not attached to the amplified DNA molecules, impurities

of the primers and conglomerates of the primers, conglomerates of fluorochromes and other substances that may be present in the post-PCR reaction volume that is injected into the capillaries. The PCR amplification with the fluorescent primers is usually responsible for the majority of the noise signal. Large variations may exist from kit to kit and from batch to batch. Other contributors to irrelevant noise signals include impurities in the DNA preparations and other chemical reagents than the STR kit, the DNA sequencer equipment, including the electronic detection and amplification system and other components of minor importance.

In multiplex PCR STR kits, a number of fluorochromes are typically used, and the balance between signal and noise between the fluorochromes vary. The signal intensities of the various STR loci also vary. Thus, systematic variations such as the STR kit, the batch, batches of other reagents, the DNA sample, the DNA sequencers with attached electronic equipment, etc., may influence the level of discrimination between positive and negative reactions in STR typing. Therefore, it is desirable to develop methods that can determine the threshold between positive and negative reactions for each of the investigated DNA samples and for each STR locus.

Systematic extra reactions such as "stutters", which are caused by infidelity of the Taq polymerase in the PCR resulting in amplification products typically 4 base pairs (bp) shorter than the true PCR products, and "pull-ups" that are caused by spectral overlap of the fluorochromes used for the detection of the PCR products must also be handled during the interpretation of the STR results. Stutters are often compensated for by ignoring results below a certain ratio of the signal of the "parental peak", i.e. the DNA fragment supposed to cause the stutter signal. Stutter filter ratios are usually decided for each STR locus based on average data from initial investigations of small numbers of samples performed by the supplier of the STR kit and, thus, not necessarily optimal for all laboratories and/or all alleles in an STR locus. Correction for pull-ups is most often done by visual inspection although IT-based expert programmes may be used to remove signals that most likely are caused by pull-up effects.

We have developed a new method for the discrimination between positive and negative STR results based primarily on analyses of the "noise component" of the data that represent true negative STR results. The positive results were further analyzed to identify and correct for stutters and pull-up effects. We used distribution analysis of the noise component in order to separate the negative and positive results. Algorithms based on regression analysis were developed to correct for stutter effects of each STR locus and pull-up effects of each sample. The method makes it possible to analyze the STR results of a sample according to the results that are specific for the sample and each STR locus.

The results of the method were compared to those obtained by the method recommended by the manufacturer of the SGM Plus kit with fixed cut-offs for positive reactions (50 rfu) and stutters.

7.2 Materials and methods

7.2.1 Data

Controlled experiments

The laboratory investigations were performed at the Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen and approved by the local ethical committee (KF-01-037/03). Genomic DNA from blood samples from two males and two females was extracted by a standard phenol-chloroform extraction method. The DNA was quantified in triplicates using the Quantifiler Human DNA Quantification kit (AB) with Human Genomic DNA Male (Promega) as the quantification-standard on an ABI Prism 7000 (AB). The median DNA concentrations were used. Each sample was diluted in water to approximately 500 pg DNA/μl. The DNA concentrations in the diluted samples were measured again in triplicates and the medians of the DNA concentrations were recorded for further use.

DNA from each of the four persons was serially diluted with water in the proportions 16:1, 8:1, 4:1, 2:1 and 1:1 for the training STR data set, cf. below. Pair-wise two-person mixtures of DNA (w/v) in proportions 16:1, 8:1, 4:1, 2:1, 1:1, 1:2, 1:4; 1:8 and 1:16 were made of DNA from each pair of the four persons for the validation STR data set, cf. below.

The amount of DNA from each person in the diluted single donor samples and the two-person mixtures was calculated based on the measured DNA concentration of each sample. The total amount of DNA ranged from 24.6 to 410 pg DNA in the dilutions of single donor samples and from 328 to 528 pg DNA in two-person mixtures. The DNA was amplified with the AmpFℓSTR SGM Plus kit (AB) in a GeneAmp 9700 PCR thermocycler (AB) as recommended by the manufacturer.

One ul of the amplificate in 15 ul HiDi Formamide (AB) was analysed on an ABI Prism 3100 Genetic Analyzer (AB) using POP4 (AB) as the polymer and 3 kV injection voltage for 6 seconds.

DNA fragments were detected and fragment sizes were estimated with GeneScan 3.7 (AB). Genotypes were assigned using GenoTyper 3.7 (AB). The data were analyzed in two ways: (1) With the method recommended by the manufacturer of the SGM Plus kit with a fixed cut-off of 50 Relative Fluorescence Units (rfu) for the discrimination between positive and negative results and the recommended stutter filter, and (2) by the presented floating threshold method. The data for the floating method were generated by GeneScan 3.7 with a detection threshold of 5 rfu. Genotypes were assigned using GenoTyper 3.7 with no stutter filter applied.

Crime scene fingernail swabs

In addition to the DNA mixtures from controlled experiments with known contributors, the methodology were also tested on samples from real crime cases. DNA transfer between a victim and suspect frequently occurs during violent crimes. Debris from fingernail swabs are routinely analysed in many crime cases such as rapes, assaults and other violent crimes. Often the contri-

bution of DNA from the suspect is limited and low peak intensities of alleles associated with the suspect's DNA profile will be produced on analysis.

Data from fingernail swabs from 98 real crime cases were analysed (1) using the standard protocol with a 50 rfu cut-off and the recommended stutter filter (2) and a 5 rfu detection threshold and no stutter filters, similar to that of the experimental data.

With DNA from a single person, the SGM Plus STR kit detects 10 STR loci and X- and Y-specific (amelogenin) DNA fragments resulting in a maximum of 22 data elements representing DNA fragments. The data from single donor dilutions (training data) were used for developing the various mathematical models, while the data from the two-person mixtures and crime case samples from fingernail swabs were used to evaluate the performance/efficiency of the mathematical models.

7.2.2 Data model

The statistical model was derived from STR data from each of four donors (training data). The model assumes the following major components of STR data:

(1) True positive results from the STRs.
(2) So-called "stutters" (cf. below) that are DNA artefacts created during the investigations.
(3) So-called "pull-ups" (cf. below) that are artefacts created during the detection of the various signals of the fluorochromes that contribute to the identification of the signals of each of the STR alleles.
(4) Back-ground noise of various kinds.

The quantitative STR data is a mixture of contributions from various sources. Apart from the signals from true alleles, the signals consist of at least two error components from (1) the PCR amplification (stutters) and (2) the measurement technique (pull-ups). Stutters are PCR products typically four base pairs (bp) shorter and to some extend four bp longer than the true PCR product ("back-stutters"). Stutters originate from primer mis-pairing in the PCR amplification creating PCR products that mimic alleles typically one repeat shorter than the true peak (Butler, 2005). In Figure 7.1, the stutter products are shown with effects on both the true peaks ("Signal + stutter") and on the noise. In the "double stutter" situation, the stutter peak is caused by stutter effects from both peaks to the right of the stutter peak. This causes the double stutter peak to be larger than single stutter peaks because a double-stutter peak is the stutter product of two peaks. However, the effect of double stutters is not directly implemented in our model, but it is accounted for in the parameter estimates used in the algorithm (Section 7.2.4).

The quantitative signals are obtained by a very sensitive photocell (CCD camera) detecting the intensity of light emitted from a fluorochrome on DNA molecules corresponding to alleles of each STR locus. The signal intensities are measured as rfu. Due to noise in the apparatus, the observed signal contains a continuous noise part that we denote background noise (peaks designated "Noise" in Figure 7.1). The light-detecting system also causes a systematic error component, namely the pull-up effects. This is caused by overlap of the spectra of the light emitted from the various fluorochromes as illustrated in Figure 7.2. The pull-up effect is observed

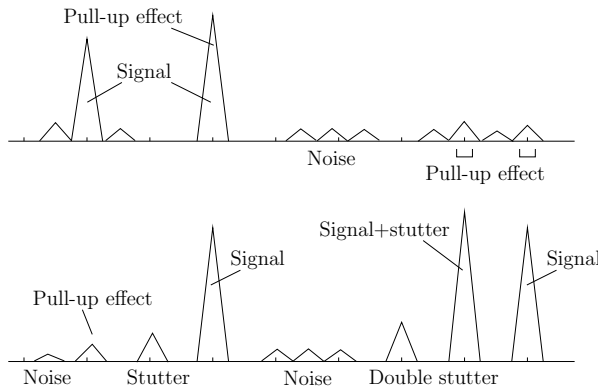

Figure 7.1: Picture of the non-signal components of a STR DNA trace.

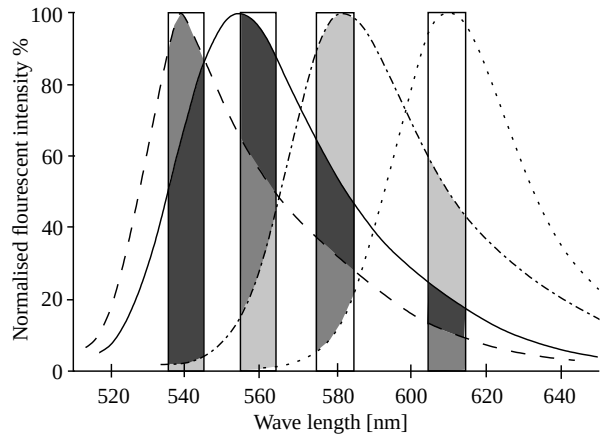

Figure 7.2: Fluorescent dye bands: Blue (dashed/semi-gray), green (solid/dark-gray), yellow (dot-dashed/light-gray) and red (dotted/white). The shaded areas under each curve indicate the amount of spectral overlap between the various dyes. Reproduced from Applied Biosystems (2000).

as an increase in the intensities of both the background noise and true peaks. The shaded areas in Figure 7.2 represent the amount of overlapping light frequencies of the four different colours (blue, green, yellow and red) used in the SGM Plus kit. The increases caused by pull-up effects are pictured in Figure 7.1 as “Pull-up effect”.

In our model, pull-up effects cannot cause stutters, whereas stutters may induce pull-up effects on other dye bands.

7.2.3 Determination of floating threshold

For each STR locus and amelogenin of a sample, we wanted to obtain a set of negative data in order to model the negative data elements and develop a threshold for discrimination of positive and negative signals based on the distribution of the negative signals. We used the training data set for the development of the mathematical model. The peak height observations (intensities of fluorescent signals) below 5 rfu were removed at the first step of analysis with the Genescan software because the software was unable to handle the large amount of data elements below 5 rfu. The remaining signals comprised both background noise and more systematic components. We removed (1) all peaks on the allelic ladder that primarily represent true alleles and (2) all off-ladder signals in pull-up positions. This ensured that the remaining data points represented true noise. The peaks designated "Noise" in Figure 7.1 illustrate the data used for the determination of the threshold.

Inspection of the data indicated that the noise followed a right-skewed distribution. In order to obtain a normal distribution, the peak heights were transformed by $\log_e$(peak height − 4.5). The distribution of the noise data fitted the log-normal distribution, and the fit was not better with distributions like the exponential, Fisher-Tippett, Pareto, Rayleigh, or Weibull distributions. Figure 7.3 shows the distribution of the observed peak heights of the noise for each locus of a sample after the data had been transformed by $\log_e$(peak height in rfu − 4.5) against a standard normal distribution in a QQ-plot. Note that the "outliers" in the upper tail of the distribution are in fact the true positive signal. The plots demonstrated that the noise (shifted by −4.5) followed a log-normal distribution with individual parameters μ_s and σ_s for each locus, s. These parameters determined the intercept and slope of the superimposed QQ-line and were estimated by

$$\hat{\sigma}_s = \frac{x_{s(q_1)} - x_{s(q_0)}}{z_{(q_1)} - z_{(q_0)}} \quad \text{and} \quad \hat{\mu}_s = x_{s(q_0)} - \hat{\sigma}_s z_{(q_0)},$$

where $x_{s(q)}$ and $z_{(q)}$ are the empirical and standard normal q-quantiles, respectively. We used these quantile estimators rather than the ordinary maximum likelihood estimators in order to increase the robustness of the method.

Figure 7.3 shows that the fit to normality was better for the higher values of $\log_e$(peak height in rfu − 4.5) than for the lower ones. The observations in the upper part of the peak heights of the noise are those of main interest for the estimation of the threshold. Thus, we chose to use the $(q_0, q_1) = (50\%, 90\%)$-interval for the estimation of the threshold. The threshold was determined by the mean plus 3.29 times the standard deviation. The locus specific threshold can be written as:

$$\text{Threshold for locus } s = \exp(3.29\hat{\sigma}_s + \hat{\mu}_s) + 4.5.$$

Approximately 99.95% of the noise will be below the threshold and, thus, will be categorized as noise. However, 0.05% of the true negative results will be above the threshold and, thus, will be categorized as positive signals. For practical purposes, the majority of such false positive assignments will be in off-ladder positions rather than in allele positions.

Figure 7.3 shows that only few noise data elements were recorded for amelogenin. This is due to the fact that the interval, in which noise could be recorded around the X- and Y-windows of

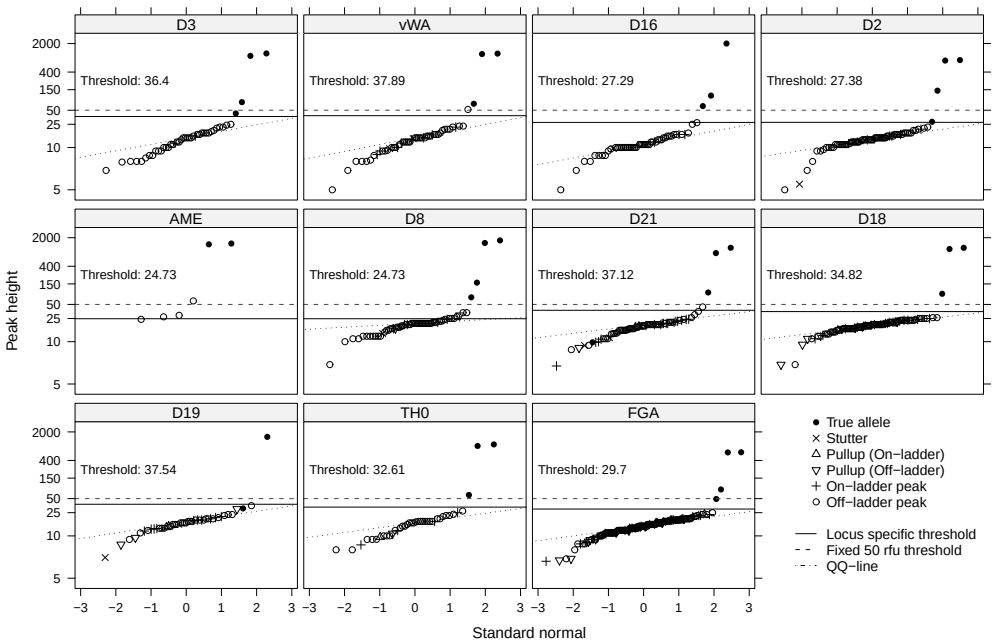

Figure 7.3: QQ-plots of the observed peaks. Note the different thresholds computed using the data of the locus itself as reference. For this particular sample, the fixed 50 rfu-threshold causes five drop-outs (in loci D3, D2, D19, D2 and FGA) and two drop-out (in loci D19 and D21) with the locus specific threshold (one true peak in locus D21 has a peak height of 22 rfu and is embedded in the noise).

amelogenin, is small. However, amelogenin and D8 are marked with the same fluorochrome and D8 alleles are only slightly longer than the DNA fragments of X- and Y-amelogenin. The distributions of noise in amelogenin and D8 were rather similar to each other and, therefore, the threshold of D8 was also used for amelogenin.

7.2.4 Stutter correction

Figure 7.4 shows three different situations involving stutters. The background noise (grey peaks) is the same in all three scenarios, but the parental peaks and, thus, the stutter peaks (black peaks) differ in sizes.

We used the training data set to develop the mathematical model based on a regression model on the peak intensities of the parental peaks. Assuming additivity of the noise and stutter product, we take into account that peaks in stutter positions in front of small peaks mainly consist of noise

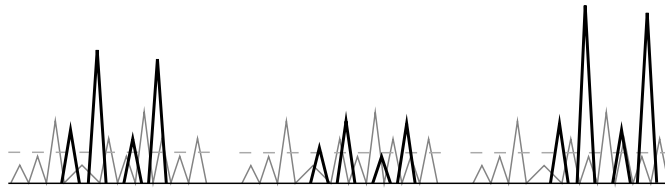

Figure 7.4: Stutter peaks caused by different parental peaks (in black). The grey peaks picture the noise and the dashed line the median of the noise.

as pictured in Figure 7.4. The model of the expected stutter height, h_{Stutter}, is given by

$$h_{\text{Stutter}} = h_{\text{Noise},s} + (\beta_s + \gamma_s \tilde{\text{bp}}_s) h_{\text{Parent}}, \tag{7.1}$$

where $h_{\text{Noise},s}$ is the known/determined median of the off-ladder peaks not in pull-up position on locus s (see Section 7.2.3) and h_{Parent} is the parental peak's height. The parameters were estimated by a weighted least square fit with weights $1/h_{\text{Parent}}$, due to the proportionality of the mean and variance of the peak heights (Tvedebrink et al., 2010). In the latter term, $\tilde{\text{bp}}_s$ is the base pair deviation from the mean base pair, $\bar{\text{bp}}_s$, on locus s, $\tilde{\text{bp}}_s = \text{bp}_s - \bar{\text{bp}}_s$. The parameter β_s is the average stutter effect at a given locus, s. By including the base pairs in the model, we are able to have different stutter fractions for various alleles within a locus, if necessary.

Table 7.1: Estimates of γ_s and β_s in the stutter model (7.1).

Locus	D3	vWA	D16	D2	D8	D21	D18	D19	TH0	FGA
$\beta_s \times 10^2$	7.209	6.714	6.101	7.712	4.996	6.359	7.031	7.252	2.031	6.508
$SE(\beta_s) \times 10^3$	0.930	0.883	0.874	0.848	0.645	0.836	0.776	1.348	1.310	1.221
$\gamma_s \times 10^2$	0.104	0.203	0.212	0.091	0.096	0.075	0.172	0.215	0.069	0.123
$SE(\gamma_s) \times 10^3$	0.112	0.118	0.137	0.064	0.069	0.140	0.099	0.215	0.184	0.139

The previously observed increase in stutter percentage as a function of allele number (Applied Biosystems, 2006) was reproduced by the positive estimates of γ_s in Table 7.1. The STR locus specific β_s parameters in Table 7.1 are in accordance with the picture in the manufacturer's kit documentation (Applied Biosystems, 2006, Figure 9-5, 9-6 and 9-7), where e.g. the average stutter effect, β_{TH0}, in TH0 is the weakest.

Figure 7.5 shows the stutter peak heights predicted by the model compared to the observed stutter peak heights. The plot demonstrates that the model in (7.1) is sufficient in order to describe the stutters. For adjacent heterozygous alleles, the base pairs typically differ by only a limited number of bp, which minimizes the effect of the length of the DNA fragment estimated by γ_s.

Additional examinations of the data also made it clear that back-stutters were present typically in the position 4 bp larger than the parental peak. The model for back-stutters and correction

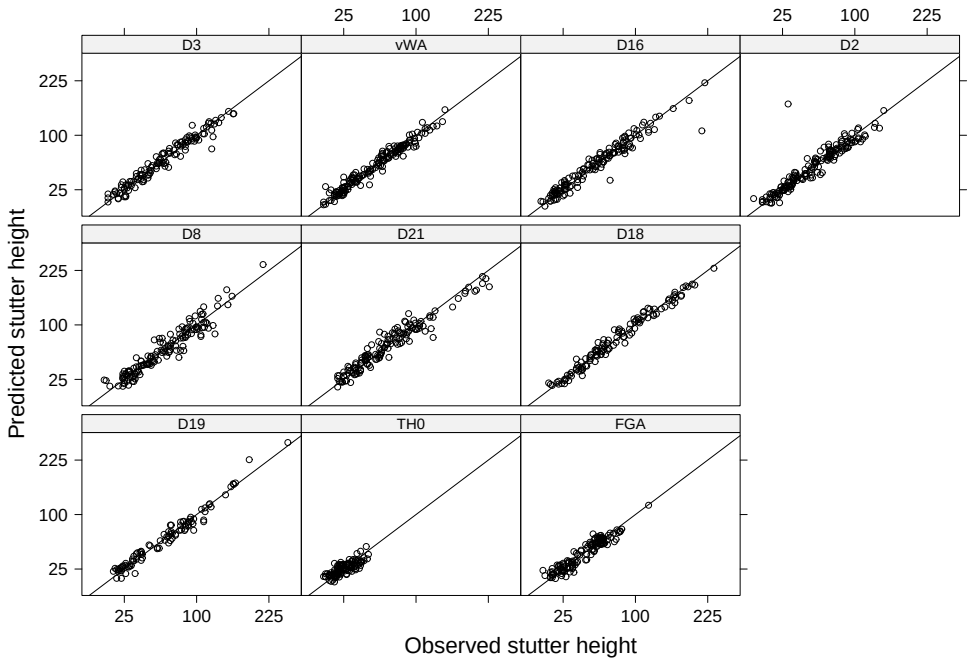

Figure 7.5: Predicted stutter peak heights plotted against observed stutter peak heights with the identity line superimposed. The scale of the plot is the variance stabilising square-root transformation.

is based on the same idea as those concerning conventional stutters with a noise level and an additional effect from the parental peak, i.e.

$$h_{\text{Backstutter}} = h_{\text{Noise},s} + \beta_s h_{\text{Parent}}. \tag{7.2}$$

Table 7.2 shows the parameter estimates. The lack of homozygous alleles in some of the loci in the actual data set implied that the estimates of β_s were insignificant for these loci. This is due to the fact that the parental peak needs to reach a certain height (typically well above 1,000 rfu) for the back-stutter to exceed the noise level. For the same reason, base pairs were not included in the back-stutter model as only a few base pair lengths were represented in the back stutter data.

Double stutters originating from two adjacent alleles separated by 4 bp in a heterozygous individual behave slightly differently from single stutters. In Appendix 7.A, we evaluate the ratio of the stutter peak to the mean of the two parental peaks. In situations where the heterozygous alleles are not adjacent (separated by more than 4 bp) or when a stutter originates from a homozygous allele, for practical purposes, we only need to consider the ratio of the stutter peak to the parental peak.

Table 7.2: Parameter estimates of the backstutter model (7.2). Loci with insignificant β_s estimates due to lack of homozygous alleles were removed.

Locus	D3	vWA	D16	D2	D8	D21	D18	FGA
$\beta_s \times 10^2$	0.233	0.211	0.428	0.187	0.293	0.615	0.560	0.638
$SE(\beta_s) \times 10^3$	0.643	0.738	0.628	0.627	0.551	0.625	0.561	0.747

7.2.5 Pull-up correction

We defined pull-ups as peaks on different dye bands within ±0.5 bp of the parental bp lengths. Only the peaks not being true alleles or possible stutters on a different dye band than the parental peak were included in the data analyses. Figure 7.1 shows an increase of the noise level (rightmost on the upper band) and a true peak (heterozygote imbalance, leftmost on the upper band). We used the training data set to develop the mathematical model based on regression.

Figure 7.6 shows examples of pull-up values as function of the values of the parental peaks for the various colours. The magnitudes of the observed pull-up effects were in accordance with the spectral overlap in Figure 7.2, i.e. the effects of green signals in the yellow spectrum and of green signals in the blue spectrum were the two largest, and yellow signals had the smallest effect in the blue spectrum.

For predictive purposes, we fitted a linear model to the observed data in Figure 7.6. Of the included data points, only a limited subset comprised detectable pull-up peaks, while the remaining observations were background noise in pull-up positions. Our model takes this into account by having a noise dependent intercept, $h_{\text{Noise},s}$, for locus s (median of the noise data described in Section 7.2.3). This approach is similar to the one used in the model for correction of stutter effects. In the formulation of the model, the notation $D \to d$ reflects that the parental peak is in fluorescent dye band D and the pull-up peak is located in the fluorescent dye band, d,

$$h_{\text{Pull-up}} = h_{\text{Noise},s} + \beta_{D\to d} h_{\text{Parent}}, \tag{7.3}$$

where parameters were estimated by a weighted least-square fit. Table 7.3 shows the parameter estimates of $\beta_{D\to d}$. The superimposed lines in Figure 7.6 were based on the parameter estimates of Table 7.3. Thus, the superimposed lines are in accordance with the spectral overlaps in Figure 7.2 except for $\beta_{G\to B}$, which is smaller than expected. This may be due to the particular alleles included in our data set.

Table 7.3: Parameter estimates of the various overlapping fluorescent dyes.

Dye → dye	$B\to G$	$B\to Y$	$G\to B$	$G\to Y$	$Y\to B$	$Y\to G$
$\beta_{D\to d} \times 10^2$	1.039	0.449	0.342	0.978	0.322	0.597
$SE(\beta_{D\to d}) \times 10^3$	0.405	0.357	0.411	0.341	0.560	0.482

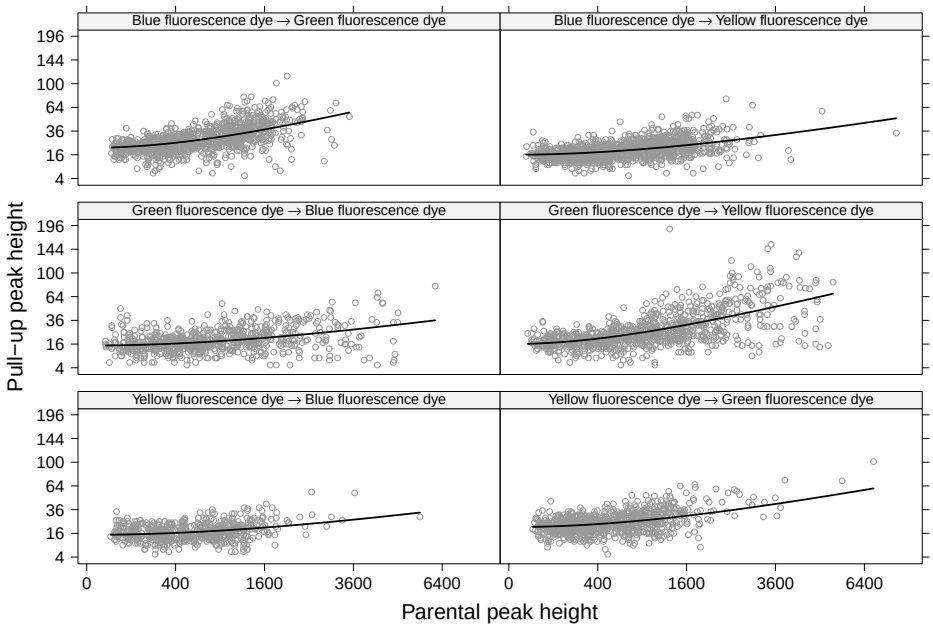

Figure 7.6: Pull-up effects stratified by overlapping fluorescent dyes. The superimposed lines indicate the estimated model. The scale of the plot is the variance stabilising square-root transformation.

7.3 Results

The parameters for correction of pull-up and stutter effects were estimated using the dilutions of non mixture samples, whereas the overall performance of the filter was based on analysis of all possible combinations of pairwise two-person mixtures of four profiles in mixture ratios ranging from 1:16 to 1:1.

The procedure of events were the following:

(1) Determination of the floating threshold: Determine the threshold and detect potential stutters, pull up effects and true peaks, i.e. alleles with peak heights above the threshold.
(2) Pull-up correction: Correcting for pull-up effects caused by peaks above the threshold determined in (1).
(3) Stutter correction: Correcting for stutter effects caused by peaks above the threshold determined in (1).
(4) Allele assignment: Assignment of alleles according to the determined floating threshold in (1) and the allelic ladder.

Note, that the corrections for pull-up effects were made before the stutter correction was applied, because stutters may cause pull-ups while pull-ups cannot make stutters.

7.3.1 DNA mixtures from controlled experiments

We used our floating threshold, stutter and pull-up correction method on 107 two-person mixtures. In Table 7.4, we summarise the performance of the overall filter. It is worth emphasising that 263 of the true alleles dropped out and that the stutter filter let 6 stutters and no backstutters slip through. In addition to the stutter peaks, another 25 (21 drop-ins and 4 pull-ups) on-ladder peaks were classified as proper peaks of the samples.

Table 7.4: Filtered and passed peaks classified by type.

Classification	Assigned negative result	Assigned positive result
True allele	263	3,308
Stutter	2,167	6
Backstutter	1,260	0
Noise	62,669	324
On-ladder	*11,619*	*21*
Off-ladder	*51,050*	*303*
Pull-up	3,825	14
On-ladder	*982*	*4*
Off-ladder	*2,843*	*10*

The remaining peaks passing the filter were all in off-ladder positions and removed from the analysis afterwards. The data were also analysed following the standard protocol of The Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen. Using the technique recommended by the manufacturer, 312 drop-outs were observed together with 27 stutters and 7 pull-up peaks. Thus, the number of drop-outs of true alleles was 16% lower with locus specific filtering than with a fixed 50 rfu threshold.

The classification tables for the two methods are listed in Table 7.5. In the classification tables each observation is categorised by its classification and actual class. In Table 7.5 the diagonals are the correctly classified observations, while the off-diagonals are the misclassified. The lower the counts in the off-diagonal cells the better is the classification methodology.

To summarize the classification table in a single value we suggest the misclassification rate, which is the total of misclassified observations to the total of correctly classified observations. From the misclassification rates (bottom lines in Table 7.5) we see that the floating threshold method yields a better classification than the fixed 50 rfu threshold.

Table 7.5: Classification tables for the two methods: Fixed 50 rfu and floating threshold.

Floating threshold

Expected		+	÷
Observed	+	3308	31
	÷	263	69,921
Misclassification rate:			0.401%

Fixed 50 rfu threshold

Expected		+	÷
Observed	+	3259	34
	÷	312	69,918
Misclassification rate:			0.473%

7.3.2 Fingernail swabs from crime cases

Data from 98 crime cases were analysed using the approach presented. The Section of Forensic Genetics, University of Copenhagen, supplied the data from the crime cases together with reference samples associated with the crime. These reference samples may explain the observed stain, but since contamination from other biological material and debris may accumulate under the fingernails, the number of "random" drop-ins may be misleading (Cook and Dixon, 2006).

From Table 7.6 we see that the number of drop-outs decreased by 220 events from 912 using the standard protocol to 692 using the samples specific setup (decrease of 24%). However, this gain in fewer drop-outs comes with a cost in more drop-ins. The standard protocol gave 15 drop-ins versus 90 using our approach. In the experiment conducted by Cook and Dixon (2006), foreign DNA were detected in 13% of the fingernail swabs taken from the participating individuals. Hence, the higher number of drop-ins using our more sensitive methodology may be caused by foreign DNA. These alleles are actually true alleles rather than drop-ins, however, this is impossible to conclude from the available data.

Table 7.6: Classification tables for the two methods: Fixed 50 rfu and floating threshold.

Floating threshold			
Expected		+	÷
Observed	+	1,460	90
	÷	692	82,497
Misclassification rate:			0.931%

Fixed 50 rfu threshold			
Expected		+	÷
Observed	+	1,240	15
	÷	912	82,572
Misclassification rate:			1.106%

The misclassification rates in Table 7.6 are more than twice the rates of Table 7.5. This is a consequence of the data being from real crime cases with many degraded samples and low amounts of DNA. Hence, the number of drop-outs is larger and so is the number of partial DNA profiles.

In addition Table 7.7 compare the drop-outs and drop-ins of the two methods. We see that 236 of the drop-outs under the standard protocol were correctly declared as true alleles using the floating threshold method. However, sixteen of the allelic drop-outs from the floating threshold method did not drop-out using standard methods. More than half of these "new" drop-outs were located in locus D3 which tends to have a higher noise level compared to the other loci in this dataset. This may be due to primer residue increasing the background noise for the shorter loci in the electrophoresis.

Table 7.7: Comparisons of the drop-ins and drop-outs produced by the two methods.

		Dropped out		Dropped in	
Fixed 50 rfu threshold		Yes	No	Yes	No
Floating threshold	Yes	676	16	7	83
	No	236	1,224	8	-

7.4 Discussion

Previously Gilder et al. (2007) indicated that using observations from the negative controls from the same run as the samples could be used to extract information about the noise level. However, their approach did not take variation between the capillaries into account. From our analysis there are significant differences between the capillaries with negative controls within each run, and also significant differences between the same capillaries with negative controls for different runs. This suggest that the noise distribution is neither constant within runs nor for the same capillary for consecutive runs. Hence, our approach where we use the sample itself in order to determine the noise distribution is recommended, as it eliminates the between run and capillary variation. Furthermore, the stratification on loci for determining the threshold clearly improves the noise filtering as indicated by Figure 7.3.

The fixed 50 rfu-threshold yields in many cases the same number of drop-outs as the locus specific floating threshold. In Figure 7.7, the box-plots show the thresholds of the 107 mixture samples. For all loci, the median of the floating threshold is lower than the fixed 50 rfu limit. Note that within each dye band, the threshold median tends to decrease with the base pair length.

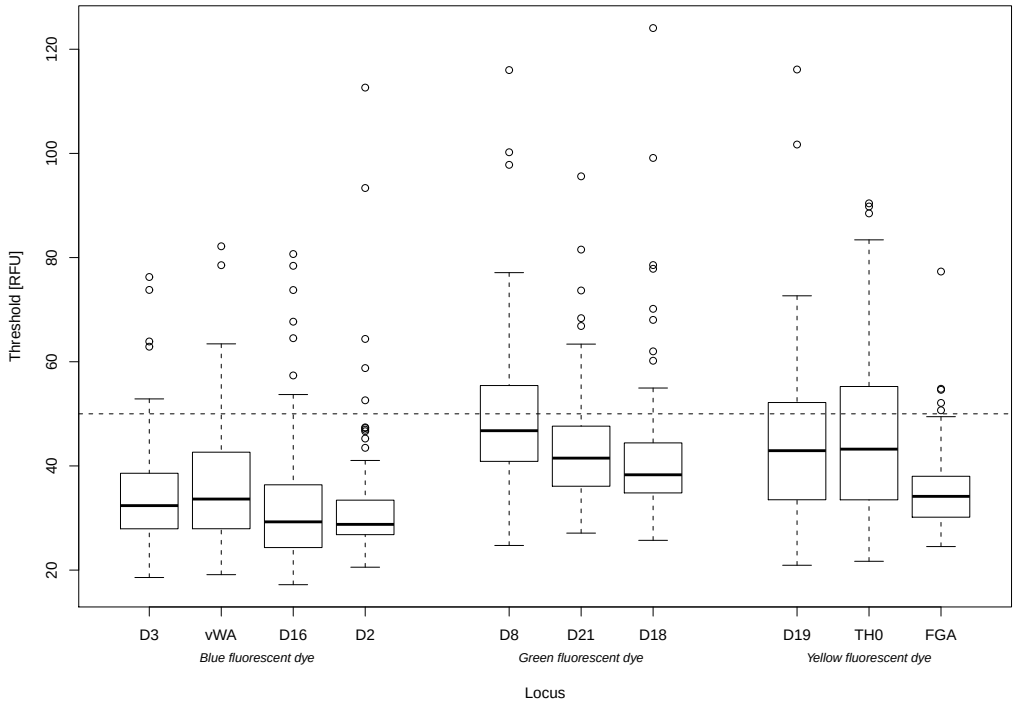

Figure 7.7: Box-plots of the estimated locus specific floating threshold for the 107 mixture cases.

An advantage of the locus specific threshold is that it enables the case worker to assess the noise level of the sample. Hence, for cases where a peak lies just below 50 rfu, the magnitude of the locus specific threshold indicates whether it is reasonable to include the peak in the signal or not.

Furthermore, in cases where the transformed peak heights, $\log_e$(peak height – 4.5), deviate substantially from normality, the data indicate that the sample may be subject to extensive noise or contamination of some kind. This may be used as sample quality diagnostic in order to determine if a re-analysis is necessary. This deviation may be observed from the QQ-plots and other usual diagnostics to validate assumptions of normality.

Since the stutter and pull-up corrections are based on a regression model, the parameters has been tuned for this specific data set. In general, the parameters must be determined for each laboratory, kit and DNA sequencer.

However, the trend in parameter magnitudes for the different pull-up directions is expected to be satisfied in general - possibly with an increase in the $\beta_{G\to B}$-parameter estimate. It is also worth emphasising the dependency on the kit used for DNA typing. The data used in our analyses were obtained using the SGM-Plus kit from Applied Biosystems. I.e., the parameters of stutter and pull-up filters are not directly applicable to other kits.

7.5 Conclusion

The methodology of regression and distributional analysis of the noise yielded satisfactory results in order to deduce a sample and investigation specific filter for STR DNA typing. Comparisons of the results with those based on the recommendations of the manufacturers indicated that the number of drop-outs for the two validation datasets decreased by 16% and 24%, respectively. Studies of different data sets supported this improvement and suggests that the methodology of the threshold determination is adequate for the noise filtering of STR quantitative data.

The filters for pull-up effects and stutters based on regression analysis trained on non-mixture data also showed applicability to mixed DNA samples. As mentioned in Section 7.4, the parameter estimates in the filter were tuned for this specific data set and the alleles of included profiles. Hence, the estimation of the parameters must be a part of a laboratory's internal quality assessment, where the consistency of the estimates over time are quality indicators.

Appendix

7.A Double stutters

In Gill et al. (2005), the authors argue that, once a stutter has been formed, its replication during subsequent PCR cycles perform as an ordinary allele. We investigate the behaviour of stutters, when we have two adjacent alleles of a heterozygous profile, i.e. what we called a double stutter.

Let h_i denote the expected value of pre-PCR peak height of allele i. Then for two adjacent alleles

n and n+1 from the same contributor, we have $h_n = h_{n+1} = h$ and $h_{n-1} = 0$ for the stutter position $n-1$. Let $\mathcal{P}$ denote the effect of one PCR-cycle, then after t PCR-cycles, the expected value of post-PCR peak heights $h_i^{(t)}$ is given by,

$$\left(h_{n+1}^{(t)}, h_n^{(t)}, h_{n-1}^{(t)}\right)^\top = \mathcal{P}^t(h, h, 0)^\top,$$

where $\boldsymbol{x}^\top$ denotes the transpose of the vector $\boldsymbol{x}$. $\mathcal{P}$ may be specified in terms of the PCR efficiency in one cycle, p, and the one-cycle stutter percentage, δ,

$$\begin{pmatrix} h_{n+1}^{(t)} \\ h_n^{(t)} \\ h_{n-1}^{(t)} \end{pmatrix} = \begin{bmatrix} 1+p & 0 & 0 \\ \delta & 1+p & 0 \\ 0 & \delta & 1+p \end{bmatrix}^t \begin{pmatrix} h \\ h \\ 0 \end{pmatrix} = \begin{bmatrix} (1+p)^t & 0 & 0 \\ t\delta(1+p)^{t-1} & (1+p)^t & 0 \\ \binom{t}{2}\delta^2(1+p)^{t-2} & t\delta(1+p)^{t-1} & (1+p)^t \end{bmatrix} \begin{pmatrix} h \\ h \\ 0 \end{pmatrix}$$

The second equality can be shown using some linear algebra. Define $\beta = t\delta/(1+p)$ to be the stutter percentage for the entire PCR process comprising t cycles. This definition ensures that the stutter percentage increases with the number of cycles as noted in the literature (Gill et al., 2000). The expression can then be rewritten as

$$\begin{pmatrix} h_{n+1}^{(t)} \\ h_n^{(t)} \\ h_{n-1}^{(t)} \end{pmatrix} \approx \begin{bmatrix} 1 & 0 & 0 \\ \beta & 1 & 0 \\ \frac{\beta^2}{2} & \beta & 1 \end{bmatrix} \begin{pmatrix} h_0 \\ h_0 \\ 0 \end{pmatrix} = \begin{pmatrix} h_0 \\ h_0(1+\beta) \\ h_0(\beta + \frac{\beta^2}{2}) \end{pmatrix} \tag{7.4}$$

where $h_0 = (1+p)^t h$ and the $\approx$ is due to $t(t-1)/2 \approx t^2/2$ from the binomial coefficient. The error induced from this approximation is negligible for $t \geq 28$ cycles.

The peak height, h_0, can be interpreted as the *actual* peak height after the PCR process. In Gill et al. (2005), the authors use $p = 0.8$ as the efficiency of a PCR cycle, hence indicating the theoretical doubling effect (requires that $p = 1$) from each cycle is not met in practice. Note, that there is a difference to the work of Gill et al. (2005) where they model the PCR process at the nucleic level. Our approach is in terms of quantitative measures of peak heights.

Often it is assumed that the peak at position n+1, $\hat{h}_{n+1}^{(t)}$, equals some "true" height, $\hat{h}$, after t PCR cycles. Due to stuttering, the peak at position n equals $\hat{h}$ plus an additional fraction, $\hat{\beta}$, from the n+1-position peak, $\hat{h}_n^{(t)} = (1+\hat{\beta})\hat{h}_{n+1}^{(t)} = (1+\hat{\beta})\hat{h}$. Furthermore, the peak height of stutter peak at position $n-1$ is $\hat{h}_{n-1}^{(t)} = \hat{\beta}\hat{h}_n^{(t)} = (\hat{\beta}+\hat{\beta}^2)\hat{h}$. This can be written using matrices as

$$\begin{pmatrix} \hat{h}_{n+1}^{(t)} \\ \hat{h}_n^{(t)} \\ \hat{h}_{n-1}^{(t)} \end{pmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \hat{\beta} & 1 & 0 \\ \hat{\beta}^2 & \hat{\beta} & 1 \end{bmatrix} \begin{pmatrix} \hat{h} \\ \hat{h} \\ 0 \end{pmatrix} \tag{7.5}$$

where the difference ($\beta^2/2 \neq \hat{\beta}^2$) between the matrices in (7.5) and (7.4) is induced by the delay of one cycle in the stutter product from the n+1-position peak to the stutter peak in position n. Hence, the relative contribution from the n+1-peak is smaller than modelled in (7.5) since when formed, the stutter peak is amplified as a regular peak (Gill et al., 2005), which is captured using (7.4).

When referring to the stutter percentage, β, we define it as the percentage of the parental peak that is transferred to the stutter peak, $\beta = h_{n-1}^{(t)}/h_n^{(t)}$. However, having two true alleles located at position n and $n+1$, we find

$$\frac{h_{n-1}^{(t)}}{h_n^{(t)}} = \frac{h_0\beta\left(1+\frac{\beta}{2}\right)}{h_0(1+\beta)} \neq \beta.$$

In this situation, the ratio of the stutter peak to the mean of the two parental peaks yields the stutter percentage,

$$\frac{h_{n-1}^{(t)}}{\frac{1}{2}\left(h_n^{(t)}+h_{n+1}^{(t)}\right)} = \frac{h_0\beta\left(1+\frac{\beta}{2}\right)}{\frac{1}{2}(h_0(1+\beta)+h_0)} = \frac{h_0\beta\left(1+\frac{\beta}{2}\right)}{h_0\left(1+\frac{\beta}{2}\right)} = \beta.$$

In situations where the heterozygous alleles are not adjacent (separated by more than 4 base pairs) or when stutter originates from a homozygous allele, we need for practical purposes only to consider the direct ratio $h_{n-1}^{(t)}/h_n^{(t)}$ in order to estimate β.

Bibliography

Applied Biosystems (2000). *GeneScan Reference Guide - Chemistry Reference for the ABI PRISM 310 Genetic Analyzer*. Applied Biosystems. Figure 'Virtual Filter Set F', pp. 4-10.

Applied Biosystems (2006). *AmpFℓSTR SGM Plus PCR Amplification Kit User's Manual*. Applied Biosystems.

Butler, J. M. (2005). *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* (2 ed.). Burlington, MA: Elsevier Academic Press Inc., U.S.

Cook, O. and L. Dixon (2006). The prevalence of mixed DNA profiles in fingernail samples taken from individuals in the general population. *Forensic Science International: Genetics 1*(1), 62–68.

Gilder, J. R., T. E. Doom, K. Inman, and D. E. Krane (2007). Run-Specific Limits of Detection and Quantitation for STR-based DNA Testing. *Journal of Forensic Science 52*(1), 97–101.

Gill, P. D., J. M. Curran, and K. Elliot (2005). A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. *Nucleic Acids Research 33*(2), 632–643.

Gill, P. D., J. Whitaker, C. Flaxman, N. Brown, and J. S. Buckleton (2000). An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International 112*(1), 17–40.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2010). Evaluating the weight of evidence using quantitative STR data in DNA mixtures. *Journal of the Royal Statistical Society. Series C, Applied statistics.* In Press.

7.6 Supplementary remarks

The three-person mixtures discussed in Section 5.10 were also analysed using the floating threshold methodology. As in Section 5.10 we discarded 17 out of the 120 samples due to preparation or run errors. For the remaining cases, the minimum amount of DNA contributed to a true allele was approximately 77.5 pg. Hence, the number of peak heights close to the limit of detection (50 rfu) is expected to be low, since experience show that with about 50 pg pre-PCR product the average peak heights are close to this limit.

When using the standard protocol with a fixed 50 rfu threshold, there was observed 8 drop-outs and 80 extra peaks not assigned to the contributors. These were distributed as 47 stutters, 10 pull-ups and 23 drop-ins. For the floating threshold there were 4 drop-outs and 85 extra peaks, which were categorised as 27 stutters, 2 pull-ups and 56 drop-ins. Hence, the performance of the two methods were almost identical with respect to the misclassification rates, which were 0.123% and 0.124%, respectively. This non-significant difference in assignment of alleles indicates that the fixed 50 rfu threshold is very reasonable for standard applications.

However, the methodology may be useful in situations were a the amount of DNA contributed by a suspect is limited. Given such circumstances the peak intensities associated with the suspect's profile may be close to the fixed limit of detection, e.g. with the majority of peak heights in the range 40 rfu to 60 rfu. Peaks below the limit of detection, 50 rfu say, would conventionally be declared as drop-outs. However, if the level of the noise supports a floating threshold limit of 30 rfu such considerations need not to be made, since no alleles would drop-out in this case. Often a case worker is able to visually detect peaks belonging to the suspect in the EPG below the limit of detection. However, lowering the limit of detection in order to include the suspect is clearly very erroneous and unfavourable to the defendant, since taken to the extreme, any DNA profile could be included in the crime related stain.

Furthermore, the method of adjusting for the contribution of stutter and pull-up effects is more accurate than just removing the peaks due to so-called masking. Keeping all relevant in the system is desirable since having a peak in stutter position that after adjustment has a peak height of 35 rfu, say, is more informative than having a NA observation due to removal of a potential stutter.

CHAPTER 8

Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out

Publication details

Co-authors: Poul Svante Eriksen[*], Helle Smidt Mogensen[†] and Niels Morling[†]

[*] *Department of Mathematical Sciences*
Aalborg University
[†] *Section of Forensic Genetics, Department of Forensic Medicine*
Faculty of Health Science, University of Copenhagen

Journal: Forensic Science International: Genetics (Under preparation)

Abstract:
DNA samples found at a scene of crime or obtained from the debris of a mass disaster accident are often subject to degradation. When using the STR DNA technology the DNA profile is observed via a so called electropherogram (EPG), where the alleles are identified as signal peaks above a signal to noise threshold. Degradation implies that these peak intensities decrease in strength for longer repeat sequences. Consequently, long STR loci possibly fail to produce peak heights above the limit of detection resulting in allelic drop-outs.

In this paper we present at method for measuring the degree of degradation of a sample and demonstrate how to incorporate this in estimating the probability of allelic drop-out. This is done by extending an existing method derived for non-degraded samples. The performance of the methodology is evaluated using data from degraded DNA where cases with varying amounts of DNA and levels of degradation are investigated.

Keywords:
Forensic genetics; STR DNA; Degraded DNA; Allelic dropout.

8.1 Introduction

This paper presents a statistical analysis of degraded STR DNA samples. The model derived in the subsequent sections is based on analysis of degraded DNA from body tissue kept under various non-optimal conditions. This implies that the DNA is affected by degradation, which is a commonly occurring event in crime cases, where evidence is collected after it has been exposed to e.g. sunlight, humidity and other degrading conditions (see e.g. Alaeddini et al., 2010). Furthermore, when identifying body remains in mass disaster cases, the samples are often found in the debris of the accident or in mass graves. Samples taken under these circumstances are often highly degraded and it is often hard to obtain full DNA profiles from longer STR loci (Schneider et al., 2004; Bender et al., 2004; Alonso et al., 2005; Dixon et al., 2006; Irwin et al., 2007; Prinz et al., 2007; Colotte et al., 2009).

In samples with degraded DNA, the signal intensities for the STR fragments decreases with the fragment length, due to the higher likelihood of the longer fragments to be degraded compared to the shorter fragments. Consequently, signals for the longest alleles are frequently missing, a phenomena called allelic drop-out. Allelic drop-out of the long alleles can also occur in samples with apparently moderate amount of DNA since the available quantification kits (Plexor, Applied Biosystems, and QHum, Qiagen) are based on amplicons less than 200 bp (Green et al., 2005), which is about half the length of the longest amplicons in e.g. the SGM Plus kit (Applied Biosystems).

In order to assign weight to the evidence in cases involving degraded samples, the case worker needs to be able to account for the fact that alleles or loci have dropped out. I.e. alleles of the true DNA profile fail to cause peak heights large enough to pass a limit of detection. Tvedebrink et al. (2009) presented a method for estimating the probability of allelic drop-out based on a logistic regression. However, the analysis was based on diluted samples from “healthy” DNA samples where degradation was absent. Here we show how to extend the drop-out model of

Tvedebrink et al. (2009) to handle degraded samples by adjusting the proxy for the amount of DNA to correct for degradation.

8.2 Materials and methods

8.2.1 Data

The data used in this study were investigated by The Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen. DNA profiles from 47 crime case samples were identified as degraded due to the decreasing signal intensities in the electropherogram (EPG) for longer fragments in the SGM Plus analysis. Of these were eight samples discarded due to obvious inhibition and for five samples the amount of DNA were limited such that the observed peak heights were in the range 40 to 15 rfu. The remaining 34 samples originated from saliva on a shirt (three samples), blood stains on paper (eight samples), blood sample from a decomposed body (five samples), a spleen from a decomposed body (one sample) and paraffin-embedded tissue (17 samples). The amounts of DNA varied from 17 pg to 1244 pg as quantified with Plexor (Applied Biosystems) and QHum (Qiagen) quantification kits. We used the methodology of Chapter 7 for filtering the raw signal, where the detection limit was set to 5 rfu in GeneScan. The Kazam macro in Genotyper was used for allele designation.

8.2.2 Model

It is well known to case workers investigating DNA from crime scenes that the DNA often is subject to degradation to some degree. The most common effect is an observable decrease in peak intensities for increasing base pair length of the amplicons. A probable explanation for this is that the longer the repeat sequence the higher the probability of a breakage in the primer binding sequence. Let p denote the probability that there is no breakage between two DNA bases (A, T, C or G). For simplicity we assume p to be constant with respect to length and the two adjacent DNA bases. That is, the probability of breakage between A and G is the same as T and C, and so on. Furthermore, by a constant probability of breakage as a function of base pair, bp, we do not assume any region of the genome to be more susceptible to breakage than others. Hence, this simple model does not include the possibility that proteins may protect some regions or segments of the DNA from degradation. Therefore, the longer the primer binding site the more possibilities exists for the occurrence of just one breakage. I.e. longer sequences increase the probability of damaged DNA:

$$\begin{aligned} P(\text{No degradation}) &= P(\text{No breakage between any base pair}) \\ &= P(\text{No breakage between a given base pair})^{\text{bp}} \\ &= p^{\text{bp}}, \end{aligned}$$

where we from the first to second line used that the probability of breakage is constant and that the probability of breakage between any two pairs is assumed independent of the constitution between all other pairs. Since $p \leq 1$ the function p^{bp} is a decreasing function of

bp. This implies that the longer amplicon (larger bp-values), the smaller is the probability of no degradation. Conversely this increases the probability of degradation as $P(\text{Degradation}) = 1 - P(\text{No degradation}) = 1 - p^{\text{bp}}$.

For healthy DNA samples the peak heights for the various loci are almost constant. This is due to the fact that there is no degradation acting in healthy samples with $p \approx 1$. This led Tvedebrink et al. (2009) to argue that the amount of DNA is well modelled using the average peak height H:

$$\text{Amount of DNA} \propto H = (n_{\text{het}} + 2n_{\text{hom}})^{-1} \sum_{i=1}^{n} h_i, \tag{8.1}$$

where $n = n_{\text{het}} + n_{\text{hom}}$ is the number of observed heterozygous and homozygous alleles in the profile. This was previously demonstrated to be a good proxy for the amount of DNA contributed to a stain (Tvedebrink et al., 2010).

However, in degraded samples one need to take the varying bp into account when modelling the mean peak height. Rather than being constant, the peak heights are affected by p and bp. By modelling the mean peak height as

$$H(\text{bp}) = cp^{\text{bp}}, \tag{8.2}$$

where c is some proportionality factor, depending e.g. on the amount of DNA in the sample, we obtain an expression for the mean peak heights in degraded samples. Note that for healthy samples $p \approx 1$ which implies that $c \approx H$ as defined in (8.1).

Hence, p may be taken as a measure of the level degradation of a given sample: the smaller p the more severe is the degradation, whereas values close to 1 indicate only moderate degradation. Taking logs on both sides of (8.2), we get:

$$\log H(\text{bp}) = \log(c) + \log(p)\text{bp} = \alpha_0 + \alpha_1 \text{bp}. \tag{8.3}$$

This implies a linear relationship between bp and $\log H(\text{bp})$. However, the assumption of linearity does not correct for the possibility of homozygocity. Hence, peak heights from homozygous loci needs to be divided by 2 in order for the model to be applicable to all loci. In Figure 8.1 we see that the model is supported by the data given in Table 8.1. Analysis of all the samples described in Section 8.2.1 indicated that linearity were satisfied for all samples (plots similar to Figure 8.1 are provided as supplementary on-line material). When applying this methodology to analyse degraded samples, it is important to verify the model fit by graphical diagnostics (as in Figure 8.1) and the R^2-statistic of the linear model in (8.3). This is due to the fact that a linear model may be fitted to any data set, but without reasonable validity the interpretation might be dubious.

This model formulation is in itself simple and intuitive. Furthermore, equation (8.3) enables direct implementation in the model of Tvedebrink et al. (2009) for estimating dropout probabilities $P(D)$, where D indicates a drop-out event. Since the probability of drop-out is primarily determined by the amount of DNA, it is natural to implement this into the model for drop-out probabilities. In Tvedebrink et al. (2009) the authors demonstrated that a logistic regression with H as explanatory variable yield an applicable model to estimate the drop-out probability for a given value of H:

$$\text{logit}\, P(D; H) = \log \frac{P(D; H)}{P(\overline{D}; H)} = \beta_{0,s} + \beta_1 \log \hat{H}, \tag{8.4}$$

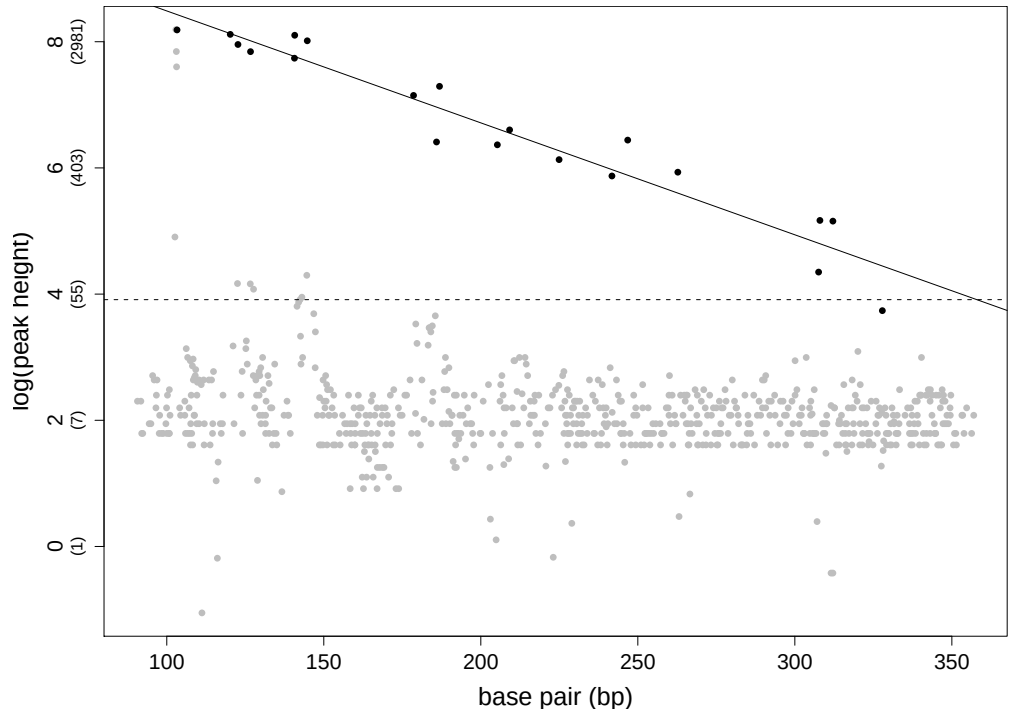

Figure 8.1: Peak heights on logarithmic scale plotted against base pair (bp). The black points are assigned true peaks by the floating threshold methodology (Chapter 7) and the grey points are assigned noise (negative signal). The numbers in brackets on the ordinate are the rfu values. The dashed horizontal line shows the fixed 50 rfu detection threshold.

where $\hat{H} = H$ and $\hat{H} = 2H$ for heterozygous and homozygous loci, respectively. The subscript s in $\beta_{0,s}$ indicate that this parameter is locus specific whereas β_1 is not (Tvedebrink et al., 2009). Since the model in (8.3) measures the degree of degradation, we may adjust the estimate of H by H(bp), such that the model for dropout also is applicable to degraded DNA samples with H(bp) as explanatory variable.

8.2.3 Implementation of degradation in drop-out probability estimation

The definition of H in Tvedebrink et al. (2009) assumes that H is estimated based on all peak height observations (see (8.1)). However, since the peak height decreases for increasing bp in a degraded DNA sample, the assumptions for the drop-out model it not satisfied. To compensate for the decrease in peak heights and thereby increase in drop-out probability, we incorporate the level of degradation in the drop-out model. Let G_S be the profile of a given individual, e.g. the suspect of a crime case. First, the peaks related to the alleles originating from G_S is determined, and the α-parameters of the degradation model, (8.3), are estimated based on a linear regression

Table 8.1: Data used in Figures 8.1 and 8.2. The drop-out probability of allele 24 in locus D2 (shaded row) is assessed in the example of Section 8.3. The column 'Corrected height' is computed using the method discussed in Section 8.6.

Dye	Locus	bp	Allele	Height	log(Height)	Corrected height
Blue	D3	120.22	14	3349.00	8.12	4790.12
Blue	D3	140.66	19	2295.52	7.74	4714.49
Blue	vWA	178.58	17	1272.69	7.15	5114.07
Blue	vWA	186.85	19	1470.00	7.29	6838.08
Blue	D16	246.78	9	627.00	6.44	8424.95
Blue	D16	262.74	13	377.00	5.93	6719.33
Blue	D2	307.57	19	77.40	4.35	3050.41
Blue	D2	327.87	24	42.00	3.74	2370.77
Green	AME	103.27	1	7188.76	8.88	7617.12
Green	D8	140.74	12	3303.01	8.10	6793.26
Green	D8	144.74	13	3026.21	8.02	6680.60
Green	D21	205.27	29	581.91	6.37	3750.30
Green	D21	209.17	30	737.10	6.60	5090.00
Green	D18	307.99	18	175.50	5.17	6967.80
Green	D18	312.12	19	173.54	5.16	7412.75
Yellow	D19	122.67	14	2853.63	7.96	4262.48
Yellow	D19	126.67	15	2546.79	7.84	4083.25
Yellow	TH0	185.90	9.3	1217.00	7.10	5566.79
Yellow	FGA	224.93	20	460.00	6.13	4198.52
Yellow	FGA	241.76	24	355.00	5.87	4364.55

of log peak height on bp. The G_S-specific regression $(\hat{\alpha}_0, \hat{\alpha}_1)$-parameters are inserted in (8.3) together with the bp-value for the allele under investigation for drop-out. For the ith allele of G_S the adjusted H-estimate is: $H(\text{bp}_i) = \exp(\hat{\alpha}_0 + \hat{\alpha}_1 \text{bp}_i)$ which we then insert in (8.4):

$$\text{logit}\, P[D_i; \hat{H}(\text{bp}_i)] = \beta_{0,s} + \beta_1 \log \hat{H}(\text{bp}_i) \tag{8.5}$$

where $\hat{H}(\text{bp}_i) = H(\text{bp}_i)$ or $\hat{H}(\text{bp}_i) = 2H(\text{bp}_i)$ depending on whether D_i represents a drop-out on a heterozygous or homozygous loci, respectively. The information about bp_i and homozygous/heterozygous locus is given by the specified profile. Hence, as for the drop-out model of Tvedebrink et al. (2009) the drop-out probabilities are determined for a specific profile since the drop-out probability depends on the observed peak heights associated to that particular profile.

8.3 Results

In Figure 8.2, the observed peak heights of Table 8.1 are plotted against their base pair lengths (the peak heights of homozygous loci are divided by 2 in Figure 8.2). The superimposed curves represents the adjusted, $H(\text{bp})$, observed mean peak heights, H, and the fixed 50 rfu detection

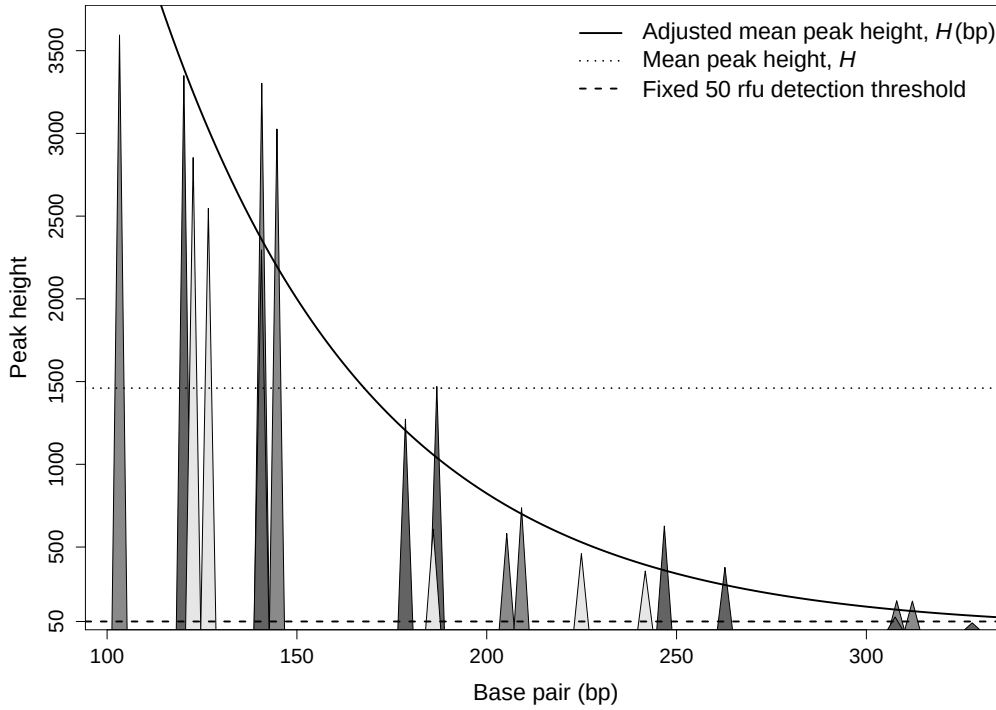

Figure 8.2: Peak heights plotted against base pair length. The solid curve show the adjusted mean peak height, the dotted line the observed mean peak height and the dashed line the fixed 50 rfu detection threshold.

threshold, respectively. Note that the profile of Table 8.1 is homozygous for Amelogenin (female) and TH0. This implies that the observed peak heights for these two loci are divided by 2 before the linear model (8.3) is fitted to the data.

In Table 8.1, the row with a grey shading show the peak below the fixed 50 rfu detection threshold which is represented by the dashed line in Figures 8.1 and 8.2. However, using the methodology of Chapter 7, it is possible to have locus specific thresholds enabling detection of all the alleles in Table 8.1. In the following we assume that the identified alleles in Table 8.1 represents a true DNA profile, e.g. identified from a blood stain left by the suspect found at the scene of crime.

The drop-out model and fitted parameters in Tvedebrink et al. (2009) are calibrated for the event $D = \{\text{peak height} < 50 \text{ rfu}\}$ for non-degraded DNA. Hence, the H-estimate must only be based on the peak heights above 50 rfu for the drop-out model to be applicable. Under the assumption that the alleles in Table 8.1 constitute the profile, i.e. the suspect profile is heterozygous for all loci but Amelogenin and TH0, we find that $H = 1460.41$ rfu. Most of the observed peak heights deviates substantially from H due to degradation (see dotted line in Figure 8.2).

For evidential computations we need the probability that allele 24 in locus D2 has dropped out.

The methodology of Tvedebrink et al. (2009) makes this computation straight forward using the estimated H-value. By plugging-in the estimated H in (8.4) and taking the inverse of the logit-function, $\text{logit}^{-1}(x) = \exp(x)/[1+\exp(x)]$, we obtain the drop-out probability $P(D_{\text{D2}_{24}}; \hat{H} = 1460.41) = 1.54 \times 10^{-6}$, where we used $\beta_{0,\text{D2}} = 18.31$ and $\beta_0 = -4.35$ from Table 2 of Tvedebrink et al. (2009). This is an extremely low drop-out probability when considering the fact that allele 19 in the same locus has a peak height of 77 rfu.

From graphical inspections of the (simplified) EPG in Figure 8.2 it is obvious that the DNA sample is subject to degradation. In order to take the degradation of the DNA into account we adjust the estimated H. The solid line in Figure 8.1 has $(\hat{\alpha}_0, \hat{\alpha}_1) = (10.262, -0.0177)$ with $R^2 = 0.931$ which together with Figures 8.1 and 8.2 and other graphical diagnostics indicate a good agreement with the model. Since the fragment length, bp, of allele 24 in locus D2 is $\text{bp}_{\text{D2}_{24}} = 327.87$ (see Table 8.1), the adjusted H-value yields $H(\text{bp}) = \exp(10.262 - 0.0177 \times 327.87) = 85.25$ rfu. This estimated peak height is reasonably close to the observed peak height (77 rfu) for the other allele in the same locus (Table 8.1). The estimated peak height is plugged into (8.5) which implies that $P[D_{\text{D2}_{24}}; \hat{H}(\text{bp}) = 85.25] = 0.26$. This drop-out probability is more reasonable than $P(D; \hat{H})$ not taking degradation into account.

Note from (8.3) we may compute p from the estimate of α_1, $\hat{p} = \exp(\hat{\alpha}_1) = \exp(-0.0177) = 0.982$. From experience (see the supplementary material) this sample is moderately degraded.

8.4 Discussion

Since most DNA samples are analysed in replicates (or at least in duplicates), an additional source of information is the consistency of the estimated degradation parameter across replicates. For replicates the amount of DNA may vary, however, this affects (in principle) only α_0, whereas α_1 should remain constant. For most of the samples analysed in this paper the were no significant difference between the levels of degradation $\hat{p}_{R_i}$ and $\hat{p}_{R_j}$ for different replicates R_i and R_j, $i \neq j$. Similarly, for samples originating from the same body tissue or fluid, the degradation pattern should be reasonably similar across samples taken from the same source of the crime scene. This were supported by the data, however, some cases had significant differences between tissue/fluid samples.

The likelihood ratio is defined as $LR = P(E|H_p)/P(E|H_d)$, where H_p and H_d are two competing hypotheses that could represents the statements of the prosecutor and defence. Let G_S be the DNA profile of the suspect, which in the example of Section 8.3 equals the profile in Table 8.1. In order to evaluate $P(E|H_p)$ where H_p claims that G_S is the donor of the observed stain, an allelic drop-out need to have occurred in order to explain the missing 24 allele in locus D2. Hence, the probability $P(D_{\text{D2}_{24}})$ enters in the numerator of LR. Thus, the smaller this probability the smaller the LR. Hence, the prosecutor will claim that degradation is present since the probability of allelic drop-out is approximately 10^5 larger when assuming degradation, compared to the non-degraded probability of allelic drop-out.

$P(E|H_d)$ is evaluated by summation over the set of possible unknown profiles with or without allelic drop-out. Whether or not it is favourable for the defence to consider unknown profiles

with drop-outs depend on the allele probabilities for the homozygous loci. That is, if

$$P(\bar{D}; 2H)P(A_iA_i) < P(D; H)P(\bar{D}; H)\sum_{j\neq i}^{k} P(A_iA_j)$$

then $P(E|H_d)$ is increased by allowing for drop-out which results in a decreased LR. This consideration applies whether or not the sample is degraded. However, the drop-out probabilities will only increase when considering degradation since $H(\text{bp}) = cp^{\text{bp}} \leq H$ and $P(D; H)$ increases as H decreases. On the other hand, the probability of alleles not dropping out is possibly larger when correcting for possible degradation, $P(\bar{D}; H(\text{bp})) < P(\bar{D}; H)$, since $H(\text{bp})$ may be larger than H for short amplicons.

8.5 Conclusion

We presented a method for the decay in the peak intensities of forensic STR loci as a function of increasing base pairs, bp. The model showed satisfactory agreement to data and is simple and intuitive. Furthermore, we demonstrated how to implement the information of degradation in the computation of the probability of allelic drop-out in the situation of degraded samples.

Bibliography

Alaeddini, R., S. J. Walsh, and A. Abbas (2010). Forensic implications of genetic analyses from degraded DNA - A review. *Forensic Science International: Genetics 4*(3), 148–157.

Alonso, A. et al. (2005). Challenges of DNA profiling in mass disaster investigations. *Croatian Medical Journal 46*(4), 540–548.

Bender, K., M. J. Farfan, and P. M. Schneider (2004). Preparation of degraded human DNA under controlled conditions. *Forensic Science International 139*(2-3), 135–140.

Bill, M. et al. (2005). PENDULUM - a guideline-based approach to the interpretation of STR mixtures. *Forensic Science International 148*, 181–189.

Colotte, M., V. Couallier, S. Tuffet, and J. Bonnet (2009). Simultaneous assessment of average fragment size and amount in minute samples of degraded DNA. *Analytical Biochemistry 388*(2), 345–347.

Dixon, L. A. et al. (2006). Analysis of artificially degraded DNA using STRs and SNPs - results of a collaborative European (EDNAP) exercise. *Forensic Science International 164*(1), 33–44.

Green, R., I. Roinestad, C. Boland, and L. Hennessy (2005). Developmental Validation of the Quantifiler™ Real-Time PCR kits for the Quantification of Human Nuclear DNA samples. *Journal of Forensic Science 50*(4), 809–825.

Irwin, J. A. et al. (2007). Application of low copy number STR typing to the identification of aged, degraded skeletal remains. *Journal of Forensic Sciences 52*(6), 1322–1327.

Prinz, M. et al. (2007). DNA Commision of the International Society for Forensic Genetics (ISFG): Recommendations regarding the role of forensic genetics for disaster victim identification (DVI). *Forensic Science International: Genetics 1*(1), 3–12.

Schneider, P. M. et al. (2004). STR analysis of artificially degraded DNA - results of a collaborative European exercise. *Forensic Science International 139*(2-3), 123–134.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2009). Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics 3*(4), 222–226.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2010). Evaluating the weight of evidence using quantitative STR data in DNA mixtures. *Journal of the Royal Statistical Society. Series C, Applied statistics*. In Press.

8.6 Supplementary remarks

Degradation affects the mean of the peak heights and areas. Since degradation is a very common situation in forensic case work, the models developed should be able to handle degradation. As with the extension of the mixture separation method to allowing for allelic drop-out, the method is extensible to correct for degradation.

Assume that the biological material contributed by the donors is of similar type, e.g. blood, tissue, body fluids, etc., and that the material has been exposed to similar conditions over an approximate identical time span. Based on these assumptions it is reasonable to assume that the level of degradation is common for the DNA and that the peak intensities may be modelled by $c_k p^{\text{bp}}$, where c_k reflects the amount of DNA contributed by the kth individual and p is common for all $k = 1, \ldots, m$.

In cases of degradation, the effect on a four peak locus might be such that the highest and lowest peak heights relate to the major component and the two alleles with intermediate peak heights belong to the minor contributor of a two-person DNA mixture. This could happen if the highest and lowest peaks are in each end of the ladder interval and the intermediate in between. Figure 8.3 shows examples of this situation for the base pair interval from 125 bp to 280 bp for the SGM Plus kit.

The plot in Figure 8.3 exemplifies a two-person DNA mixture with $p = 0.98$ and amounts of DNA corresponding approximately to an 1:2 mixture. That is, the expected peak heights of heterozygous loci are given by $h_{s,i}^{(k)} = c_k p^{\text{bp}_{s,i}^{(k)}}$, which implies that $\sum_{i=1}^{4} \log h_{s,i} = 2\alpha_0^{(+)} + \alpha_1 \text{bp}_{s,+}$, where $\alpha_0^{(+)} = \alpha_0^{(1)} + \alpha_0^{(2)}$, $\alpha_0^{(i)} = \log c_i$ and $\alpha_1 = \log p$.

Hence, a regression of $\sum_{i=1}^{4} \log h_{s,i}$ on $\text{bp}_{s,+}$ would give estimates of $(\alpha_0^{(+)}, \alpha_1)$. However, additivity of peak heights on natural-scale does not transfer to additivity on log-scale. Homozygous allele peak heights are $\log h_{s,1}^{(k)} = \log 2 + \alpha_0^{(k)} + \alpha_1 \text{bp}_{s,1}^{(k)}$ and shared alleles has $\log h_{s,i'} = \log(c_1 + c_2) + \alpha_1 \text{bp}_{s,i'}^{(\cdot)}$, where in particular the shared alleles implies that the regression of $\sum_{i=1}^{n_s} \log h_{s,i}$ on $\text{bp}_{s,+}$ would yield locus dependent intercept.

Thus, different means for estimating p for DNA mixtures need to be considered. By simulating a large number (e.g. 1,000) DNA mixtures with known profiles, amounts of DNA and level of degradation, p_0, it was possible numerically to compare the performance of different estimators of p. Of the investigated methods a regression of log to the mean of peak heights, $\log \bar{h}$, on the mean of base pairs, $\bar{\text{bp}}_s$, yielded a good approximation based on simulations with a 95%-confidence interval of $(-8.58 \times 10^{-5}, 4.62 \times 10^{-5})$ for the difference between p_0 and $\hat{p}$.

The relevant observation window for the loci included in the SGM Plus kit (AB) starts around 100 bp. Using this off-set the observed peak intensities may be adjusted for degradation by compensating by the fitted decay. Given $\hat{\boldsymbol{\alpha}}$ and the peak height h it is possible to compute the degradation corrected peak height $\tilde{h}$. By multiplying the observed peak heights by $\exp[-\alpha_0(\text{bp} - 100)]$ the effect of degradation is inverted resulting in less imbalances between loci, $\tilde{h}_{s,i} = h_{s,i} \exp[-\hat{\alpha}_1(\text{bp}_{s,i} - 100)]$.

In the example of Section 8.3, $\hat{\alpha}_1 = -0.0177$ and by using the approach above we get the peak

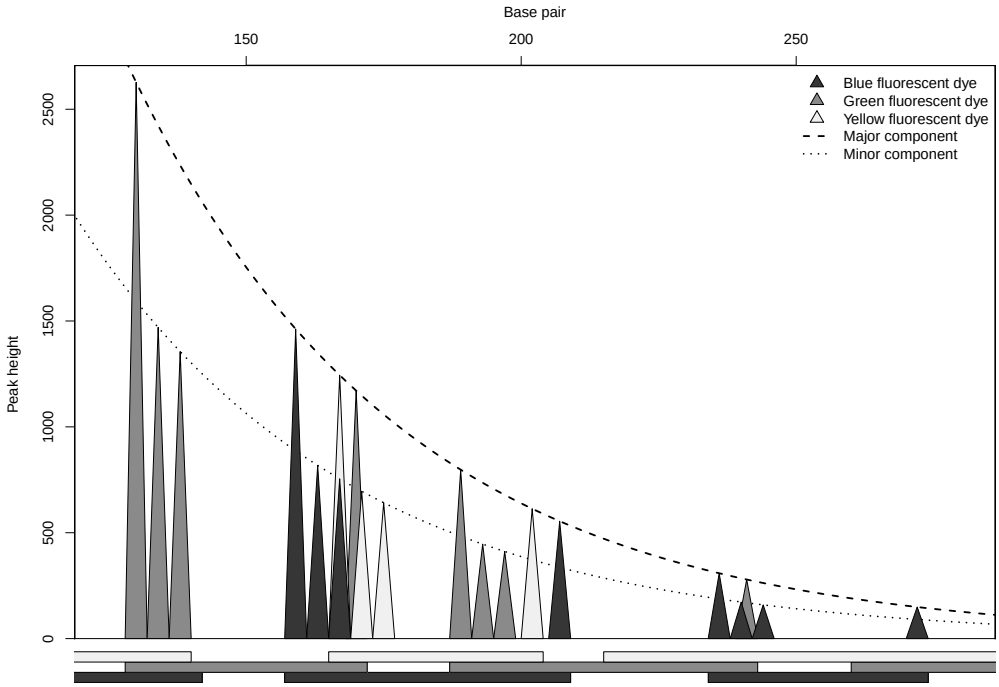

Figure 8.3: Degradation of a two-person DNA mixture. The highest and lowest peaks belong to the major component. The shaded areas below the first axis show the range of the allelic ladder for the various STR loci in the 125-280 bp window of the SGM-Plus kit (Applied Biosystems, AB).

heights reported in the 'Corrected height'-column of Table 8.1. There is still evidence of peak height imbalances within loci. However, the heterozygote balance, *Hb*, which is the ratio of the heterozygous peak heights (see e.g. Bill et al., 2005), is improved by the correction, where the range of *Hb* for the observed peak heights is (0.54, 0.99) it is (0.74, 0.98) after the peak height correction.

In order for the models for DNA mixtures of Chapters 4 and 5 to be valid, the proportionalities of peak heights and peak areas need to be preserved. However, the correction of peak heights is also applicable to peak areas, hence $\tilde{a}_{s,i} = a_{s,i} \exp[-\hat{\alpha}_1(\text{bp}_{s,i} - 100)]$ which ensures the same proportionality as before the correction. Therefore, no changes are needed in the sets $\mathcal{J}_i'$ for the mixture separator when the peak intensities are adjusted for degradation.

In the next chapter the model for degraded DNA is combined with the models from the previous chapters in a 'unifying likelihood ratio'. That is, a likelihood ratio were all the discussed complications can be included and accounted for when assessing the weight of the DNA evidence in crime cases.

CHAPTER 9

Epilogue

9.1 Conclusion

In the preceding seven chapters (Chapters 2-8) the core content of this present PhD thesis has been presented. The main focus of the PhD project has been to develop statistical models applicable to the quantitative part of the STR analysis and in particular DNA mixtures. However, since the genetic part (qualitative allelic data) of the evidence constitutes the fundamental inputs in evidential weight calculations, it was difficult not to treat this topic. This lead to the interest for IBD and the effect of population structures when computing the evidential weight. As pointed out by one of the reviewers of the paper in Chapter 2 ('Overdispersion in allelic counts and θ-correction in forensic genetics' to appear in Theoretical Population Biology) does the forensic databases not constitute the databases of interest. More general population surveys should be used when making inference about θ, e.g. random samples taken from well-defined subpopulations on a high resolution. For the Danish population this could be samples taken from small villages or islands since these subpopulations may cause large allelic divergence and thus yield θ-estimates in the higher end of the plausible range (Balding, 2005). From Figure 2.1 we saw that this in practise would lead to conservative evaluation of the evidence. Furthermore, this is equivalent to the fact that the probability of a "random match" (a profile match of two unrelated individuals) increases with θ.

For the quantitative part, the work was initiated by assuming no complications of stutters, pull-up effects, allelic drop-out or degradation. Under these settings it was possible to derive two models for DNA mixtures, where the simplest of the two were wrapped into a greedy algorithm which

efficiently separated DNA mixtures. For the particular data used in the paper the algorithm was at least as successful as three experienced case workers. However, the analysis did also emphasise that the results should be interpreted with caution. This was especially important for samples close to 1:1-mixture proportion and when the interest was about the minor profile. The analysis of three-person mixtures repeated this picture where the success rate for the 1:2:4-mixture proportion was rather low for the mid and minor profiles.

Having done this, a natural extension of the models was to handle allelic drop-out and degraded DNA samples as these phenomena are frequently occurring in real crime case work. From the remarks of Sections 6.5.1 and 8.6 it was demonstrated how the presented models may be combined in order to handle these complications. In the remarks it was only exemplified how to modify the statistical model and mixture separating algorithm for two-person DNA mixtures. However, the cases of more contributors follow along the same lines. The work with allelic drop-out also made it evident that there were possibilities for refinement of the determination of the signal-to-noise ratio. The use of a fixed threshold may in some cases discard important information regarding the distribution of the noise component from the measurement technique. The model proposed to determine this threshold was based on a simple analysis of quartiles in order to estimate the parameters of the log-normal distribution. However, for some situations this approach seemed to be too simple as a sudden increase of the noise level was detected for a short bp-interval. This temporally increase in the background noise caused non-linearity in the QQ-plots and in some cases increased the variance estimate substantially. Loess-curves were investigated to handles this non-linearity. However, they did not improve the overall performance significantly. Further work may suggest ways to adjust for this fact, but one has to focus the attention on newer typing kits, as these should have better signal-to-noise ratios than the SGM-Plus kit (Applied Biosystems).

9.2 Weight of evidence calculations

In the preceding chapters it has been demonstrated how to incorporate the quantitative part of the STR typing results in the likelihood ratio approach. The principle was to assign a weight to each quantitative term of the *LR*, where the weight should reflect the compliance between the expected and observed peak intensities. Terms with minor disagreement (e.g. due to measurement errors) should receive a large weight whereas profile combinations leading to substantial differences would be weighted by a quantity close to zero.

This extendability of the *LR* is one of the many arguments for using this approach rather than the "Random man not excluded"-approach (often abbreviated RMNE-approach in the literature). I will not discuss the philosophical differences or many advantages of *LR* over RMNE, since these are irrelevant at this point. However, it should be noted that the models discussed above is of no use when assessing the weight of evidence through RMNE. In line with many others (e.g. Evett and Weir, 1998; Balding, 2005; Buckleton et al., 2005; Gill et al., 2006; Buckleton and Curran, 2008) I strongly recommend the *LR*-approach in evidential calculations carried out in forensic genetics.

The *LR* is formed by evaluating the evidence (crime scene evidence and identified profiles) under

competing hypothesis, often denoted H_p and H_d for the prosecutor and defence hypotheses. Since H_p and H_d are only mutually exclusive, and not exhaustive, one needs to recall that there are several *LR*s - one for every (H'_p, H'_d)-pair of hypotheses. Hence, the fact that the *LR* favours H_p over H_d does not imply that there cannot exist a H'_d for which $LR' = P(E|H_p)/P(E|H'_d) < 1$ (Balding, 2005).

The extensions of the *LR* derived in Chapters 4 and 5 only considered cases assuming no allelic drop-outs. However, as previously argued does this assumption often fail together with the 'no degradation'-assumption. Hence, for proper inclusion of the available data and applicability to most types of crime cases, the *LR* needs to be extended further. Let $\mathcal{Q} = (\mathcal{Q}_{\text{mis}}, \mathcal{Q}_{\text{obs}})$ and $\mathcal{G} = (\mathcal{G}_{\text{mis}}, \mathcal{G}_{\text{obs}})$, where the subscripts refer to dropped-out and observed alleles. That is, $\mathcal{Q}_{\text{obs}}$ are the observed peak intensities, whereas $\mathcal{Q}_{\text{mis}}$ denotes the event of an allelic drop-out, i.e. the peak failed to be detected. Similarly are $\mathcal{G}_{\text{obs}}$ and $\mathcal{G}_{\text{mis}}$ the associated types of alleles, where the need for $\mathcal{Q}_{\text{mis}}$ and $\mathcal{G}_{\text{mis}}$ is induced by the hypothesis under consideration.

Given a specific hypothesis the set of plausible profile combinations $\mathcal{C}$ is induced. That is, if the prosecutors hypothesis H_d claims that the observed crime scene stain originates from a victim, V, and suspect S then $\mathcal{C}_p = \{(G_V, G_S)\}$, where respectively G_V and G_S are the profiles of V and S. In connection to this hypothesis the defence states that "The observed crime scene stain originates from the victim and an unknown profile" then $\mathcal{C}_d = \{G_U : (G_V, G_U) \equiv H_d\}$, with G_U being the profile of the unknown contributor U. This definition of $\mathcal{C}_d$ does not limit G_U to be consistent with $(\mathcal{Q}_{\text{obs}}, \mathcal{G}_{\text{obs}})$, hence drop-out of U's alleles is allowed with this formulation. Note that this definition of $\mathcal{C}$ is different from that of Sections 4.3 and 5.5 where the plausible profiles in $\mathcal{C}$ needed to be consistent with the observed alleles, i.e. no allelic drop-outs were allowed.

Additionally, drop-ins, stutters and pull-up peaks possibly causes more alleles to be observed than those of the true contributors. However, as claimed in Section 5.9 are stutters (and pull-up peaks) profile independent. Hence, given the peak intensity information in allele position n it is (in principle) possible to predict and adjust for the stutter contribution to the peak in position $n-1$. Similarly, the pull-up contribution can be removed from peaks with overlapping bp-values. However, not all such peaks were successfully removed as 6 stutters and 4 pull-ups were observed above the signal-to-noise threshold (Table 7.4) for the two-person mixtures, and for the three-person mixtures 27 stutters and 2 pull-ups were detected. Hence, peaks other than those belonging to the true donors must be incorporated in the 'unifying' model to be consistent with the observed data.

9.3 Unifying likelihood ratio

The evaluation of the *LR* consists of computing the probability of the evidence under the two hypothesis and form their ratio. Since the $\mathcal{C}$-sets are discrete the probability $P(\mathcal{E}|H)$ may be evaluated using the law of total probability $P(\mathcal{E}|H) = \sum_{\boldsymbol{G}\in\mathcal{C}} P(\mathcal{E}|\boldsymbol{G})P(\boldsymbol{G})$, where $\boldsymbol{G}$ is short for the profiles involved, e.g. $\boldsymbol{G} = (G_V, G_S)$ under the prosecutors hypothesis in the example above. In order to discuss the evidential weight there need to be at least one identified DNA profile, namely the suspect's profile G_S. For general purposes let $\boldsymbol{K}$ be the known DNA profiles associated with the case, e.g. $\boldsymbol{K} = (G_V, G_S)$ above. Then the evidence $\mathcal{E}$ consists of $\mathcal{E}_c$ and $\boldsymbol{K}$,

where $\mathcal{E}_c$ were the crime scene evidence including both the quantitative and qualitative parts, $\mathcal{E}_c = (\mathcal{Q}, \mathcal{G})$. First we note that the crime scene evidence, $\mathbb{E}$, and the known profiles, $\boldsymbol{K}$ are assumed conditionally independent given $(\mathcal{G}, \boldsymbol{G})$. That is, given $\boldsymbol{G} \in \mathcal{C}$ and $\mathcal{G}$ the known profiles $\boldsymbol{K}$ has no influence on the crime scene stain. Hence, using the definition of conditional probability we can for $\boldsymbol{G} \in \mathcal{C}$ factorise $P(\mathcal{E}|\boldsymbol{G})$ as:

$$P(\mathcal{E}|\boldsymbol{G}) = P(\mathcal{E}_c, \boldsymbol{K}|\boldsymbol{G}) = P(\mathcal{Q}, \mathcal{G}, \boldsymbol{K}|\boldsymbol{G}) = P(\mathcal{Q}|\mathcal{G}, \boldsymbol{G})P(\mathcal{G}|\boldsymbol{K}, \boldsymbol{G})P(\boldsymbol{K}|\boldsymbol{G}). \tag{9.1}$$

In (9.1) the $P(\mathcal{Q}|\mathcal{G}, \boldsymbol{G})$-term measures the agreement of the observed and expected peak intensities under some model. If the detected alleles in $\mathcal{G}$ equals those of $\boldsymbol{G}$ neither drop-out nor drop-in (including stutters and pull-ups) have caused missing or additional alleles to be present in the signal. Hence, $P(\mathcal{Q}|\mathcal{G}, \boldsymbol{G}) = P(\mathcal{Q}|\boldsymbol{G})$ may be evaluated by the one of models as presented in Chapters 4 or 5, and since $\mathcal{G} = (\mathcal{G} \cap \boldsymbol{G}) = \boldsymbol{G}$, i.e. the profiles are consistent, $P(\mathcal{G}|\boldsymbol{G}) = 1$.

However, in cases with stutters, pull-ups or drop-ins present $\mathcal{Q}$ is split into two parts ascribed respectively to $\boldsymbol{G} = \mathcal{G} \cap \boldsymbol{G}$ and $\bar{\boldsymbol{G}} = \mathcal{G} \setminus (\mathcal{G} \cap \boldsymbol{G})$. The evaluation is done by $P(\mathcal{Q}|\mathcal{G}, \boldsymbol{G}) = P(\mathcal{Q}_{\bar{\boldsymbol{G}}}|\mathcal{Q}_{\boldsymbol{G}}, \mathcal{G}, \boldsymbol{G})P(\mathcal{Q}_{\boldsymbol{G}}|\boldsymbol{G})$, where in cases of possible stuttering $P(\mathcal{Q}_{\bar{\boldsymbol{G}}}|\mathcal{Q}_{\boldsymbol{G}}, \mathcal{G}, \boldsymbol{G})$ assigns probability to this event. In this thesis such models have not been discussed, however, a logistic regression (similar to that of the drop-out model) may be derived, where the explanatory variable for stutters and pull-ups would be the parental peak's intensities. For drop-ins (additional peaks not possible to categorise as stutters or pull-ups), the noise level of the sample might be an appropriate covariate.

Furthermore, if $\boldsymbol{G}$ implies allelic drop-out $\mathcal{Q}$ can be decomposed into $(\mathcal{Q}_{\text{mis}}, \mathcal{Q}_{\text{obs}})$ and the quantitative term then factorises further $P(\mathcal{Q}|\mathcal{G}, \boldsymbol{G}) = P(\mathcal{Q}_{\text{mis}}|\mathcal{Q}_{\text{obs}}, \mathcal{G}, \boldsymbol{G})P(\mathcal{Q}_{\text{obs}}|\mathcal{G}, \boldsymbol{G})$. The probability of an allelic drop-out, $P(\mathcal{Q}_{\text{mis}}|\mathcal{Q}_{\text{obs}}, \mathcal{G}, \boldsymbol{G})$, is computed given the observations and information about the sample's genotypes. An allelic drop-out is equivalent to the event that the peak height is less than the limit of detection. Hence, $P(\mathcal{Q}_{\text{mis}}|\cdot)$ could be evaluated by $\int_0^T P(h|\cdot)\,\mathrm{d}h$, where T and h are the limit of detection and peak height, respectively. However, the drop-out model of Chapter 6 is an approximation to this integral and since it is easier to compute we use $P(D; H)$ to quantify $P(\mathcal{Q}_{\text{mis}}|\mathcal{Q}_{\text{obs}}, \mathcal{G}, \boldsymbol{G})$.

Thus combining (9.1) with the extension for drop-outs and additional alleles compared to $\boldsymbol{G}$ the 'unifying likelihood ratio' can be defined as:

$$\begin{aligned} LR &= \frac{P(\mathcal{E}|H_p)}{P(\mathcal{E}|H_d)} \\ &= \frac{\sum_{\boldsymbol{G} \in \mathcal{C}_p} P(\mathcal{Q}_{\text{mis}}|\mathcal{Q}_{\text{obs}}, \boldsymbol{G})P(\mathcal{Q}_{\text{obs},\bar{\boldsymbol{G}}}|\mathcal{Q}_{\text{obs},\boldsymbol{G}}, \mathcal{G}, \boldsymbol{G})P(\mathcal{Q}_{\text{obs},\boldsymbol{G}}|\boldsymbol{G})P(\mathcal{G}|\boldsymbol{K}, \boldsymbol{G})P(\boldsymbol{K}|\boldsymbol{G})P(\boldsymbol{G})}{\sum_{\boldsymbol{G}' \in \mathcal{C}_d} P(\mathcal{Q}_{\text{mis}}|\mathcal{Q}_{\text{obs}}, \boldsymbol{G}')P(\mathcal{Q}_{\text{obs},\bar{\boldsymbol{G}}}|\mathcal{Q}_{\text{obs},\boldsymbol{G}}, \mathcal{G}, \boldsymbol{G}')P(\mathcal{Q}_{\text{obs},\boldsymbol{G}}|\boldsymbol{G}')P(\mathcal{G}|\boldsymbol{K}, \boldsymbol{G}')P(\boldsymbol{K}|\boldsymbol{G}')P(\boldsymbol{G}')} \end{aligned} \tag{9.2}$$

This LR is constructed such that it (in principle) is applicable in all possible scenarios arising from crime cases.

For the example above with H_p: (G_V, G_S) and H_d: (G_V, G_U) the known profiles are thus $\boldsymbol{K} = (G_V, G_S)$. Assume that G_S has alleles not present in $\mathcal{G}$ implying that allelic drop-out must have

occurred if the suspect is a true contributor to the stain. Furthermore, all alleles in $\mathcal{G}$ is accounted for by (G_V, G_S). Then the *LR* is given by:

$$\begin{aligned} LR &= \frac{P(\mathcal{Q}|\mathcal{G}, G_V, G_S)P(\mathcal{G}|G_V, G_S)P(G_V, G_S)}{\sum_{G_U \in \mathcal{C}_d} P(\mathcal{Q}|\mathcal{G}, G_V, G_U)P(\mathcal{G}|G_V, G_U)P(G_S|G_V, G_U)P(G_U, G_V)} \\ &= \frac{P(\mathcal{Q}_{\text{mis}}|\mathcal{Q}_{\text{obs}}, \mathcal{G}, G_V, G_S)P(\mathcal{Q}_{\text{obs}}|\mathcal{G}, G_V, G_S)P(\mathcal{G}_{\text{mis}}, \mathcal{G}_{\text{obs}}|G_V, G_S)}{\sum_{G_U \in \mathcal{C}_d} P(\mathcal{Q}_{\text{mis}}|\mathcal{Q}_{\text{obs}}, \mathcal{G}, G_V, G_U)P(\mathcal{Q}_{\text{obs}}|\mathcal{G}, G_V, G_U)P(\mathcal{G}_{\text{mis}}, \mathcal{G}_{\text{obs}}|G_V, G_U)P(G_U|G_V, G_S)}, \end{aligned}$$

where $P(\mathcal{G}_{\text{mis}}, \mathcal{G}_{\text{obs}}|G_V, G_S) = 1$ since $(G_V, G_S) \equiv (\mathcal{G}_{\text{mis}}, \mathcal{G}_{\text{obs}})$. Assume further that $\mathcal{C}_d = \{G_U : (G_V, G_U) \equiv \mathcal{G}_{\text{obs}}\}$, i.e. the set of possible unknown profiles is restricted to be consistent with the observed alleles when combined with G_V. Thus, $P(\mathcal{G}_{\text{obs}}, \mathcal{G}_{\text{mis}}|G_V, G_U) = 1$ and *LR* reduces further:

$$LR = \frac{P(\mathcal{Q}_{\text{mis}}|\mathcal{Q}_{\text{obs}}, \mathcal{G}, G_V, G_S)P(\mathcal{Q}_{\text{obs}}|\mathcal{G}, G_V, G_S)}{\sum_{G_U \in \mathcal{C}_d} P(\mathcal{Q}_{\text{obs}}|G_V, G_U)P(G_U|G_V, G_S)}$$

9.4 Future research

9.4.1 Replicates

When a sample is taken from a crime scene the number of molecules may be limited, e.g. does dead hair follicles only contain limited amounts of DNA and similarly for 'touch DNA' which is biological material transferred by physical contact (Gill and Buckleton, 2010a). Let N be the number of DNA molecules present after extraction and n be the number of replicates, R_i, made based on the N molecules. For the n replicates to be comparable in terms of drop-outs (and possibly stutters and contamination) it is desirable for the amount of DNA to be evenly distributed among $R_1, \ldots, R_n$, e.g. for $n = 3$ one could imagine to have approximately 30% in each replicate leaving 10% of the extracted DNA in the tube.

Let π_A denote the aliquot proportion, then this sampling scheme implies that $R_1 \sim \text{bin}(N, \pi_A)$ and $(R_i|R_1, \ldots, R_{i-1}) \sim \text{bin}(N - \sum_{j=1}^{i-1} R_j, \pi_A/[1 - (i-1)\pi_A])$ for $j = 2, \ldots, n$. It is easy to verify that this construction yields the expected values as desired: $\mathbb{E}(R_1) = N\pi_A$ and

$$\mathbb{E}(R_i) = \mathbb{E}[\mathbb{E}(R_i|R_1, \ldots, R_{i-1})] = \frac{[N - (i-1)N\pi_A]\pi_A}{1 - (i-1)\pi_A} = N\pi_A.$$

Furthermore, this implies that $(\boldsymbol{R}, Q) = (R_1, \ldots, R_n, Q) \sim \text{mult}(N, \{\mathbf{1}\pi_A, 1 - n\pi_a\})$, where Q is the remaining extract. Assume that there need to be M molecules of an allele prior to PCR in order to be detected by the CCD camera in the electrophoresis machine post-PCR. Hence, for the allele to be detected in each replicate we require that $R_i > M$ for all i:

$$P(R_1 > M, R_2 > M, \ldots, R_n > M) = \sum_{r_1 > M}^{N} \sum_{r_2 > M}^{N-r_1} \cdots \sum_{r_n > M}^{N-r_+} P(R_1 = r_1, R_2 = r_2, \ldots, R_n = r_n), \quad (9.3)$$

where $r_+ = \sum_{i=1}^{n-1} r_i$. This probability depends on several factors but most importantly $\delta = N - nM$. For small δ the probability that the allele has dropped out in at least one of the replicates is considerable, and for negative δ we are sure to have drop-outs. However, when $\delta >> 0$ the probability of drop-outs in any of the replicates is minimal, i.e. when the amount of DNA is large all the replicates should have all alleles present.

In low template DNA (LT-DNA, formerly known as Low Copy Number DNA, LCN-DNA, Gill and Buckleton (2010a)) it is common to use the 'biological model' to form a so-called 'consensus profile' (Buckleton et al., 2005, Chapter 8). That is, only alleles present in at least two replicates are reported in the consensus profile (Gill et al., 2000). However, from the probability in (9.3) it is for small N very likely that an allele present in some replicates is absent in others. Hence, the definition of a consensus profile may not be the best approach when it is expected that the replicates will show different alleles for small amounts of DNA, which is the case for LT-DNA. A better method would be to model the negative correlation between peak intensities of replicates.

In the left panel of Figure 9.1 the probability that the consensus profile excludes a true allele is plotted for two and three replicates against the total amount of extracted DNA, i.e. $2 \times P(R_1 < M, R_2 > M)$ and $3 \times P(R_1 < M, R_2 < M, R_3 > M)$, where permutation of replicates induce the multiplication of weights. It is assumed that in order to trigger the observation of an allele using a 50 rfu threshold it is required to have 50 pg of DNA material prior to PCR. Furthermore, for the two replicate case all of the extracted DNA is used in equal amounts. For the three replicate situation it is intended to assign 30% of the total DNA to each replicate.

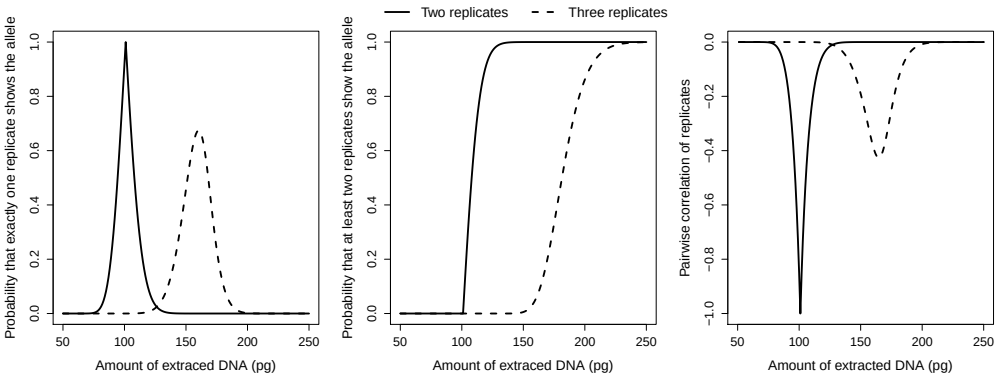

Figure 9.1: Left: Probability that the 'consensus profile' excludes a true allele for two and three replicates. Centre: Probability that the allele will be included in the 'consensus profile' for two and three replicates. Right: Pairwise correlation between consensus profile inducing replicates.

From the curves in the left panel of Figure 9.1 it is evident that for small and large amounts of DNA the probabilities are effectively zero. For the small values this is because neither of the replicates have observed alleles (no allele in consensus profile due to drop-out in both replicates), whereas for the large values it is because the allele is observed in all replicates (allele in

consensus profile). The maximum are respectively at 101 pg and 160 pg while the ranges where the probabilities are larger than 10^{-3} are 75-136 pg and 112-201 pg, for two and three replicates.

In the centre panel of Figure 9.1 the probability that the consensus profile will include the allele for two and three-replicates is plotted against the amount of extracted DNA. For an allele to be present in the consensus profile it must be detected at least twice:

$$P(R_1 > M, R_2 > M) \qquad \text{and} \qquad 3 \times P(R_1 < M, R_2 > M, R_3 > M) + P(R_1 > M, R_2 > M, R_3 > M)$$

To be 99.9% certain that an allele is present in the consensus profile the minimum required amount of extracted DNA are 140 pg for two replicates and 245 pg for three replicates using the assumed aliquot sampling scheme.

Define the indicator variables T_i which are 1 if $R_i > M$ and 0 otherwise. Hence, T_i indicates whether replicate i triggers the observation of an allele above the threshold. The "consensus inducing correlations" are thus $\mathbb{C}\text{or}(T_1, T_2)$ for two replicates and similarly $\mathbb{C}\text{or}(T_1, T_2|T_3 = 0)$ for three replicates. The latter correlation is naturally subject to permutation of replicates, but since the amount of DNA for one replicate, here R_3, is less than M, the two other replicates need to show the allele for it to be included in the consensus profile. The right panel of Figure 9.1 shows the negative correlations as expected due to the limited amount of DNA. For two replicates the pairwise correlation is approximately equal to the negative probability of the left panel of Figure 9.1.

The general picture from the model and analysis of replicates indicate that the concept of the consensus profile (or biological model) is flawed, due to the disproportion between expected peak intensities and consensus profile construction. However, it should be added that the figures above are computed without taking measurement error, PCR efficiency variation, quantification inaccuracy, etc. into account. A more refined model should include these and other factors to be applicable to real STR data.

9.4.2 The number of contributors

When evaluating DNA mixtures a source of uncertainty is the number of contributors. Lauritzen and Mortera (2002) derived an upper bound on the number of unknown contributors worth considering (typically) under H_d. That is, the bound b is computed such that if the number of unknown profiles x is larger than $\lceil b \rceil$, the evidence is less favourable to the defendant than with $x = \lceil b \rceil$. However, this bound is computed without taking the quantitative part of the evidence into consideration and may therefore yield an inaccurate bound for $LR = P(\mathcal{Q}, \mathcal{G}, \boldsymbol{K}|H_p)/P(\mathcal{Q}, \mathcal{G}, \boldsymbol{K}|H_d)$.

9.4.3 Distribution of $\max_{\boldsymbol{G}} L(\mathcal{Q}|\boldsymbol{G})$ - optimisation over a discrete space

In relation to the problem above, it is relevant to be able to quantify the distribution of $L(\mathcal{Q}|\boldsymbol{G})$. How does one measure the significance in the $L(\mathcal{Q}|\boldsymbol{G})$-value when changing the number of contributors m? And how is this related to the mixture proportions $\boldsymbol{\alpha}$? For a fixed combination of

profiles, going from m to $m-1$ contributors is equivalent to setting $\alpha_1 = 0$. However, since the greedy algorithm searches over all possible combination in the discrete space $\mathcal{G}$, it may be inappropriate to rely on asymptotic theory or other common approaches to test H_0: $\alpha_1 = 0$ against H_1:$\alpha_1 > 0$.

9.4.4 Estimation of $P(D)$ using the floating threshold methodology

In the drop-out model of Chapter 6 the limit of detection threshold was fixed at 50 rfu. However, if the STR signal is assigned positive and negative by the floating threshold methodology (Chapter 7), the threshold is not fixed and the previous definition of a dropout, $D = \{h < 50\}$, does not apply. The definition of the drop-out probability on page 170 as an integral may be used in this setting. That is, the quantitative data is spilt into two disjoint partitions where the noise part (off-ladder observation not in pull-up position) is used to determine T and is therefore independent of the quantitative signal in the remaining part. Hence, it would be possible to estimate a mean, μ_h, and standard deviation, σ_h, for the peak heights and evaluate $P(D;\mu_h,\sigma_h) = P(h < T;\mu_h,\sigma_h) = \int_0^T f(h;\mu_h,\sigma_h)\,\mathrm{d}h$.

9.4.5 Evaluating the entire signal

As mentioned in Chapter 1 the use of threshold or limit of detection imply the possibility for drop-out. In that chapter the argument for using a threshold strategy in this thesis were to limit the set of possible combinations that were needed to evaluate *LR*. However, it may be possible to evaluate the entire STR signal by including all observations above a given limit, 5 rfu say. This would lead to more complicated expressions for the *LR*, however with a gain in conceptual clarity since assignment of positive/negative alleles is superfluous. Using this methodology, especially the $P(\mathcal{E}|H_d)$ could imply a summation over a huge set which would be computationally intense.

However, the terms in $P(\mathcal{E}|H_p)$ and $P(\mathcal{E}|H_d)$ that would have numerical impact on the *LR* would be those including the observed alleles with the strongest signals. Often this would be those associated with the alleles in $\boldsymbol{K}$. However this need not to be the case, but searching for a best matching pair of profiles would still be possible. For the evaluation of *LR* to be operational, it might be necessary to use importance sampling in order to evaluate the sum in the denominator since fewer known profiles is specified by H_d than by H_p. Assume that the hypothesis H_d states that the observed crime scene stain was a two-person DNA mixture, then correcting for stutters and pull-up effects, it may be possible to determine a best matching pair of profiles $\hat{\boldsymbol{G}}$. This best matching configuration is then applicable as "reference profiles" for importance sampling similar to the construction in Section 5.6.

Let $\mathcal{E}$ denote the signal obtained from the EPG based on a crime related sample, e.g. a sample taken from a scene of crime. When evaluating the sample we are interested in $P(\mathcal{E}|H_a)$ for some H_a-hypothesis. H_a induces a discrete set of DNA profiles and we denote this $\mathcal{C}_a = \{\boldsymbol{G} : \boldsymbol{G} \equiv H_a\}$. Furthermore, H_a may specify further evidence in terms of DNA profiles of identified individuals. Let $\boldsymbol{K}$ denote the common set of known profiles of the two hypotheses evaluated in the *LR*. For example, in a two-person DNA mixture $\boldsymbol{K}$ may be the profiles of a victim and the suspect,

$\boldsymbol{K} = (G_V, G_S)$. Thus the likelihood ratio is $LR = P(\mathcal{E}, \boldsymbol{K}|H_p)/P(\mathcal{E}, \boldsymbol{K}|H_d)$. This LR is evaluated by summing in both numerator and denominator over profiles in $\mathcal{C}_p$ and $\mathcal{C}_d$, respectively. That is, $P(\mathcal{E}, \boldsymbol{K}|H_a) = \sum_{\boldsymbol{G}\in\mathcal{C}_a} P(\mathcal{E}, \boldsymbol{K}|\boldsymbol{G})P(\boldsymbol{G})$.

We assume that given $\boldsymbol{G}$ no other profiles affect the observed signal. In particular this is true for the known profiles, $\boldsymbol{K}$. Hence, $\mathcal{E}$ and $\boldsymbol{K}$ are conditionally independent given $\boldsymbol{G}$: $P(\mathcal{E}, \boldsymbol{K}|\boldsymbol{G}) = P(\mathcal{E}|\boldsymbol{G})P(\boldsymbol{K}|\boldsymbol{G})$. For each set of profiles $\boldsymbol{G} \in \mathcal{C}_a$ a set of stutters and on-ladder pull-up peaks are induced. Let $\boldsymbol{S_G}$ and $\boldsymbol{P_G}$ denote these "derivatives", where $\boldsymbol{S_G}$ includes both stutters (first, second, third, etc.) and back-stutters. Furthermore, for each $\boldsymbol{G}$ the allelic ladder, L, is known and fixed.

Given $\boldsymbol{G}$ the observed signal, $\mathcal{E}$, may decomposed into five parts that constitute a STR signal:

- Off-ladder noise, $\mathcal{E}_n^{\bar{L}}$ which are all intensity observations in off-ladder position and not in possible pull-up position. $\mathcal{E}_n^{\bar{L}}$ is fixed for all $\boldsymbol{G}$ since the it only rely on the fixed ladder, L.
- The signal due to the proposed profiles in $\boldsymbol{G}$: $\mathcal{E}_{\boldsymbol{G}}$.
- The signal due to stutters induced by profiles in $\boldsymbol{G}$: $\mathcal{E}_{\boldsymbol{S_G}}$.
- The signal due to pull-up peaks induced by profiles in $\boldsymbol{G}$ and $\boldsymbol{S_G}$: $\mathcal{E}_{\boldsymbol{P_G}}$.
- On-ladder noise, $\mathcal{E}_n^{L}$ which are all on-ladder observations not ascribed to $\boldsymbol{G}$ and its derivatives.

Using this decomposition we have for $\boldsymbol{G} \in \mathcal{C}_a$:

$$\begin{aligned} P(\mathcal{E}|\boldsymbol{G}) &= P(\mathcal{E}_n^{L}|\mathcal{E}_{\boldsymbol{P_G}},\mathcal{E}_{\boldsymbol{S_G}},\mathcal{E}_{\boldsymbol{G}},\mathcal{E}_n^{\bar{L}},\boldsymbol{G})P(\mathcal{E}_{\boldsymbol{P_G}}|\mathcal{E}_{\boldsymbol{S_G}},\mathcal{E}_{\boldsymbol{G}},\mathcal{E}_n^{\bar{L}},\boldsymbol{G})P(\mathcal{E}_{\boldsymbol{S_G}}|\mathcal{E}_{\boldsymbol{G}},\mathcal{E}_n^{\bar{L}},\boldsymbol{G})P(\mathcal{E}_{\boldsymbol{G}}|\mathcal{E}_n^{\bar{L}},\boldsymbol{G})P(\mathcal{E}_n^{\bar{L}}|\boldsymbol{G}) \\ &= P(\mathcal{E}_n^{L}|\mathcal{E}_n^{\bar{L}},\boldsymbol{G})P(\mathcal{E}_{\boldsymbol{P_G}}|\mathcal{E}_{\boldsymbol{S_G}},\mathcal{E}_{\boldsymbol{G}},\mathcal{E}_n^{\bar{L}},\boldsymbol{G})P(\mathcal{E}_{\boldsymbol{S_G}}|\mathcal{E}_{\boldsymbol{G}},\mathcal{E}_n^{\bar{L}},\boldsymbol{G})P(\mathcal{E}_{\boldsymbol{G}}|\mathcal{E}_n^{\bar{L}},\boldsymbol{G})P(\mathcal{E}_n^{\bar{L}}), \end{aligned} \tag{9.4}$$

where $P(\mathcal{E}_n^{\bar{L}}|\boldsymbol{G}) = P(\mathcal{E}_n^{\bar{L}})$ since it is fixed for all profiles $\boldsymbol{G}$ and thus cancels out when forming the likelihood ratio. It is likely that some of the terms in (9.4) can be simplified due to conditional independence given $\boldsymbol{G}$. For example, may the on-ladder noise, $\mathcal{E}_n^{L}$, be independent of the off-ladder noise, $\mathcal{E}_n^{\bar{L}}$, given $\boldsymbol{G}$ when the parameters of $P(\mathcal{E}_n^{\bar{L}})$ is determined, i.e. $P(\mathcal{E}_n^{L}|\mathcal{E}_n^{\bar{L}}, \boldsymbol{G}) = P(\mathcal{E}_n^{L}|\mathcal{E}_n^{\bar{L}}, \boldsymbol{G})$. The LR is formed by a hypothesis specific ratio of the expression in (9.4):

$$LR = \frac{\sum_{\boldsymbol{G}\in\mathcal{C}_d} P(\mathcal{E}_n^{L}|\mathcal{E}_n^{\bar{L}},\boldsymbol{G})P(\mathcal{E}_{\boldsymbol{P_G}}|\mathcal{E}_{\boldsymbol{S_G}},\mathcal{E}_{\boldsymbol{G}},\mathcal{E}_n^{\bar{L}},\boldsymbol{G})P(\mathcal{E}_{\boldsymbol{S_G}}|\mathcal{E}_{\boldsymbol{G}},\mathcal{E}_n^{\bar{L}},\boldsymbol{G})P(\mathcal{E}_{\boldsymbol{G}}|\mathcal{E}_n^{\bar{L}},\boldsymbol{G})P(\boldsymbol{K}|\boldsymbol{G})P(\boldsymbol{G})}{\sum_{\boldsymbol{G}'\in\mathcal{C}_d} P(\mathcal{E}_n^{L}|\mathcal{E}_n^{\bar{L}},\boldsymbol{G}')P(\mathcal{E}_{\boldsymbol{P_{G'}}}|\mathcal{E}_{\boldsymbol{S_{G'}}},\mathcal{E}_{\boldsymbol{G}'},\mathcal{E}_n^{\bar{L}},\boldsymbol{G}')P(\mathcal{E}_{\boldsymbol{S_G}}|\mathcal{E}_{\boldsymbol{G}'},\mathcal{E}_n^{\bar{L}},\boldsymbol{G}')P(\mathcal{E}_{\boldsymbol{G}'}|\mathcal{E}_n^{\bar{L}},\boldsymbol{G}')P(\boldsymbol{K}|\boldsymbol{G}')P(\boldsymbol{G}')}$$

As in Section 9.3 we consider a two-person DNA mixture with known victim profile G_V and suspect profile G_S where H_p:(G_V, G_S) and H_d: (G_V, G_U). Due to limited space we define $\boldsymbol{G}_{V,S} = (G_V, G_S)$ and $\boldsymbol{G}_{U,S} = (G_U, G_S)$, then the likelihood ratio is

$$LR = \frac{P(\mathcal{E}_n^{L}|\mathcal{E}_n^{\bar{L}},\boldsymbol{G}_{V,S})P(\mathcal{E}_{\boldsymbol{P_{G_{V,S}}}}|\mathcal{E}_{\boldsymbol{S_{G_{V,S}}}},\mathcal{E}_{\boldsymbol{G}_{V,S}},\mathcal{E}_n^{\bar{L}},\boldsymbol{G}_{V,S})P(\mathcal{E}_{\boldsymbol{S_{G_{V,S}}}}|\mathcal{E}_{\boldsymbol{G}_{V,S}},\mathcal{E}_n^{\bar{L}},\boldsymbol{G}_{V,S})P(\mathcal{E}_{\boldsymbol{G}_{V,S}}|\mathcal{E}_n^{\bar{L}},\boldsymbol{G}_{V,S})}{\sum_{G_U\in\mathcal{C}_d} P(\mathcal{E}_n^{L}|\mathcal{E}_n^{\bar{L}},\boldsymbol{G}_{V,U})P(\mathcal{E}_{\boldsymbol{P_{G_{V,U}}}}|\mathcal{E}_{\boldsymbol{S_{G_{V,U}}}},\mathcal{E}_{\boldsymbol{G}_{V,U}},\mathcal{E}_n^{\bar{L}},\boldsymbol{G}_{V,U})P(\mathcal{E}_{\boldsymbol{S_{G_{V,U}}}}|\mathcal{E}_{\boldsymbol{G}_{V,U}},\mathcal{E}_n^{\bar{L}},\boldsymbol{G}_{V,U})P(\mathcal{E}_{\boldsymbol{G}_{V,U}}|\mathcal{E}_n^{\bar{L}},\boldsymbol{G}_{V,U})}.$$

Bibliography

Alaeddini, R., S. J. Walsh, and A. Abbas (2010). Forensic implications of genetic analyses from degraded DNA - A review. *Forensic Science International: Genetics 4*(3), 148–157.

Alonso, A. et al. (2005). Challenges of DNA profiling in mass disaster investigations. *Croatian Medical Journal 46*(4), 540–548.

Applied Biosystems (2000). *GeneScan Reference Guide - Chemistry Reference for the ABI PRISM 310 Genetic Analyzer*. Applied Biosystems. Figure 'Virtual Filter Set F', pp. 4-10.

Applied Biosystems (2006). *AmpFℓSTR SGM Plus PCR Amplification Kit User's Manual*. Applied Biosystems.

Ayres, K. L. (2000). A two-locus forensic match probability for subdivided populations. *Genetica 108*, 137–143.

Balding, D. J. (2003). Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology 63*, 221–230.

Balding, D. J. (2005). *Weight-of-evidence for Forensic DNA Profiles*. Chichester, West Sussex: John Wiley & Sons, Ltd.

Balding, D. J. and J. S. Buckleton (2009). Interpreting low template DNA profiles. *Forensic Science International: Genetics 4*(1), 1–10.

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

Balding, D. J. and R. A. Nichols (1995). A method for quantifying differentiation between

populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica 96*, 3–12.

Balding, D. J. and R. A. Nichols (1997). Significant genetic correlations among caucasians at forensic DNA loci. *Heredity 78*(6), 583–589.

Barndorff-Nielsen, O. E. and D. R. Cox (1994). *Inference and Asymptotics*. Number 52 in Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Bender, K., M. J. Farfan, and P. M. Schneider (2004). Preparation of degraded human DNA under controlled conditions. *Forensic Science International 139*(2-3), 135–140.

Bill, M. et al. (2005). PENDULUM - a guideline-based approach to the interpretation of STR mixtures. *Forensic Science International 148*, 181–189.

Box, G. E. P. and N. R. Draper (1987). *Empirical model-builing and response surfaces*. Wiley.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*, 1–3.

Buckleton, J. S. and J. M. Curran (2008). A discussion of the merits of random man not excluded and likelihood ratios. *Forensic Science International: Genetics 2*, 343–348.

Buckleton, J. S., C. M. Triggs, and S. J. Walsh (2005). *Forensic DNA evidence interpretation*, pp. 217–274. Boca Raton, FL: CRC Press.

Budowle, B. and T. R. Moretti (1999). Genotype profiles for six population groups at the 13 CODIS short tandem repeat core loci and other PCR-based loci. *Forensic Science Communications*.

Butler, J. M. (2005). *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* (2 ed.). Burlington, MA: Elsevier Academic Press Inc., U.S.

Clayton, T. M., J. P. Whitaker, R. Sparkes, and P. D. Gill (1998). Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International 91*, 55–70.

Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution 23*(1), 72–84.

Cockerham, C. C. (1973). Analysis of gene frequencies. *Genetics 74*(4), 679–700.

Colotte, M., V. Couallier, S. Tuffet, and J. Bonnet (2009). Simultaneous assessment of average fragment size and amount in minute samples of degraded DNA. *Analytical Biochemistry 388*(2), 345–347.

Cook, O. and L. Dixon (2006). The prevalence of mixed DNA profiles in fingernail samples taken from individuals in the general population. *Forensic Science International: Genetics 1*(1), 62–68.

Cowell, R. G. (2009). Validation of an STR peak area model. *Forensic Science International: Genetics 3*(3), 193–199.

Cowell, R. G., S. L. Lauritzen, and J. Mortera (2007a). A gamma model for DNA mixture analyses. *Bayesian Analysis 2*(2), 333–348.

Cowell, R. G., S. L. Lauritzen, and J. Mortera (2007b). Identification and separation of DNA mixtures using peak area information. *Forensic Science International 166*, 28–34.

Cowell, R. G., S. L. Lauritzen, and J. Mortera (2010). Probabilistic expert systems for handling artifacts in complex DNA mixtures. *Forensic Science International: Genetics*. In Press.

Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics 29*(2), 357–372.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. Chapman and Hall Ltd.

Curran, J. M. (2008). A MCMC method for resolving two person mixtures. *Science & Justice 48*, 168–177.

Curran, J. M., J. S. Buckleton, C. M. Triggs, and B. S. Weir (2002). Assessing uncertainty in DNA evidence caused by sampling effects. *Science and Justice 42*(1), 29–37.

Curran, J. M., C. M. Triggs, J. S. Buckleton, and B. S. Weir (1999). Interpreting DNA mixtures in structured populations. *Journal of Forensic Science 44*(5), 987–995.

Curran, J. M. and T. Tvedebrink (2010a). DNAtools - a R package for forensic DNA database analysis. *Journal of Computational Statistics*. Manuscript in preparation.

Curran, J. M. and T. Tvedebrink (2010b). *DNAtools: Statistical functions for analysing forensic DNA databases*. R package version 0.1.

Curran, J. M., S. J. Walsh, and J. S. Buckleton (2007). Empirical testing of estimated DNA frequencies. *Forensic Sciences International: Genetics 1*, 267–272.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Dixon, L. A. et al. (2006). Analysis of artificially degraded DNA using STRs and SNPs - results of a collaborative European (EDNAP) exercise. *Forensic Science International 164*(1), 33–44.

Donnelly, P. (1995a). Match probability calculations for multi-locus DNA profiles. *Genetica 96*, 55–67.

Donnelly, P. (1995b). Nonindependence of matches at difference loci in DNA profiles: quantifying the effect of close relatives on the match probability. *Heredity 75*, 26–34.

Evett, I. W., P. D. Gill, and J. A. Lambert (1998). Taking account of peak areas when interpreting mixed DNA profiles. *Journal of Forensic Sciences 43*(1), 62–69.

Evett, I. W. and B. S. Weir (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sunderland, MA: Sinauer Associates.

Fields, C. A. and A. H. Welsh (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society. Series B, Statistical methodology 69*(3), 369–390.

Gilder, J. R., T. E. Doom, K. Inman, and D. E. Krane (2007). Run-Specific Limits of Detection and Quantitation for STR-based DNA Testing. *Journal of Forensic Science 52*(1), 97–101.

Gill, P. D. et al. (1998). Interpreting simple STR mixtures using allele peak areas. *Forensic Science International 91*(1), 41–53.

Gill, P. D. et al. (2006). DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Science International 160*(2-3), 90–101.

Gill, P. D. and J. S. Buckleton (2010a). A universal strategy to interpret DNA profiles that does not require a definition of low-copy-number. *Forensic Science International: Genetics 4*(4), 221–227.

Gill, P. D. and J. S. Buckleton (2010b). Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. *Journal of Forensic Sciences 55*(1), 265–268.

Gill, P. D., J. M. Curran, and K. Elliot (2005). A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. *Nucleic Acids Research 33*(2), 632–643.

Gill, P. D., J. Whitaker, C. Flaxman, N. Brown, and J. S. Buckleton (2000). An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International 112*(1), 17–40.

Green, P. J. and J. Mortera (2009). Sensitivity of inferences in forensic genetics to assumptions about founding genes. *Annals of Applied Statistics 3*(2), 731–763.

Green, R., I. Roinestad, C. Boland, and L. Hennessy (2005). Developmental Validation of the Quantifiler™ Real-Time PCR kits for the Quantification of Human Nuclear DNA samples. *Journal of Forensic Science 50*(4), 809–825.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science 28*(706), 49–50.

Harrell Jr., F. E. (2001). *Regression Modeling Strategies*. Springer.

Holsinger, K. E. (1999). Analysis of genetic diversity in geographically structure populations: A bayesian perspective. *Hereditas 130*, 245–255.

Holsinger, K. E. and B. S. Weir (2009). Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews. Genetics 10*(9), 639–650.

Irwin, J. A. et al. (2007). Application of low copy number STR typing to the identification of aged, degraded skeletal remains. *Journal of Forensic Sciences 52*(6), 1322–1327.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1997). *Discrete Multivariate Distributions*. Wiley.

Lange, K. (1993). Match probabilities in racially admixed populations. *American Journal of Human Genetics 52*, 305–311.

Lange, K. (1995a). Applications of the Dirichlet distribution to forensic match probabilities. *Genetica 96*, 107–117.

Lange, K. (1995b). *Mathematical and Statistical Methods for Genetic Analysis* (2 ed.). Springer.

Laurie, C. and B. S. Weir (2003). Dependency effects in multi-locus match probabilities. *Theoretical Population Biology 63*, 207–219.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

Lauritzen, S. L. and J. Mortera (2002). Bounding the number of contributors to mixed DNA stains. *Forensic Science International 130*(2-3), 125–126.

Little, R. and D. Rubin (2002). *Statistical Analysis with missing data* (2 ed.). Wiley.

Maimon, G. (2010). *A Bayesian approach to the statistical interpretation of DNA evidence*. Ph. D. thesis, Department of Mathematics and Statistics, McGill University, Montreal, Canada.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. Chapman and Hall.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika 49*(1-2), 65–82.

Mueller, L. D. (2008). Can simple populations genetic models reconcile partial match frequencies observed in large forensic databases? *Journal of Genetics 87*(2), 101–107.

Neerchal, N. K. and J. G. Morel (2005). An improved method for the computation of maximum likelihood estimates for multinomial overdispersion models. *Computational Statistics & Data Analysis 49*, 33–43.

Nichols, R. A. and D. J. Balding (1991). Effects of population structure on DNA fingerprint analysis in forensic science. *Heredity 66*, 297–302.

Paul, S. R., U. Balasooriya, and T. Banerjee (2005). Fisher information matrix for the Dirichlet-multinomial distribution. *Biometrical Journal 47*(2), 230–236.

Perlin, M. W. and B. Szabady (2001). Linear mixture analysis: A mathematical approach to resolving mixed DNA samples. *Journal of Forensic Science 46*(6), 1372–1378.

Petricevic, S. et al. (2009). Validation and development of interpretation guidelines for low copy number (LCN) DNA profiling in New Zealand using the AmpFℓSTR SGM Plus(TM) multiplex. *Forensic Science International: Genetics In Press, Corrected Proof*.

Phillips, C., T. Tvedebrink, et al. (2010). Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel. *Forensic Science International: Genetics*. In Press.

Prinz, M. et al. (2007). DNA Commision of the International Society for Forensic Genetics (ISFG): Recommendations regarding the role of forensic genetics for disaster victim identification (DVI). *Forensic Science International: Genetics 1*(1), 3–12.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rannala, B. and J. A. Hartigan (1996). Estimating gene flow in island populations. *Genetical Research 67*, 147–158.

Robert, C. P. and G. Casella (2004). *Monte Carlo Statistical Methods* (2 ed.). Springer.

Samanta, S., Y.-J. Li, and B. S. Weir (2009). Drawing inferences about the coancestry coefficient. *Theoretial Population Biology 75*, 312–319.

Schneider, P. M. et al. (2004). STR analysis of artificially degraded DNA - results of a collaborative European exercise. *Forensic Science International 139*(2-3), 123–134.

Song, Y. S. and M. Slatkin (2007). A graphical approach to multi-locus match probabilitiy computation: Revisiting the product rule. *Theoretical Population Biology 72*, 96–110.

Troyer, K., T. Gilroy, and B. Koeneman (2001). A nine STR locus match between two apparent unrelated individuals using AmpFℓSTR Profiler Plus™ and COfiler™. *Proceedings of the Promega 12th International Symposium on Human Identification*.

Tvedebrink, T. (2009). *dirmult: Estimation in Dirichlet-Multinomial distribution*. R package version 0.1.3.

Tvedebrink, T. (2010). Overdispersion in allelic counts and θ-correction in forensic genetics. *Theoretical Population Biology*. In Press.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2008). Amplification of DNA mixtures - Missing data approach. *Forensic Science International: Genetics Supplement Series 1*, 664–666.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2009). Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Science International: Genetics 3*(4), 222–226.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2010a). Evaluating the weight of evidence using quantitative STR data in DNA mixtures. *Journal of the Royal Statistical Society. Series C, Applied statistics*. In Press.

Tvedebrink, T., P. S. Eriksen, H. S. Mogensen, and N. Morling (2010b). Identifying contributors of DNA mixtures by of quantitative information of STR typing. *Journal of Computational Biology*. Accepted for publication.

Ukoumunne, O. C., A. C. Davison, M. C. Gulliford, and S. Chinn (2003). Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. *Statistics in Medicine 22*, 3805–3821.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (4 ed.). Springer.

Votaw, D. F. (1948). Testing compound symmetry in a normal multivariate distribution. *Annals of Mathematical Statistics 19*(4), 447–473.

Wang, T., N. Xue, and J. D. Birdwell (2006). Least-square deconvolution: A framework for interpreting short tandem repeat mixtures. *Journal of Forensic Science 51*(6), 1284–1297.

Weinberg, W. (1908). Über den nachweis der vererbung beim menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg 64*, 368–382.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sinauer Associates, Inc.

Weir, B. S. (2004). Matching and partially-matching DNA profiles. *Journal of Forensic Science 49*(5), 1–6.

Weir, B. S. (2007). The rarity of DNA profiles. *The Annals of Applied Statistics 1*(2), 358–370.

Weir, B. S. and C. C. Cockerham (1984). Estimating *F*-statistics for the Analysis of Population Structure. *Evolution 38*(6), 1358–1370.

Weir, B. S. and W. G. Hill (2002). Esimating F-statistics. *Annual Review of Genetics 36*, 721–750.

Wright, S. (1951). The genetical structure of populations. *Annals of eugenics 15*, 323–354.

Zhou, H. and K. Lange (2010). MM algorithms for some discrete multivariate distributions. *Journal of Computational and Graphical Statistics*. In Press.