Statistical Analysis of DNA Profiles

Robyn L. McClelland

Department of Mathematics and Statistics

McGill University, Montreal

March, 1994

A Thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements of the degree of Master of Science.

Abstract

DNA profiles have become an extremely important tool in forensic investigations, and a match between a suspect and a crime scene specimen is highly incriminating. Presentation of this evidence in court, however, requires a statistical interpretation, one which reflects the uncertainty in the results due to measurement imprecision and sampling variability. No consensus has been reached about how to quantify this uncertainty, and the literature to date is lacking an objective review of possible methods.

This thesis provides a survey of approaches to statistical analysis of DNA profile data currently in use, as well as proposed methods which seem promising. A comparison of frequentist and Bayesian approaches is made, as well as a careful examination of the assumptions required for each method.

Résumé

Les profils d'A[illegible] des outils extrêmement importants dans les enquêtes criminelles. La [illegible] entre un suspect et un spécimen recueilli sur la scène du crime [illegible] hautement incriminante. Toutefois, compte tenu de l'imprécision des mesures et de la variabilité de l'échantillonnage, la démonstration de ce type de preuve devant le tribunal implique une interprétation des statistiques reflétant l'incertitude des résultats. Jusqu'à présent, il n'y a pas eu de consensus sur la manière de quantifier cette incertitude et la littérature ne fournit pas de revue objective des méthodes possibles.

Cette thèse fournit un apperçu des différentes approches présentement utilisées pour l'analyse statistique des profils d'ADN. Elle suggère de plus les méthodes les plus prometteuses. Enfin, une comparaison entre les approches fréquentiste et Bayesienne est présentée aussi qu'un examen détaillé des hypothèses requises pour chacune de ces méthodes.

Acknowledgements

I would like to thank Professor David Wolfson, whose constant guidance and encouragement made this thesis a pleasure to write.

Contents

Chapter 1

Introduction

As DNA profile data presents challenging statistical problems, presentation of this type of evidence in court has been severely debated. Frequently, the DNA evidence may be deemed inadmissible on the grounds that there is no acceptable way to analyze the data. Several possible methods of analysis have been proposed. The purpose of this thesis is to provide a detailed review of these proposals, and of the assumptions on which they are based. In Chapter 2 an overview of the laboratory techniques is provided, along with the basic genetics necessary to understand the procedure.

In Chapter 3 an outline of the frequentist and Bayesian approaches to this problem is given. The details of these methods are given in Chapters 4 and 5 respectively. In Chapter 6 a discussion of the possibility of population substructure and its potential impact on some critical assumptions is presented. Chapter 7 provides a comparison of the various methods which were discussed in Chapters 4 and 5. Appendix A briefly describes three statistical concepts which may be unfamiliar, but which are used in some of the analysis. Appendix B gives a full derivation of a set of estimation equations used by Devlin, Risch, and Roeder (1991). This derivation is nontrivial and was not supplied by Devlin et al.

Chapter 2

An Outline of DNA Profiling

A basic understanding of the cell and its components is fundamental to the understanding of DNA profile analysis. It is the cells which store and transmit the genetic information which makes each individual unique.

Genetic information is stored within the nucleus of each cell. The information is coded in the large molecules known as the nucleic acids, RNA (ribonucleic acid) and DNA (deoxyribonucleic acid). RNA is dispersed throughout the nucleus and the surrounding material known as the cytoplasm. In contrast, DNA is primarily restricted to the nucleus. Each normal cell in an organism will carry a full copy of the genetic information.

Both DNA and RNA are composed of repeating units known as nucleotides. Each nucleotide is a combination of a deoxyribose sugar, a phosphoric acid, and one of four possible nitrogen bases : adenine, cytosine, guanine, or thymine. The only difference between nucleotides is the nitrogen base present, hence the four types of nucleotides are denoted A, C, G, and T respectively, depending on the base. The nucleotides are joined together sequentially (by chemical bonds) and may occur in any order. There is thus an endless array of possible sequences which accounts for the vast degree of genetic variation in human beings.

The physical structure of DNA was described by the Watson-Crick model (1953).

It was discovered that the DNA molecule is not made up simply of one long string of nucleotides but of two such strings bonded together. The two strands twist around around one another in a spiral formation, or a "double helix". The adenine (A) nucleotides in one strand pair only with the thymine (T) nucleotides in the other strand. Similarly, the guanine (G) nucleotides in one strand pair with the cytosine (C) nucleotides in the other. The length of these strands are thus referred to in terms of base pairs. As a result of this preferential base pairing, if the nucleotide sequence of one strand is known then the sequence of the complementary strand is also known.

In each cell of the human body, within the nucleus, there are 23 *pairs* of chromosomes, known as homologous pairs. There is only one DNA molecule present per chromosome, and it may be visualized as a long fiber crossing the length of the chromosome and folding back on itself many times. A *gene* is simply a stretch of this DNA fiber, which occupies a specific position on the chromosome known as its *locus*. A particular locus will be represented twice per cell, once on each chromosome of the homologous pair which carries it. Each chromosome of a pair will carry the same loci, but they do not necessarily have the same genes present at each matching locus. Alternative forms of the same gene are known as *alleles*, and they differ from one another only in the sequence of nucleotide base pairs. If each chromosome in a pair carries the same allele then the pair is said to be *homozygous* at that locus. If two different alleles are found, the pair is *heterozygous*.

An important aspect of the DNA in eukaryotic cells (found in most plants and animals) is the appearance of certain sequences of base pairs which repeat themselves in tandem. These sequences are of a variety of lengths and fall into two classes, moderately repetitive and highly repetitive. In the highly repetitive group the sequence may repeat over a million times. This group is frequently called *satellite DNA*. The term *minisatellite* is used to refer to any short DNA sequence which repeats itself.

Certain loci were discovered to be hypervariable between individuals due to the differing number of repeats of the minisatellites. For a particular locus, a different number of repeats corresponds to a different allele. These loci are often called variable number of tandem repeat (VNTR) loci. This lead to the discovery that a characteristic pattern, based on the number of repeats at each locus, could be constructed for each individual. Moreover, this "fingerprint", or "profile", could be determined from something as small as a hair root or as degraded as an old blood stain. As will be seen, although not unique, careful selection of loci can ensure that an individual's profile occurs with low frequency in the population.

The main method used to produce these profiles is known as Southern blotting. The DNA is isolated from the samples provided and then exposed to a protein known as a restriction enzyme which cuts the DNA at specific sites into many smaller fragments (the restriction enzyme will not cut the repeating sequence). These fragments will have a characteristic length depending on how many minisatellite repeats they contain. The fragments are placed in an electrophoretic gel, in which the shorter fragments will travel farther than the longer fragments. The result is a line of fragments in order of length in base pairs. This line is moved onto a nylon membrane where it is fixed in place, and radioactive probes are used to reveal the positions of the fragments on the membrane. The fragments will appear as bands and the position of the band (i.e. how far it traveled in the gel) in theory reveals its length (a series of fragments of known size is also run through the gel, and the positions of these markers are compared to the positions of the fragments being examined). This banded pattern will be the same for a given individual regardless of the kind of tissue present in the sample.

Two types of probes may be used in the above procedure, single locus probes (SLP's) and multi-locus probes (MLP's). An SLP will reveal only those bands which

are present at one specific locus. Thus, either two bands will show for an individual heterozygous at that locus, or one band for a homozygous individual. An MLP reveals bands from several loci at once, but is a less sensitive technique than the single locus probe. Several single locus probes may be combined to produce a profile for multiple loci.

Many variations on the above technique exist, but the underlying principles are similar. In addition, a technique known as PCR (polymerase chain reaction) analysis is frequently used. This method involves amplification of the hypervariable sequences and is much less time consuming than the southern blotting approach. It is particularly useful when the sample is degraded since it can be used on very small quantities of DNA.

The problem with the above methods is that the translation of "distance fragment travels in gel" to "length of fragment in repeats" is imprecise. There is measurement error involved which may be larger (in terms of base pairs) than the size of the repeat sequence, and misclassification of alleles can occur. In addition, there are certain technical difficulties which may arise. For example, extremely small fragments have been known to migrate off the end of the electrophoretic gel, making them undetectable. Also, two bands which are very similar in length, though not identical, may appear as a single band instead of two distinct bands. This phenomenon is known as *coalescence*. An additional concern is the occurrence of *band shifting* which causes the fragments to appear a uniform amount smaller or larger than they should.

When the DNA from a crime specimen is compared with that of a criminal the two profiles are lined up beside one another. For example, consider the DNA evidence for a hypothetical murder investigation depicted in Figure 1. A specimen (e.g. a bloodstain) was obtained from the crime scene and it can easily be seen by comparing the DNA profiles of the specimen and victim that the specimen was not contributed

by the victim. It may therefore be used to place the suspect at the scene of the crime. Two suspects are apprehended and their DNA profiles are constructed and compared to that of the specimen. It can be seen from Figure 1 that Suspect 1 may be excluded from the investigation, while Suspect 2 would be declared a "match".

A careful study of the statistical methods leading to the declaration of a match and the implications of this declaration are, in part, the topic of this thesis. Other statistical viewpoints that do not depend on "match declarations" are also discussed.

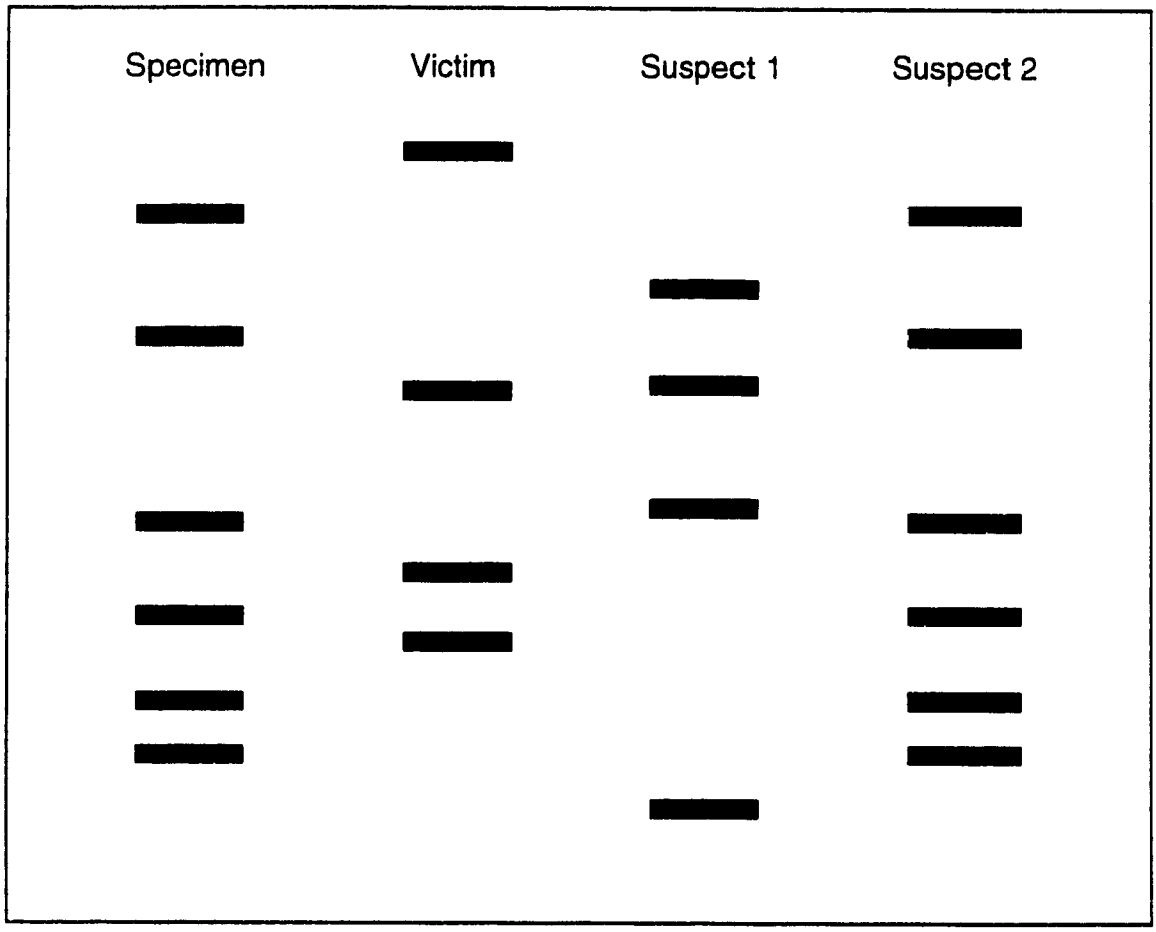

Figure 1. DNA Profiles for a Hypothetical Murder Investigation

Chapter 3

Statistical Approaches to DNA Profile Analysis

Consider a criminal investigation in which a single bloodstain is found at the scene of the crime. A suspect is apprehended and DNA profiles of both suspect and specimen are constructed using methods described in Chapter 2. How can the information contained in the DNA evidence best be presented in a courtroom setting? Predictably, two schools of thought have arisen, one which feels that the DNA evidence should be evaluated in a frequentist manner, and one which promotes Bayesian methods. These approaches are outlined in the following sections.

3.1 Frequentist Approach

In this approach it is first necessary to declare whether there is a match between the two profiles. Since each allele is measured with error, two measurements that are different, but very similar, may still be two measurements of the same allele. Two allele measurements are said to match if they meet some predetermined matching criterion which is based on the distribution of the measurement error (see Section 4.3). Suppose that at a particular locus the DNA of the suspect has allele measurements x_1 and x_2 and the DNA of the specimen has allele measurements y_1 and y_2. There

is said to be a match at that locus if either x_1 matches y_1 and x_2 matches y_2, or x_1 matches y_2 and x_2 matches y_1, since it is impossible to tell from which of the two chromosomes the measurements arise. The DNA profiles of suspect and specimen are said to match if there is a match at every locus which was examined. Otherwise an exclusion is declared and the suspect is released.

It is conceivable however, that the two DNA profiles match by "pure chance" and not because the suspect is actually the guilty party. In order to make a proper presentation of the DNA evidence in court the probability of this event must be estimated. That is, the probability that the DNA profile of an individual chosen at random from the population would match that of the crime specimen. If this probability is "extremely small" (frequently on the order of $1x10^{-5}$) then the suspect is deemed to have committed the crime " beyond a reasonable doubt". To describe the approach more formally some notation is introduced, and some crucial assumptions are outlined. Define the following events :

M = the event that the DNA profile of an individual chosen at random from the population matches the DNA profile of the specimen,

M_i = the event that the two alleles found at locus i for an individual chosen at random from the population match those found at locus i in the specimen.

We have,

$$M = \bigcap_{i=1}^{L} M_i$$

where,

L = total number of loci examined to construct the DNA profiles.

Let Y represent the DNA profile data obtained from the specimen. The probability which must be estimated is $P(M|Y)$.

3.1.1 Presentation of Evidence

When the evidence is finally presented in the courtroom, the estimated probability $\widehat{P}(M|Y)$ is presented. The jury must then decide, based on $\widehat{P}(M|Y)$ and all other available evidence, whether they believe the suspect is guilty or innocent. Technically speaking they are actually carrying out a test of the hypotheses,

H_o : the suspect was chosen at random from the population (i.e. the suspect is innocent),

versus,

H_a : the suspect was not chosen at random from the population (i.e. the suspect is guilty).

The data used is the DNA profile evidence and the test statistic is $\widehat{P}(M|Y)$. The null hypothesis is rejected for "extremely small" values of the test statistic.

3.1.2 Selecting the Loci

Since there are a multitude of loci available it is necessary to decide which ones are the best suited for analysis. This decision must occur at the *design stage* of the experiment (i.e. before viewing the DNA evidence), and thus will be discussed at this early stage. The following comments on this important topic are due to Lange (1991).

Consider a single locus. In constructing the DNA profiles it is desirable to include a locus which has a maximal ability to exonerate an innocent suspect. Let e denote the exclusion probability, which is the probability that a randomly chosen individual will not share a matching genotype with the criminal at a particular locus. It is desired to maximize the exclusion probability. Suppose further that there are n possible alleles at

this locus, with lengths $A_1, \ldots, A_n$, and with population relative frequencies p_j, $j = 1, \ldots, n$ respectively. It follows that,

$$\begin{aligned} e &= \textstyle\sum_{k=1}^{n-1} \sum_{j=k+1}^{n} 2p_j p_k(1 - 2p_j p_k) + \sum_{j=1}^{n} p_j^2(1 - p_j^2) \\ &= \textstyle\sum_{k=1}^{n-1} \sum_{j=k+1}^{n} 2p_j p_k - 2\sum_{k=1}^{n-1} \sum_{j=k+1}^{n} 2p_j^2 p_k^2 + \sum_{j=1}^{n} p_j^2(1 - p_j^2) \\ &= \textstyle\sum_{k=1}^{n} \sum_{j=1}^{n} p_j p_k - \sum_{j=1}^{n} p_j^2 - 2\{\sum_{k=1}^{n} \sum_{j=1}^{n} p_j^2 p_k^2 - \sum_{j=1}^{n} p_j^4\} \\ &\quad + \textstyle\sum_{j=1}^{n} p_j^2(1 - p_j^2) \\ &= 1 - 2\textstyle\sum_{k=1}^{n} \sum_{j=1}^{n} p_j^2 p_k^2 + \sum_{j=1}^{n} p_j^4 \\ &= 1 - 2(\textstyle\sum_{j=1}^{n} p_j^2)^2 + \sum_{j=1}^{n} p_j^4 \end{aligned}$$

The exclusion probability must be maximized with respect to each p_k, $k = 1, \ldots, n$, subject to the constraint that $\sum_{k=1}^{n} p_k = 1$. By using Lagrange multipliers one must solve.

$$-8p_k \sum_{j=1}^{n} p_j^2 + 4p_k^3 + \lambda p_k = 0$$

which implies,

$$-2\sum_{j=1}^{n} p_j^2 + p_k^2 + \frac{\lambda}{4} = 0,$$

or equivalently,

$$p_k = \sqrt{2\sum_{j=1}^{n} p_j^2 - \frac{\lambda}{4}}\,. \tag{3.1}$$

Solving 3.1 simultaneously with the constraint $\sum_{k=1}^{n} p_k = 1$ yields,

$$p_k = \frac{1}{n} \quad for\ k = 1, \ldots, n.$$

In other words, loci should be chosen which have alleles that are as close to equally frequent in the population as possible.

In addition to its ability to exonerate an innocent suspect it is also desired that a locus have maximal ability to incriminate a guilty one. That is, under the hypothesis of guilt the inclusion probability should be maximized. However, it is assumed that the probability of obtaining a match given that the suspect is guilty is one, hence there is nothing to be maximized. Instead Lange considers the Kullback-Leibler

information (Kullback and Leibler, 1951) computed under the assumption of guilt (see Appendix A). This is also a measure of the ability of a locus to point to a guilty person. For a particular locus with n possible alleles let $\tilde{X}$ denote the genotype of the suspect/specimen, and let the operator $\mathcal{E}$ be the expectation under the assumption of guilt. We then define the Kullback-Leibler information,

$$\begin{aligned}
K_{\tilde{X}}(H_G, H_I) &= \mathcal{E}\{\log \tfrac{f_G(\tilde{X})}{f_I(\tilde{X})}\} \\
&= \textstyle\sum_{j=1}^{n} p_j^2 \log \frac{p_j^2}{(p_j^2)^2} + \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} 2p_j p_k \log \frac{2p_j p_k}{(2p_j p_k)^2} \\
&= -\textstyle\sum_{j=1}^{n} p_j^2 \log p_j^2 - 2\sum_{j=1}^{n-1} \sum_{k=j+1}^{n} p_j p_k \log 2p_j p_k \\
&= -\textstyle\sum_{j=1}^{n} p_j^2 \log p_j^2 - \{\sum_{j=1}^{n} \sum_{k=1}^{n} p_j p_k \log p_j p_k - \sum_{j=1}^{n} p_j^2 \log p_j^2\} \\
&= -\textstyle\sum_{j=1}^{n} \sum_{k=1}^{n} p_j p_k \log p_j p_k \\
&= -2\textstyle\sum_{j=1}^{n} p_j \log p_j
\end{aligned}$$

This quantity must be maximized with respect to p_k, $k = 1, \ldots, n$, subject to the constraint $\sum_{k=1}^{n} p_k = 1$. Using Lagrange multipliers the quantity to be maximized is,

$$K_{\tilde{X}}(H_G, H_I) = -2\sum_{j=1}^{n} p_j \log p_j + 2\lambda(\sum_{j=1}^{n} p_j - 1).$$

This yields,

$$\frac{\partial K}{\partial p_k} = -\log p_k - 1 + \lambda = 0$$

which implies,

$$\log p_k = \lambda - 2. \tag{3.2}$$

Equation 3.2 is solved simultaneously with the constraint $\sum_{k=1}^{n} p_k = 1$ to obtain $p_k = \frac{1}{n}$, for $k = 1, \ldots, n$.

Thus under both of the hypotheses of guilt and of innocence the same conclusion is reached, although through consideration of different criteria. The best locus to include in a profile is that for which the alleles are equally frequent.

3.1.3 Current Approach

The most commonly used approach to analyzing DNA profile data is known as "match-binning", which relies on allele probabilities estimated by "binning". The viewpoint is frequentist. The technique of binning will be outlined in Chapter 4 as well as the procedure for calculating $P(M|Y)$ based on binned allele probability estimates.

3.2 Bayesian Method

The Bayesian approach to the forensic identification problem is substantially different from that of match binning. Consider the same scenario as depicted previously. That is, a suspect is apprehended during a forensic investigation and the DNA profiles of both the suspect and a crime scene specimen are constructed. It is not attempted to determine whether or not the two profiles match. Instead it is attempted to assess the posterior odds of guilt given the evidence. Define the following events,

G = the event that the suspect is guilty,

I = the event that the suspect is innocent,

and let,

X = all the DNA profile evidence = $(x, y, z_1, \ldots, z_n)$ where x, y, $z_1, \ldots, z_n$ are respectively the measurements taken at a specific locus for the suspect, the specimen, and a simple random sample of n individuals drawn from a reference population, and,

E = all other available evidence.

It is desired to find the posterior probability of guilt. That is, the probabilty of G based on all the available evidence. According to Bayes' Theorem, the posterior odds

of guilt is proportional to the prior odds of guilt. The constant of proportionality, denote it R, is known as the likelihood ratio, or the Bayes factor:

$$\frac{P(G|X,E)}{P(I|X,E)} = R\frac{P(G|E)}{P(I|E)}, \qquad where \quad R = \frac{P(X|G,E)}{P(X|I,E)}.$$

Deciding on an appropriate prior is a task for the jury. During a trial it is essential that the relationship between prior and posterior odds be made as clear as possible. As will be discussed in a later section it may not even be necessary for the jury to choose a specific prior, simply presenting the likelihood ratio with an appropriate explanation may be acceptable. Alternatively, Devlin, Risch,and Roeder (1991) also suggest the use of a frequentist decision rule of the form,

$$R \in (0, \frac{1}{K}] \quad conclude\ H_I$$

$$R \in (\frac{1}{K}, K] \quad inconclusive$$

$$R \in [K, \infty) \quad conclude\ H_G.$$

They claim that the constant K may be chosen so that the probability of a Type I error is very small. The statistical problem is thus deciding how to estimate R. Three different approaches are considered in Chapter 5. These are due to Berry (1991), Berry, Evett and Pinchin (1992), and Devlin, Risch and Roeder (1991,1992).

3.2.1 Selecting the Loci and Restriction Enzymes

Devlin, Risch, and Roeder (1992) use the term "system" to refer to a particular locus/restriction enzyme combination. It is desired to include in the DNA profiles those systems which provide the most information. In particular, the probability of rejecting H_I when the suspect is innocent should be as close to 0 as possible, while the probability of rejecting H_I when the suspect is guilty should be as close to 1 as possible. Define, $\gamma(\tilde{x}, \tilde{y}|Z, S)$ to be the probability of rejecting H_I in favour of

H_G for a particular system S, where $\tilde{x}$ and $\tilde{y}$ represent the DNA measurements for suspect and specimen respectively and $Z = (z_1, \ldots, z_n)$ is the reference sample of measurements. Ideally, for given realizations $\tilde{x}$ and $\tilde{y}$,

$$\gamma(\tilde{x}, \tilde{y}) = \begin{cases} 1 & \textit{under } H_G \\ 0 & \textit{under } H_I. \end{cases}$$

Similarly, define the following loss function,

$$Loss[\gamma(\tilde{x}, \tilde{y}|Z)] = \begin{cases} (1 - \gamma(\tilde{x}, \tilde{y}|Z))^2 & \textit{under } H_G \\ (\gamma(\tilde{x}, \tilde{y}|Z))^2 & \textit{under } H_I. \end{cases}$$

Squared error is chosen to give more weight to large deviations since these will likely correspond to false conclusions. Define the corresponding risk functions as,

$$R(H_G) = \mathcal{E}[(1 - \gamma(\tilde{x}, \tilde{y}|Z))^2 | H_G]$$

$$R(H_I) = \mathcal{E}[(\gamma(\tilde{x}, \tilde{y}|Z))^2 | H_I],$$

where with a slight abuse of notation $\tilde{x}$ and $\tilde{y}$ are now regarded as random variables and not as realizations (the averaging being done over the possible data values). The Bayes risk is then,

$$r(\beta) = (1 - \beta)R(H_I) + \beta R(H_G)$$

where β is the prior probabilty of guilt (the further averaging being done over the two possible hypotheses). The Bayes' risk may be compared between competing systems, the system with the lower Bayes' risk being the appropriate choice.

In the next chapter match-binning is examined in detail.

Chapter 4

Match-binning

In match-binning, the range of possible allele sizes is divided into classes (called bins), and the sample *bin* frequency is used as an estimate for the frequency of each allele within the bin. These "binned frequencies" may then be used to estimate the probability of a match at a particular locus for the "match-binning" approach (they do not necessarily have to be used in this manner, however this is the current practice by most forensic laboratories).

4.1 Assumptions

The following assumptions are necessary in order to use match-binning

Random Mating It is assumed that individuals mate without regard to genotype, a phenomenon known as random mating. For example, if there is information about the alleles inherited from the mother of a particular individual, this yields no information about which alleles may have been inherited from the father. In other words, the maternal and paternal chromosomes are independent. The most important consequence of this assumption is that, regardless of the observed allele frequencies for a fixed population in any particular generation, after one generation of random mating a state known as Hardy-Weinberg equilibrium

(HWE) is reached.

Suppose that the true population frequency of alleles A_1 and A_2 are p_1 and p_2 respectively. If the population is in HWE then the expected relative frequencies of the genotypes A_1A_1, A_1A_2, and A_2A_2 are respectively p_1^2, $2p_1p_2$, and p_2^2 and these values will not change from one generation to the next. (Note that the observed frequencies will always differ slightly from those expected under HWE, but the expected frequencies will remain constant between generations.)

Suppose that alleles A_j and A_k are found at locus i in the DNA profiles of both the suspect and specimen. Let p_j and p_k denote the respective relative population frequencies of these alleles. Assuming HWE one may write,

$$P(M_i|A_j, A_k) = \begin{cases} 2p_jp_k & if \quad j \neq k \quad (suspect\ heterozygous) \\ {p_j}^2 & if \quad j = k \quad (suspect\ homozygous) \end{cases}$$

Henceforth the terms "random mating assumption" and "HWE assumption" will be used interchangeably as is common in the literature, although the situation is actually that the random mating assumption leads to the state known as HWE. The assumption of random mating is a highly contentious issue and its validity will be discussed further in Chapter 6. It is standard practice to attempt to derive a conservative estimation procedure which deliberately biases the outcome in favour of the defendant in order to compensate for the uncertainty arising from these assumptions and other possible sources of error such as sampling variability which will be encountered. In addition, for the binning approach it is actually the bin relative frequencies that are of interest and not the true allele relative frequencies. The HWE assumption is assumed to hold for the bin relative frequencies as well since they are simply estimates of the true allele relative frequencies. Henceforth the notation p_j will refer to the true population relative frequency of alleles *in bin j*, instead of the true population relative frequency of allele A_j.

Linkage Equilibrium (LE) It will be assumed that the genotypes observed at one locus are independent of those observed at other loci. This will be referred to as the linkage equilibrium assumption although actually the assumption of independence between loci leads to a state known as LE. In this state, given the true allele relative frequencies in the population, the expected relative frequencies of the various possible genotypes remain constant between generations. Assuming LE, P($M|Y$) may be written as,

$$P(M|Y) = P(\bigcap_{i=1}^{L} M_i|Y) = \prod_{i=1}^{L} P(M_i|Y)$$

An important consequence of the LE assumption is that only the single locus case need be considered. The results from distinct loci will simply be multiplied.

Caution must be exercised in applying the LE assumption since loci which are located on the same chromosome are quite likely to be "linked" (not independent), particularly if they are positioned close together. This will be discussed further in Chapter 6.

4.2 Outline of Match-binning Procedure

The basic steps in match-binning are as follows,

(i) Determine the matching criterion.

(ii) Determine whether there is a match or an exclusion. (If there is an exclusion nothing further is required.)

(iii) Establish the bin boundaries.

(iv) Estimate the bin frequencies.

(v) Use the assumption of HWE to estimate $P(M_i|Y)$ as

$$\hat{P}(M_i|Y) = \begin{cases} 2\hat{p}_j\hat{p}_k & if \;\; j \neq k \;\; (suspect\; heterozygous) \\ \hat{p}_j^2 & if \;\; j = k \;\; (suspect\; homozygous) \end{cases}$$

(vi) Use the assumption of LE to estimate $P(M|Y)$ as

$$\hat{P}(M|Y) = \prod_{i=1}^{L} \hat{P}(M_i|Y).$$

The actual process of "binning" involves only steps (iii) and (iv), in which the allele relative frequencies are estimated. All six steps are described in detail in the following sections.

4.3 Determining the Match Criterion

Suppose that two allele measurements $x_i = A_i + \varepsilon_{i_1}$ and $y_i = B_i + \varepsilon_{i_2}$ are obtained from the suspect and specimen respectively. It must be decided whether these are two measurements of the same allele. In terms of hypothesis testing it is desired to test,

$$H_o \; : \; A_i = B_i$$

versus,

$$H_a \; : \; A_i \neq B_i.$$

It is assumed that ε_{i_1} and ε_{i_2} are independent, with $\varepsilon_{i_1} \sim \mathcal{N}(0, \sigma_{i_1}^2)$ and $\varepsilon_{i_2} \sim \mathcal{N}(0, \sigma_{i_2}^2)$.

Under H_o, $\mathcal{E}(x_i - y_i) = 0$ and $var(x_i - y_i) = var(x_i) + var(y_i) = 2\sigma_i^2$, where σ_i denotes the common standard deviation of x_i and y_i. Hence an appropriate test statistic (given that σ_i is unknown) is,

$$T = \frac{x_i - y_i}{\sqrt{2}\hat{\sigma}_i}$$

which is assumed to follow a standard normal distribution. The null hypothesis is rejected for large values of the test statistic. That is, for $|T| \geq z_{(\frac{\alpha}{2})}$ where $z_{(\frac{\alpha}{2})}$ is the $100(1-\frac{\alpha}{2})$ percentile point of the standard normal distribution. It remains to estimate the measurement error standard deviation σ_i.

Berry, Evett and Pinchin (1992) approach this problem by taking duplicate measurements of a large sample (218) of fragment lengths (alleles). Suppose that x_{i_1} and x_{i_2} are two measurements of an allele of length A_i, i=1,...,218. That is,

$$\begin{aligned} x_{i_1} &= A_i + \varepsilon_{i_1} \\ x_{i_2} &= A_i + \varepsilon_{i_2} \qquad i = 1, \ldots, 218 \end{aligned}$$

where ϵ_{i_1} and ϵ_{i_2} are the errors associated with each measurement with $var(\varepsilon_{i_1}) = var(\varepsilon_{i_2}) = \sigma_i^2$. For each allele of length A_i, the measurement error will have a different distribution. It is assumed that the standard deviation of the measurement error (σ_i) is directly proportional to the true allele length A_i, and this assumption appears to be supported empirically (Baird et al., 1986). Since the true allele length is unknown, it is estimated as the average of the two measurements. Therefore on the one hand the estimate of σ_i is given by,

$$\hat{\sigma}_i = c\frac{(x_{i_1} + x_{i_2})}{2} \qquad \textit{for } i = 1, \ldots, 218, \;\; c \textit{ an unknown constant.}$$

On the other hand, using the sample standard deviation based on two observations yields an estimate of,

$$\hat{\sigma}_i' = \frac{|x_{i_1} - x_{i_2}|}{\sqrt{2}}$$

Equating the two estimates yields,

$$\frac{|x_{i_1} - x_{i_2}|}{\sqrt{2}} = c\frac{(x_{i_1} + x_{i_2})}{2}$$

or rearranging,

$$\sqrt{2}\frac{|x_{i_1} - x_{i_2}|}{x_{i_1} + x_{i_2}} = c.$$

For each estimated allele size this quantity was calculated and the mean value of c was found by Berry et al. to be 0.008. Baird et al. (1986) estimate c = .0012 based on 70 duplicate measurements.

Thus if $|x_i - y_i| \leq 0.008 z_{\frac{\alpha}{2}} \sqrt{2} \frac{(x_i + y_i)}{2}$ then a match between the two alleles being measured is declared. The choice of $z_{(\frac{\alpha}{2})}$ varies between laboratories but it is frequently chosen to be an integer k, with $\alpha \approx 0.05$. For example, at one time Lifecodes used $k = 2$ (Baird et al., 1986) but as of 1989 were using k=3 (Berry, 1991). This match criterion is then applied to the entire profile, one locus at a time, and if there is a match at every individual locus examined then it is said that the profiles are a match. Henceforth it will be assumed that a match has been declared.

4.4 Establishing the Boundaries of the Bins

In general, if a bin is centered around an allele of size A_i, the boundaries of the bin should not be closer than 2 standard deviations of the measurement error on either side. This is to ensure that approximately 95% of the measurements which could be due to allele A_i are included in the bin. The exact positioning of the boundaries seems to be somewhat arbitrary. In Section 3.1.2 it was shown that the best loci to include in the profiles are those for which the alleles are equally frequent. In the match-binning approach the allele relative frequencies are estimated by the corresponding bin relative frequencies which implies that in the ideal situation the bins would be equiprobable, and loci should be chosen with this in mind. Alternatively, bin boundaries may be adjusted so as to more closely approximate such a uniform allele frequency distribution.

The *fixed bin approach* sets the bin boundaries in advance and may use the same set of bins for different investigations. By contrast the *floating bin approach* centers a bin around any allele found in the DNA profile of the suspect/specimen. For

each case the floating bin technique requires tabulation of an entirely new set of bin frequencies. It is important to note that "matching" and "falling in the same bin" are not precisely equivalent. Alleles which fall in adjacent bins may match, while alleles falling at opposite ends of the same bin may not necessarily meet the matching criterion.

4.5 Estimating the Bin Frequencies

In order to estimate the probability of a randomly chosen individual having a DNA profile which matches that of the suspect (i.e. $P(M|V)$), it is first necessary to estimate the individual allele relative frequencies. The estimation process has two stages. The first stage, as mentioned previously, consists of estimating the individual allele relative frequencies by the corresponding bin relative frequencies. The second stage involves estimation of the bin relative frequencies using a simple random sample from the population. For each bin number j, $j = 1, \ldots, b$, a point estimator ($\hat{p}_j$) of the bin relative frequency is obtained. These estimators are shown to be unbiased and to have small variance for large N. In addition the covariance between two bin relative frequency estimates is derived since, as will be shown, this is an indication of the bias in the estimate of heterozygote relative frequency. Confidence intervals for the bin relative frequency estimates are also derived since the length of these intervals gives an indication of the validity of the estimators. Further it has been suggested by Lander (1989) that the upper 99% confidence limit be used in place of the point estimate in an attempt to be conservative.

4.5.1 Point Estimator for p_j

A simple random sample of N individuals is taken from the population. Each subject is examined at locus i to see which alleles are present on the two chromosomes with

that locus. For each of the 2N chromosomes in the sample, the observation recorded is the index number of the bin into which the allele at locus i falls.

Suppose there are b bins in total. Suppose further that allele A_j falls into bin j. Let X_j denote the total number of observations for bin j. Under the current assumptions $(X_1, \ldots, X_b)$ has a multinomial distribution with parameters p_j, $j = 1, \ldots, b$, and 2N. Thus,

$$\begin{aligned} \mathcal{E}(X_j) &= 2Np_j \qquad \forall\, j \in (1, \ldots, b) \\ Var(X_j) &= 2Np_j(1-p_j) \quad \forall\, j \in (1, \ldots, b) \end{aligned}$$

The maximum likelihood point estimator for p_j is :

$$\hat{p}_j = \frac{X_j}{2N} \quad \forall\, j \in (1, \ldots, b). \tag{4.1}$$

For each bin, the estimated probability obtained from 4.1 is assigned to all of the allelic types in the bin. To be conservative, if an allele in the suspect/specimen profile lies on the boundary of two adjacent bins, or if it could belong to another bin due to measurement error, it is assigned to the higher frequency bin. Standard multinomial theory gives the following results :

$$\begin{aligned} \mathcal{E}(\hat{p}_j) &= p_j \qquad \forall\, j \in (1, \ldots, b), \quad (\hat{p}_j \textit{ unbiased}) \\ Var(\hat{p}_j) &= \tfrac{p_j(1-p_j)}{2N} \quad \forall\, j \in (1, \ldots, b), \\ Cov(\hat{p}_j, \hat{p}_k) &= \tfrac{-p_j p_j}{2N} \quad \forall\, j \in (1, \ldots, b), \quad j \neq k. \end{aligned}$$

Note that the variance is an indication of the bias in the estimated homozygote frequencies. This is easily seen since (under HWE),

$$\begin{aligned} Var(\hat{p}_j) &= \mathcal{E}(\hat{p}_j^2) - [\mathcal{E}(p_j)]^2 \quad \forall\, j \in (1, \ldots, b) \\ &= \mathcal{E}(\hat{p}_j^2) - p_j^2 \\ &= \mathcal{E}\{\textit{estimated homozygote frequency}\} \\ &\quad -\{\textit{true homozygote frequency}\} \end{aligned}$$

Similarly, the covariance is an indicator of the bias in estimated heterozygote

frequencies since (under HWE),

$$\begin{aligned} Cov(\hat{p}_j, \hat{p}_k) &= \mathcal{E}(p_j p_k) - \mathcal{E}(p_j)\mathcal{E}(p_k) \quad \forall\, j \in (1,\ldots,b) \\ &= \tfrac{1}{2}\mathcal{E}(2\hat{p}_j\hat{p}_k) - 2p_j p_k \\ &= \tfrac{1}{2}\mathcal{E}\{estimated\ heterozygote\ frequency\} \\ &\quad -\{true\ heterozygote\ frequency\} \end{aligned}$$

Hence the estimated heterozygote and homozygote frequencies are asymptotically unbiased.

4.5.2 Confidence Intervals for p_j

Method I

Confidence intervals for p_j, j=1,...,b, are usually determined using the method of Goodman (1965). Denoting the lower and upper endpoints of the confidence interval by L_j and U_j respectively, Goodman's method yields,

$$L_j = \frac{B + 2X_j - \sqrt{B(B + 4X_j(\frac{2N-X_j}{2N}))}}{2(2N+B)} \quad \forall\, j \in (1,\ldots,b) \tag{4.2}$$

$$R_j = \frac{B + 2X_j + \sqrt{B(B + 4X_j(\frac{2N-X_j}{2N}))}}{2(2N+B)} \quad \forall\, j \in (1,\ldots,b) \tag{4.3}$$

where $B = \chi^2_1(\frac{\alpha}{b})$ is the $100(1-\frac{\alpha}{b})$ percentile point of a chi-square distribution with 1 degree of freedom. These intervals can be derived in the following way. A set of b confidence intervals are required (one for each p_j, j=1,...,b) such that the simultaneous coverage probability of all b intervals is at least $1-\alpha$. Let,

C_j = confidence interval for p_j, *for* $j \in (1,\ldots,b)$.

Now let each C_j, j=1,...,b , have coverage probability $(1-\frac{\alpha}{b})$.Then it follows from a simple Bonferroni inequality that,

$$P(\bigcap_{j=1}^{b} c_j) \;\geq\; 1 - \sum_{j=1}^{b} P(\bar{c}_j) \;=\; 1-\alpha,$$

where c_j denotes the event that p_j falls within confidence interval C_j. Now,

$$\frac{\sqrt{2N}(\hat{p}_j - p_j)}{\sqrt{p_j(1-p_j)}} \xrightarrow{\mathcal{D}} Z \quad as \quad N \to \infty \quad where \quad Z \sim N(0,1)$$

which implies that as $N \longrightarrow \infty$,

$$P(\frac{2N(\hat{p}_j - p_j)^2}{p_j(1-p_j)} > \chi_1^2(\frac{\alpha}{b})) = \frac{\alpha}{b}.$$

Therefore, set L_j and R_j equal to the two solutions of the quadratic equation in p_j,

$$\frac{2N(\hat{p}_j - p_j)^2}{p_j(1-p_j)} = \chi_1^2(\frac{\alpha}{b}) \tag{4.4}$$

to obtain an interval such that

$$P(L_j < p_j < R_j) = 1 - \frac{\alpha}{b}.$$

Solving 4.4 yields,

$$p_j = \frac{2\hat{p}_j + \frac{B}{2N} \pm \sqrt{B(\frac{2\hat{p}_j}{N} + \frac{B}{4N^2} - \frac{2\hat{p}_j^2}{N})}}{2(\frac{B}{2N}+1)}, \quad where \quad B = \chi_1^2(\frac{\alpha}{b})$$

which upon the substitution $\hat{p}_j = \frac{X_j}{2N}$, yields after simplification,

$$p_j = \frac{B + 2X_j \pm \sqrt{B(B + 4X_j\frac{2N-X_j}{2N})}}{2(2N+B)} \qquad \forall\, j \in (1,\ldots,b), \;\; B = \chi_1^2(\frac{\alpha}{b}).$$

These are the interval endpoints given by equations 4.2 and 4.3.

Method II

Goodman's intervals are the most frequently used although a simpler, and perhaps more familiar, set of $(1-\frac{\alpha}{b})\%$ confidence intervals are given by,

$$\hat{p}_j \pm Z_{\frac{\alpha}{2b}}\sqrt{\frac{\hat{p}_j(1-\hat{p}_j)}{2N}} \quad j \in (1,\ldots,b)$$

where $Z_{(\frac{\alpha}{2b})}$ is the $100(1-\frac{\alpha}{2b})$ percentile point of the standard normal distribution. These intervals are symmetric about the point estimates (whereas Goodman's intervals are not) but they may yield endpoints which are less than zero or greater than one.

To illustrate the differences in the methods just described consider the following example.

Table 1 Confidence Intervals for p_j, $j = 1, \ldots, 5$, N=25

Bin	Range	X_j	Point Estimate	95% LCL Method I	95%UCL Method I	95% LCL Method II	95% UCL Method II
1	0 - 900	12	0.24	0.1211	0.4198	0.0844	0.3956
2	901-990	8	0.16	0.0682	0.3315	0.0265	0.2935
3	991-1050	17	0.34	0.1955	0.5220	0.1674	0.5126
4	1051-1200	6	0.12	0.0447	0.2843	0.0016	0.2381
5	1201-1300	7	0.14	0.0561	0.3082	0.0136	0.2661

For this example there are N=25 subjects hence 2N=50 independent observations, all of which are made at the same (hypothetical) locus. The range of possible allele sizes has been divided into 5 bins. The range column indicates the boundary sizes of the bins in base pairs. It is standard practice to pool the low frequency bins to obtain a minimum of five occurrences in each bin. This is to avoid having bins with exceptionally small bin frequency estimates, since certain rare alleles may not have appeared in the sample. The bin ranges and counts are completely hypothetical. In practice there would be many more bins and the observed number in each bin would usually be lower.

4.6 Relaxing the HWE Assumption

The assumption of HWE is crucial to all calculations performed thus far, and the issue of whether this is reasonable is contentious. It is therefore worthwhile to consider

the consequences of *falsely* assuming HWE. It will be shown that although the *bin frequency* estimators are unchanged (and still unbiased), the estimators of homozygote/heterozygote frequencies will be altered. This is indicated through the variance and covariance of the bin frequency estimators respectively. That is,

$$\begin{aligned} \mathcal{E}(\hat{p}_j^2) &= var(\hat{p}_j) + \mathcal{E}(\hat{p}_j)^2 = var(\hat{p}_j) + p_j^2 \quad forall\ j,k = 1,\ldots,b \\ \mathcal{E}(2\hat{p}_j\hat{p}_k) &= 2cov(\hat{p}_j,\hat{p}_k) + 2p_jp_k \quad \forall\, j,k = 1,\ldots,b,\ j \neq k. \end{aligned}$$

It is thus important to know how the variance and covariance are affected by an inaccurate assumption of HWE.

The HWE assumption allows us to consider all of the 2N trials to be independent. In particular it ensures that two trials on the same subject, one at each chromosome with the locus in question, will be independent. Dropping the HWE assumption means that there is only independence between trials on *different* individuals. The X_j's no longer come from a multinomial distribution. Mendell and Simon (1984) show how a departure from HWE affects the variance (and covariance) of the estimates. A more general method for calculating the variance is introduced – one which does not depend on $(X_1,\ldots,X_b)$ having a multinomial distribution.

The probability p_j may be expressed as follows,

$$p_j = p_{jj} + \frac{1}{2}p_{jh} \quad \forall\, j,h \in (1,\ldots,b), \quad j \neq h \tag{4.5}$$

where,

p_{jj} = probability that a randomly chosen individual is homozygous for an allele in bin j (note that $p_{jj} = p_j^2$ under HWE), and,

p_{jh} = probability that a randomly chosen individual is heterozygous for an allele in bin j and an allele from *any* other bin.

Now let,

X_{jj} = number of individuals in the sample who are homozygous for an allele in bin j, and,

X_{jh} = number of individuals in the sample who are heterozygous for an allele in bin j and an allele from *any* other bin.

Since there is now only one observation being recorded for each subject there is independence between all (N) trials. The random variables X_{jj} and X_{jh} come from a multinomial distribution hence the maximum likelihood estimators for p_{jj} and p_{jh} are given by,

$$\hat{p}_{jj} = \frac{X_{jj}}{N}, \quad \hat{p}_{jh} = \frac{X_{jh}}{N}, \quad j,h \in (1,\ldots,b), \ \ j \neq h$$

which gives the estimator,

$$\hat{p}_j \ = \ \hat{p}_{jj} + \frac{1}{2}\hat{p}_{jh} \ = \frac{X_{jj}}{N} + \frac{X_{jh}}{2N}, \quad j,h \in (1,\ldots,b), \ \ j \neq h.$$

This is equivalent to counting all the individual occurrences of alleles in bin j as before, hence the estimator is unchanged. Standard multinomial theory yields the results,

$$\begin{aligned} \mathcal{E}(\hat{p}_j) &= p_{jj} + \tfrac{1}{2}p_{jh} = p_j \quad (\hat{p}_j \textit{ unbiased}) \\ Cov(\hat{p}_{jh},\hat{p}_{km}) &= \tfrac{-p_{jh}p_{km}}{N}, \quad j,h,k,m \in (1,\ldots,b) \end{aligned}$$

where $j \neq k$, and $h = k$, $m = j$ cannot occur simultaneously.

The following theorem indicates that although the bin frequency estimates remain unbiased, the variance of these estimates is increased.

Theorem 1

$$Var(\hat{p}_j) = \underbrace{\frac{p_j(1-p_j)}{2N}}_{\substack{\textit{variance} \\ \textit{under} \\ \textit{HWE}}} + \underbrace{\frac{p_{jj}-p_j^2}{2N}}_{\substack{\textit{function of the difference} \\ \textit{between the true probability} \\ \textit{of homozygosity and} \\ \textit{that predicted under} \\ \textit{HWE}}} \qquad \forall\, j \in (1,\ldots,b)$$

Proof

$$\begin{aligned} Var(\hat{p}_j) &= Var(\hat{p}_{jj} + \frac{1}{2}\hat{p}_{jh}) \\ &= Var(\hat{p}_{jj}) + \frac{1}{4}Var(\hat{p}_{jh}) + 2Cov(\hat{p}_{jj}, \frac{\hat{p}_{jh}}{2}) \\ &= \frac{p_{jj}(1-p_{jj})}{N} + \frac{p_{jh}(1-p_{jh})}{4N} - \frac{p_{jj}p_{jh}}{N} \end{aligned} \tag{4.6}$$

Now $p_j = p_{jj} + \frac{1}{2}p_{jh}$ implies $p_{jh} = 2(p_j - p_{jj})$ and hence straightforward substitution into 4.6 yields the result.

Theorem 2 $\forall\ j,k \in (1,\ldots,b)$

$$Cov(\hat{p}_j, \hat{p}_k) = \underbrace{\frac{-p_j p_k}{2N}}_{\substack{covariance \\ under \\ HWE}} + \underbrace{\frac{p_{jk} - 2p_j p_k}{4N}}_{\substack{function\ of\ the\ difference \\ between\ the\ true\ probability \\ of\ heterozygosity\ and \\ that\ predicted\ under \\ HWE}}$$

Proof

$$\begin{aligned} Cov(\hat{p}_j, \hat{p}_k) &= Cov(\hat{p}_{jj} + \frac{1}{2}\hat{p}_{jh},\ \hat{p}_{kk} + \frac{1}{2}\hat{p}_{km}) \\ &\quad where\ \ j,h,k,m \in (1,\ldots,b),\ j \neq h,\ k \neq m,\ j \neq k \\ &= \mathcal{E}[(\hat{p}_{jj} + \frac{1}{2}\hat{p}_{jh})(\hat{p}_{kk} + \frac{1}{2}\hat{p}_{km})] - \mathcal{E}(\hat{p}_{jj} + \frac{1}{2}\hat{p}_{jh})\mathcal{E}(\hat{p}_{kk} + \frac{1}{2}\hat{p}_{km}) \\ &= \mathcal{E}(\hat{p}_{jj}\hat{p}_{kk}) + \frac{1}{2}\mathcal{E}(\hat{p}_{jj}\hat{p}_{km}) + \frac{1}{2}\mathcal{E}(\hat{p}_{jh}\hat{p}_{kk}) + \frac{1}{4}\mathcal{E}(\hat{p}_{jh}\hat{p}_{km}) \\ &\quad -(\mathcal{E}(\hat{p}_{jj}) + \frac{1}{2}\mathcal{E}(\hat{p}_{jh}))(\mathcal{E}(\hat{p}_{kk}) + \frac{1}{2}\mathcal{E}(\hat{p}_{km})) \end{aligned}$$

$$= Cov(\hat{p}_{jj}, \hat{p}_{kk}) + \frac{1}{2}Cov(\hat{p}_{jj}, \hat{p}_{km}) + \frac{1}{2}Cov(\hat{p}_{jh}, \hat{p}_{kk}) + \frac{1}{4}Cov(\hat{p}_{jh}, \hat{p}_{km}) \quad (4.7)$$

$$where\ j \neq h,\ k \neq m,\ j \neq k$$

Expanding 4.7 further yields,

$$\begin{aligned} Cov(\hat{p}_j, \hat{p}_k) &= Cov(\hat{p}_{jj}, \hat{p}_{kk}) + \frac{1}{2}Cov(\hat{p}_{jj}, \hat{p}_{km}) + \frac{1}{2}Cov(\hat{p}_{jh}, \hat{p}_{kk}) \\ &+ \frac{1}{4}Cov(\hat{p}_{jh}, \hat{p}_{km}) + \frac{1}{2}Cov(\hat{p}_{jj}, \hat{p}_{kj}) + \frac{1}{2}Cov(\hat{p}_{jk}, \hat{p}_{kk}) + \\ &\frac{1}{4}Cov(\hat{p}_{jk}, \hat{p}_{km}) + \frac{1}{4}Var(\hat{p}_{jk}) + \frac{1}{4}Cov(\hat{p}_{jh}, \hat{p}_{kj}) \end{aligned} \quad (4.8)$$

$$\forall\ j,h,k,m \in (1,\ldots,b),\quad j \neq k,\ h \neq j,\ h \neq k,\ m \neq k,\ m \neq j.$$

Now recall that,

$$Cov(\hat{p}_{jh}, \hat{p}_{km}) = -\frac{p_{jh}p_{km}}{N}\ ,$$

where $j \neq k$ and $h = k, m = j$ cannot occur simultaneously. Thus 4.8 becomes,

$$\begin{aligned} Cov(\hat{p}_j, \hat{p}_k) &= -\frac{1}{N}[p_{jj}p_{kk} + \frac{1}{2}p_{jj}p_{km} + \frac{1}{2}p_{jj}p_{kj} + \frac{1}{2}p_{jh}p_{kk} + \frac{1}{2}p_{jk}p_{kk} + \\ &\frac{1}{4}p_{jh}p_{km} + \frac{1}{4}p_{jk}p_{km} + \frac{1}{4}p_{jh}p_{kj} - \frac{1}{4}p_{jk}(1 - p_{jk})], \end{aligned}$$

$$\forall\ j,h,k,m \in (1,\ldots,b),\quad j \neq k,\ h \neq j,\ h \neq k,\ m \neq k,\ m \neq j$$

$$\begin{aligned} &= -\frac{1}{N}[p_{jj}(p_{kk} + \frac{1}{2}p_{km} + \frac{1}{2}p_{kj}) + \frac{1}{2}p_{jh}(p_{kk} + \frac{1}{2}p_{km} + \frac{1}{2}p_{kj}) + \\ &\frac{1}{2}p_{kj}(p_{kk} + \frac{1}{2}p_{km} + \frac{1}{2}p_{jk}) - \frac{1}{4}p_{jk}] \end{aligned} \quad (4.9)$$

$$\forall\ j,h,k,m \in (1,\ldots,b),\quad j \neq k,\ h \neq j,\ h \neq k,\ m \neq k,\ m \neq j$$

Further,

$$\begin{aligned} p_j &= p_{jj} + \frac{1}{2}p_{jh}, \quad \forall\, h \neq j \\ &= p_{jj} + \frac{1}{2}p_{jh} + \frac{1}{2}p_{jk}, \quad \forall\, h \neq j,\ h \neq k \\ p_k &= p_{kk} + \frac{1}{2}p_{km}, \quad \forall\, m \neq k \\ &= p_{kk} + \frac{1}{2}p_{km} + \frac{1}{2}p_{kj} \quad \forall\, m \neq k,\ m \neq j \end{aligned} \tag{4.10}$$

Substitution of 4.10 into 4.9 yields,

$$\begin{aligned} Cov(\hat{p}_j, \hat{p}_k) &= -\frac{1}{N}[p_{jj}p_k + \frac{1}{2}p_{jh}p_k + \frac{1}{2}p_{jk}p_k - \frac{1}{4}p_{jk}] \\ &= -\frac{1}{N}[p_j p_k - \frac{1}{4}p_{jk}] \end{aligned} \tag{4.11}$$

Substitution of $-p_j p_k = \frac{-p_j p_k}{2} - \frac{2p_j p_k}{4}$ into 4.11 yields the desired result.

Clearly the penalty for falsely assuming HWE depends on the extent to which HWE conditions are violated.

4.7 Remarks

Although match-binning (primarily the fixed bin method) is the most commonly used approach to DNA profile analysis, it has many drawbacks. Not the least of these is the necessity of declaring a definitive match or exclusion. Suppose for example that two DNA profiles are identical except at one locus where the allele measurements do not meet the matching criterion. Regardless of the number of loci that were used to construct the profile it would be necessary to declare an exclusion. It is

possible however that the one nonmatch could have been due to DNA degradation or other technical considerations (see Chapter 2). In addition, a match between two measurements which barely meet the matching criterion is given the same weight as a match between two that are virtually indistinguishable. Clearly some information is being lost. It would seem desirable to have an alternative approach in which it would not be necessary to declare a match or exclusion. Instead the available DNA evidence would simply be expressed in a quantitative manner. Such an approach is presented in the following chapter.

Chapter 5

Bayesian Approaches

Each method outlined in this chapter is based on a distinct set of assumptions, although the same Bayesian procedure is followed. Each section will outline the assumptions, the procedure for estimating allele relative frequencies, and the resulting likelihood ratio. A final section on evaluating the goodness of fit of the models is included which applies to any of the proposed procedures.

5.1 A Density Estimation Approach

Although the problem of declaring a definitive match or exclusion is avoided by the Bayesian approach, it is still necessary to estimate the allele distribution in order to estimate R. Instead of discretizing the problem as done in the match-binning approach, Berry (1991) assumes a continuous allele distribution and applies a density estimation technique.

5.1.1 Model and Assumptions

Initially only a single measurement will be considered (i.e. the DNA of the suspect and specimen is examined at one chromosome locus only). The approach will be extended to consider two measurements at the same locus. Let x and y be measurements of alleles with true lengths A and B, in the suspect and specimen respectively, at a

particular locus. Although Berry does not explicitly specify a relationship between true allele lengths and measurements, it is instructive to consider a model of the form,

$$x = A + \varepsilon_A$$

$$y = B + \varepsilon_B.$$

It will be assumed that,

(i) the population is in LE and HWE with random mating,

(ii) $\mathcal{E}(x) = A$ and $\mathcal{E}(y) = B$,

(iii) the measurement errors ε_A and ε_B are independent both within a locus and between different loci, and,

(iv) the measurements are lognormally distributed with constant standard deviation c. That is,

$$\log(x) \sim \mathcal{N}(\nu, c^2)$$

$$\log(y) \sim \mathcal{N}(\mu, c^2).$$

The values ν and μ *reflect* the true values of the allele sizes in the suspect and specimen respectively. If the suspect is guilty then $\nu = \mu$.

If the true allele sizes A and B are not equal then assumption (iii) is justifiable since x and y are then measurements of distinct alleles from DNA examined on two seperate gels. On the other hand, if the suspect is guilty then x and y are two measurements of the same allele, and hence the measurement errors are linked *through the common allele.* Conditional on the true allele size however, they are independent. That is,

$$p(x, y|I) = p(x|I)p(y|I) \tag{5.1}$$

and,

$$p(x,y|\mu,G) \;=\; p(x,y|\mu,\mu=\nu) \;=\; p(x|\mu,\mu=\nu)p(y|\mu,\mu=\nu). \qquad (5.2)$$

A set of duplicate measurements analyzed by Berry support his lognormality assumption. They also support the assumption of a normal distribution with standard deviation proportional to the allele size (the latter assumption was detailed in Chapter 4), but the lognormality assumption is chosen to simplify calculations.

5.1.2 Estimating the Allele Distribution

To obtain an estimate for the likelihood ratio R one must first obtain an estimate for the posterior distribution of the allele sizes. Berry approaches the problem as follows.

Let $Z = (z_1, \ldots, z_n)$ be the allele measurements obtained at a specific locus taken from a simple random sample of $\frac{n}{2}$ individuals from an appropriate reference population (issues surrounding the choice of this reference population will be discussed in Chapter 6). Under assumption (iv), $\log(z_i) \sim \mathcal{N}(\mu_i, c^2)$ for given $\mu_1, \ldots, \mu_n$. Let μ_i, $i = 1, \ldots, n$ be independent and identically distributed (i.i.d.) random variables, and let $H(\cdot|Z)$ denote the (posterior) conditional density of the population of actual allele lengths given the sample of reference population measurements Z. An estimate of $H(\cdot|Z)$ must be obtained. Berry suggests four possibilities using a smoothing approach with normal kernels. The first estimator attempts only to take into account the effect of measurement error.

By assumption $\log(z_i) \sim \mathcal{N}(\mu_i, c^2)$, i = 1,...,n, and hence for each i a reasonable estimate for μ_i is $\log(z_i)$. The z_i are measured with error, however, and hence it would be inappropriate to use the empirical distribution of the $\log(z_i)$, i=1,...,n, as an estimate of $H(\cdot|Z)$. Instead the contribution of each z_i is taken to be a normal distribution centered about $\log(z_i)$ and these are averaged across all i, i=1,...,n.

Define,

$$\hat{H}_1(\cdot|Z) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{N}(\log(z_i), c^2).$$

This estimator however, does not adjust for the sampling variability. Certain alleles may be underrepresented in the sample (or may not be present at all) and this could bias the likelihood ratio in favour of guilt.

As an attempt to compensate for the sampling variability Berry proposed the following estimator,

$$\hat{H}_2(\cdot|Z) = \frac{n^*}{n+n^*}U(N_1, N_2) + \frac{1}{n+n^*}\sum_{i=1}^{n}\mathcal{N}(\log(z_i), c^2)$$

where $U(N_1, N_2)$ is a uniform density on $[N_1, N_2]$. This range is chosen so as to include all possible values of $\log(z_i)$. Clearly $\hat{H}_2(\cdot|Z)$ is the same type of estimator as $\hat{H}_1(\cdot|Z)$ except that it includes a uniform base to ensure that the rare allele frequencies will not be underestimated. The value of n^* may be chosen according to what it is felt the minimum allele frequency should be.

An alternative method for dealing with sampling variability proposed by Berry is of the following form,

$$\hat{H}_3(\cdot|Z) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{N}(\log(z_i), (bc)^2), \quad where\ b \geq 1.$$

This estimator is of the same form as $\hat{H}_1(\cdot|Z)$ with the smoothing parameter b increased. This allocates more probability to the lower frequency alleles (the "tails" of the allele distribution). Since $\hat{H}_3(\cdot|Z)$ reduces to $\hat{H}_1(\cdot|Z)$ when b=1, setting b=1 accounts for measurement error and setting $b > 1$ accounts for sampling variability as well. Berry suggests values of b in the range $1 \leq b \leq 5$.

As a fourth possibility, Berry combines all the above estimators,

$$\hat{H}_4(\cdot|Z) = \frac{n^*}{n+n^*}U(N_1, N_2) + \frac{1}{n+n^*}\sum_{i=1}^{n}\mathcal{N}(\log(z_i), (bc)^2).$$

The use of smoothers in this density estimation process has been criticized on the grounds that it is not conservative. Although the probability of rare alleles is being boosted to adjust for sampling error, this is done at the expense of the higher probability alleles. That is, the probability estimates of the more common alleles are lower as a result of the smoothing. Since it is more likely that the profile of the suspect will contain these more common alleles, the procedure puts the defence at a disadvantage in this sense. An alternative smoothing technique, suggested to Berry by Evett (see Berry, 1991), involves smoothing that increases the lower frequency estimates but does not decrease any of the higher frequencies. This of course would result in a total density greater than one, but would definitely be conservative.

A suggestion made by Chernoff in his comments on the Berry (1991) article, is to estimate the probability of alleles which do not appear in the sample using the method of Robbins (1968). According to Robbins the *total* probability which should be allocated to unobserved events (alleles in this case) in an experiment with n possible outcomes may be estimated by the number of singleton outcomes in an experiment with n+1 trials. That is, the number of times there is only one observed event in a category (in the forensic setting a category refers to a bin, and an event refers to observing an allele measurement in that bin). It can be shown that this estimator is unbiased. Evett suggests that this estimator could be used to determine the size of uniform base to use in Berry's $\hat{H}_2(\cdot|Z)$.

5.1.3 Estimating the Likelihood Ratio

The chosen estimator of $H(\cdot|Z)$ may now be used to estimate the likelihood ratio R in the following way. (Since all probabilities are calculated conditional on the additional evidence E, this will be suppressed from the calculations.)

If the assumption is made that the event G is independent of the reference sample

Z (a reasonable assumption), the likelihood ratio R can be rewritten as follows,

$$\begin{aligned} R &= \frac{p(X|G)}{p(X|I)} = \frac{p(x,y,Z|G)}{p(x,y,Z|I)} \\ &= \frac{p(x,y,Z,G)}{p(G)}\frac{p(I)}{p(x,y,Z,I)} = \frac{p(x,y,Z,G)}{p(G)p(Z)}\frac{p(I)p(Z)}{p(x,y,Z,I)} \\ &= \frac{p(x,y,Z,G)}{p(G,Z)}\frac{p(I,Z)}{p(x,y,Z,I)} = \frac{p(x,y|G,Z)}{p(x,y|I,Z)} \\ &= \frac{p(x,y|\nu=\mu,Z)}{p(x,y|Z)}. \end{aligned}$$

Consider the numerator first,

$$p(x,y|\nu=\mu,Z) = \int p(x,y|\nu,\mu,\nu=\mu)dH(\mu|Z)$$

which by 5.2 becomes,

$$\begin{aligned} p(x,y|\nu=\mu,Z) &= \int p(x|\nu,\mu,\nu=\mu)p(y|\nu,\mu,\nu=\mu)dH(\mu|Z) \\ &= \int p(x|\mu,\nu=\mu)p(y|\mu,\mu=\nu)dH(\mu|Z) \\ &= \int f_{\log(x)}(\log(x))\frac{\partial(\log(x))}{\partial x}f_{\log(y)}(\log(y))\frac{\partial\log(y)}{\partial y}dH(\mu|Z) \\ &= \int \frac{1}{xy}\frac{1}{2\pi c^2}exp\{[-\frac{1}{2c^2}[(\log(x)-\mu)^2+(\log(y)-\mu)^2]\}dH(\mu|Z) \end{aligned}$$

where $f_{\log(x)}(\cdot)$ and $f_{\log(y)}(\cdot)$ denote the common marginal densities of $\log(x)$ and $\log(y)$, letting x and y refer to both the random variables and the observations for simplicity. Similarly, using 5.1 the denominator may be written as,

$$p(x,y|Z) = p(x|Z)p(y|Z) = \{\int \frac{1}{x}p(x|\mu)dH(\mu|Z)\}\{\int \frac{1}{y}p(y|\nu)dH(\nu|Z)\}.$$

Substituting an estimator for $H(\cdot|Z)$ into the expressions for numerator and denominator yields an estimator of R. Berry demonstrates that when $\hat{H}_3(\cdot|Z)$ is chosen as the estimator of the allele distribution then R reduces to a relatively simple form (see Berry, 1991, p.183).

5.1.4 Extension to Two Independent Measurements at a Single Locus

Let x_1 and x_2 be the measurements of two alleles at a particular locus in the suspect and similarly let y_1 and y_2 be the measurements for the specimen. By assumption (iii) two measurements performed on the same person are independent (e.g. x_1 and x_2 independent). Realistically this will not often be the case, and factors which affect the measurement error of one allele will affect the other allele at that same locus. That is, the measurement errors will tend to be positively correlated. For example, if one allele size is overestimated this may indicate that all the measurements are being overestimated by the same amount (a phenomenon known as band shifting). In section 5.2 the case of two correlated measurements will be considered.

During the DNA analysis it is not possible to distinguish between the maternal and paternal chromosomes and hence even if the suspect is guilty it is not known whether x_1 and y_1 are two measurements of the same allele or whether x_1 actually belongs with y_2. Let R_{ij} denote the likelihood ratio given the data x_i and y_j. In this case there are two possible pairings : x_1 with y_1 and x_2 with y_2, or x_1 with y_2 and x_2 with y_1. Each of these pairings is equally likely and hence the overall likelihood ratio (denote it R) for the locus is,

$$R = \frac{\frac{1}{2}p(x_1,y_1|G,Z)p(x_2,y_2|G,Z) + \frac{1}{2}p(x_1,y_2|G,Z)p(x_2,y_1|G,Z)}{\frac{1}{2}p(x_1,y_1|I,Z)p(x_2,y_2|I,Z) + \frac{1}{2}p(x_1,y_2|I,Z)p(x_2,y_1|I,Z)}.$$

However, in order to use the likelihood ratios previously defined ($R_{ij}, i = 1,2, j = 1,2$) Berry defines the overall likelihood ratio R as,

$$\begin{aligned} R &= \tfrac{1}{2}R_{11}R_{22} + \tfrac{1}{2}R_{12}R_{21} \\ &= \tfrac{1}{2}\frac{p(x_1,y_1|G,Z)p(x_2,y_2|G,Z)}{p(x_1,y_1|I,Z)p(x_2,y_2|I,Z)} + \tfrac{1}{2}\frac{p(x_1,y_2|G,Z)p(x_2,y_1|G,Z)}{p(x_1,y_2|I,Z)p(x_2,y_1|I,Z)}. \end{aligned}$$

If several single locus probes are used (as is usually the case), then if the measurements taken at distinct loci are independent the likelihood ratios for each locus

are simply calculated as above and are multiplied together.

5.1.5 Remarks

The Bayesian approach works quite naturally in the forensic setting, where errors in measurement *do* occur, and Berry's method has many advantages over that of match-binning. It does rely heavily however on the assumption that for a given individual, two measurements taken at the same locus are independent. This is not likely to be the case and in fact studies have shown (e.g. Berry, Evett and Pinchin, 1992) that these measurements may be highly correlated. The next section describes a method which attempts to take this correlation into account.

5.2 Taking Measurement Error Correlation Into Account

Empirical studies indicate that there may not be independence between measurements of alleles at the same locus. In fact the measurement errors may be highly correlated, in which case a new approach is needed. This method (Berry, Evett, and Pinchin, 1992) attempts to take this correlation into account by estimating the joint distribution of two measurements taken at the same locus.

5.2.1 Assumptions

Once again consider only a single locus and suppose that,

$$x_1 = A_1 + \varepsilon_{x_1}$$

$$x_2 = A_2 + \varepsilon_{x_2}$$

are the two measurements taken for the suspect at that locus, and similarly that,

$$y_1 = B_1 + \varepsilon_{y_1}$$

$$y_2 = B_2 + \varepsilon_{y_2}$$

are the corresponding measurements for the specimen.

The following assumptions will be made.

(i) There is linkage equilibrium in the population.

(ii) The standard deviation of a single measurement error is directly proportional to the allele size. That is, $sd(\varepsilon_{x_i}) = cA_i$, and $sd(\varepsilon_{y_i}) = cB_i$ for i=1,2.

(iii) Measurement errors are independent between different individuals but are not necessarily independent between two measurements taken at the same locus for one individual. That is, $(\varepsilon_{x_1}, \varepsilon_{x_2})$ is assumed to be independent of $(\varepsilon_{y_1}, \varepsilon_{y_2})$ but $corr(\varepsilon_{x_1}, \varepsilon_{x_2}) = corr(\varepsilon_{y_1}, \varepsilon_{y_2}) = \xi$ where ξ is not necessarily zero.

(iv) The joint distribution of two measurement errors (within a locus) is bivariate normal with zero mean. Equivalently,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, c^2 \begin{pmatrix} \mu_1^2 & \rho\mu_1\mu_2 \\ \rho\mu_1\mu_2 & \mu_2^2 \end{pmatrix}\right)$$

Similarly,

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, c^2 \begin{pmatrix} \nu_1^2 & \rho\nu_1\nu_2 \\ \rho\nu_1\nu_2 & \nu_2^2 \end{pmatrix}\right).$$

Assumptions (i) and (ii) were outlined in detail in Chapter 4 and the estimator of the constant of proportionality for assumption (ii) is $\hat{c} = .008$. Berry et al. estimate the measurement error correlation ξ in assumption (iii) from a sample of 218 individuals exhibiting double band patterns. For each individual, duplicate measurements were made of each of the two bands. For a particular individual let x_{11} and x_{12} denote the two measurements of one allele, and let x_{21} and x_{22} denote the two measurements of the other band. The differences $x_{11} - x_{12}$ and $x_{21} - x_{22}$ reflect the size of the measurement error. Now, for any two random variables v and w, the correlation between

them (corr(v,w)) may be estimated by the slope of the least squares regression line for the standardized variables. This is clear since, letting $sd(\cdot)$ denote the *estimated* standard deviation,

$$\widehat{corr}(v,w) = \frac{\Sigma(v-\bar{v})(w-\bar{w})}{sd(v)sd(w)}$$

while the slope ($\hat{\beta}$) of the least squares regression line for the standardized variables ($w = \hat{\alpha} + \hat{\beta}v$) is given by,

$$\hat{\beta} = \frac{\Sigma(\frac{v}{sd(v)} - \frac{\bar{v}}{sd(v)})(\frac{w}{sd(w)} - \frac{\bar{w}}{sd(w)})}{\sqrt{\Sigma(\frac{w}{sd(w)} - \frac{\bar{w}}{sd(w)})^2}} = \frac{\Sigma((v-\bar{v})(w-\bar{w})}{sd(v)sd(w)} = \widehat{corr}(v,w)$$

A scatterplot of the standardized differences,

$$(\frac{\sqrt{2}(x_{11}-x_{12})}{c(x_{11}+x_{12})}, \frac{\sqrt{2}(x_{21}-x_{22})}{c(x_{21}+x_{22})})$$

shows a large positive correlation which is estimated (by the slope of the least squares regression line) as $\hat{\xi} = 0.904$. This phenomenon of a large positive correlation between measurement errors is quite common in laboratory work. The plot is well approximated by a bivariate normal distribution which supports assumption (iii).

An important point is that assuming that there is correlation between measurement errors does *not* violate an assumption of HWE. The assumption of HWE merely states that the true population allele frequencies are independent and makes no statements about the measurement error. For this method the assumption of HWE is omitted simply because it is not necessary.

5.2.2 Estimating the Likelihood Ratio

Let $Z = \{(z_{11}, z_{21}), (z_{12}, z_{22}), \ldots, (z_{1n}, z_{2n})\}$ be a reference sample of fragment pairs taken at random from the population. If individual i is homozygous then $z_{1i} = z_{2i}$. This is completely analogous to the reference sample in the approach of Berry (of Section 5.1) except that since the two measurements for one individual are now

correlated they must be considered in pairs. Also let E once again denote all evidence other than the DNA evidence. The likelihood ratio is given by,

$$R = \frac{p(\tilde{x}, \tilde{y}|G, Z, E)}{p(\tilde{x}, \tilde{y}|I, Z, E)} \tag{5.3}$$

where the hypotheses of guilt and innocence are once again denoted by G and I respectively, and where,

$$\tilde{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

denotes the fragment lengths obtained from the suspect, and

$$\tilde{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

denotes the corresponding fragment lengths obtained from the specimen. For convenience the additional evidence E is suppressed from the notation since it only influences the posterior likelihood through a multiplicative factor. Further let,

$$\tilde{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \tilde{\nu} = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}$$

represent the corresponding mean vectors of $\tilde{x}$ and $\tilde{y}$ respectively. It is important to keep in mind that here $\tilde{\mu}$ and $\tilde{\nu}$ refer to the true allele lengths, whereas in Section 5.1 they were the means of a lognormal distribution which merely *reflected* the true allele lengths.

Let the joint density of x_1 and x_2 be denoted by $f(\tilde{x}|\tilde{\mu})$ where $f(\cdot|\cdot)$ is a bivariate normal density. That is,

$$f(\tilde{x}|\tilde{\mu}) = \frac{1}{|\Sigma|^{\frac{1}{2}}(2\pi)} \exp -\frac{1}{2}(\tilde{x} - \tilde{\mu})^T(\Sigma)^{-1}(\tilde{x} - \tilde{\mu})$$

where $\Sigma = c^2 \begin{pmatrix} \mu_1^2 & \xi\mu_1\mu_2 \\ \xi\mu_1\mu_2 & \mu_2^2 \end{pmatrix}$. Similarly let the joint density of y_1 and y_2 be denoted $f(\tilde{y}|\tilde{\nu})$.

Consider first the numerator of R, (see equation 5.3) where the likelihood is being evaluated under the hypothesis of guilt. Under this hypothesis $\tilde{\mu} = \tilde{\nu}$. Since it is assumed that measurement errors are independent between individuals, conditional on the true allele sizes $\tilde{\mu}$, the measurements $\tilde{x}$ and $\tilde{y}$ are independent (see Section 5.1.1). Using the independence of the reference sample and the allele measurements $\tilde{x}$ and $\tilde{y}$ the numerator of R becomes,

$$\begin{aligned} p(\tilde{x}, \tilde{y}|G, Z) &= \int\int p(\tilde{x}, \tilde{y}|\tilde{\mu}, \tilde{\nu}, G, Z) dH(\tilde{\mu}, \tilde{\nu}|G, Z) \\ &= \int\int p(\tilde{x}|\tilde{\mu}, G, Z) p(\tilde{y}|\tilde{\mu}, G, Z) dH(\tilde{\mu}|G, Z) \\ &= \int\int f(\tilde{x}|\tilde{\mu}) f(\tilde{y}|\tilde{\mu}) dH(\tilde{\mu}|Z). \end{aligned}$$

All that is required for evaluation of this expression is an estimate of the allele frequency distribution $H(\cdot|Z)$, which will be discussed in the next section. A similar approach may be used to deal with the denominator of R which is evaluated under the hypothesis of innocence. Under I, $\tilde{x}$ and $\tilde{y}$ are independent (see Section 5.1.1).

$$\begin{aligned} p(\tilde{x}, \tilde{y}|I, Z) &= p(\tilde{x}|I, Z) p(\tilde{y}|I, Z) \\ &= p(\tilde{x}|Z) p(\tilde{y}|Z) \\ &= \int\int p(\tilde{x}|\tilde{\mu}, Z) dH(\tilde{\mu}|Z) \int\int p(\tilde{y}|\tilde{\nu}, Z) dH(\tilde{\nu}|Z) \end{aligned}$$

Using the LE assumption, the values of R obtained for the various loci in the profiles may be multiplied together to obtain the overall likelihood ratio.

5.2.3 Estimating the Allele Distribution

The same methods are used as in Section 5.1, except that instead of smoothing with univariate normal kernels, bivariate normal kernels are used. This yields an estimator of the form,

$$\hat{H}(\cdot|Z) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{N}\left(\begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix}, (bc)^2 \begin{pmatrix} z_{1i}^2 & 0 \\ 0 & z_{2i}^2 \end{pmatrix} \right)$$

where b is the smoothing parameter. As previously described, b=1 accounts for measurement error while $b > 1$ adjusts for sampling variability as well. The correlation between z_{1i} and z_{2i} is taken to be zero although this is not usually the case. This is done as sampling variability is by far the dominant reason for performing the smoothing (the reference sample will by necessity be much smaller than desired). The correlation between z_{1i} and z_{2i}, caused by measurement error, is insignificant in comparison and is thus omitted to simplify calculations. A further attempt to adjust for sampling variability is made by including the fragment lengths of the suspect in the reference sample Z. This avoids the situation in which the fragment length of the suspect falls in a region where there were no observed fragment lengths in the reference sample. This estimator of the allele distribution may now be substituted for $H(\cdot|Z)$ in the numerator and denominator of R.

5.2.4 Single Band Profiles

Up to this point all methods have assumed that if the suspect/specimen exhibits a single band pattern then it is that of a true homozygote rather than a heterozygote with one band which for some reason was not visible. The likelihood ratio estimation procedure may be adjusted to take this phenomenon into account.

Suppose there is a situation in which the suspect has fragment length measurements (x_1, x_2) with $x_1 < x_2$, while the specimen has only one measurement y, which is close to x_2. Berry, Evett, and Pinchin (1992) give a complicated formula for calculating the likelihood ratio R in this situation. It may be derived as follows. First, the numerator of R is considered, which is evaluated under the hypothesis that the suspect is guilty. That is, it is necessary to compute the likelihood that the specimen actually had two bands (close to (x_1, x_2)), one of which did not appear. By the Law

of Total Probability the numerator of R becomes,

$$p(\tilde{x}, y|G, Z) = \int\int p(\tilde{x}, y|G, Z, \tilde{\mu}) dH(\tilde{\mu}|Z).$$

Using first the conditional independence under G of $\tilde{x}$ and y given the true allele lengths $\tilde{\mu}$, and second the assumption that the allele measurements are independent of the reference set (Z) and the hypothesis of guilt, yields,

$$\begin{aligned} p(\tilde{x}, y|G, Z) &= \int\int p(\tilde{x}|G, Z, \tilde{\mu}) p(y|G, Z, \tilde{\mu}) dH(\tilde{\mu}|Z) \\ &= \int\int f(\tilde{x}|\tilde{\mu}) p(y|\tilde{\mu}) dH(\tilde{\mu}|Z), \end{aligned} \tag{5.4}$$

where $f(\cdot|\cdot)$ refers specifically to a normal density function and $p(\cdot|\cdot)$ refers simply to the probability of an event. Let,

$m(\tilde{\mu})$ = the probability of observing a double band pattern when alleles of lengths μ_1 and μ_2 are measured.

Since y is close to x_2 it will be assumed that y is a measurement of an allele of length μ_2. With this notation 5.4 becomes,

$$p(\tilde{x}, y|G, Z) = \int\int f(\tilde{x}|\tilde{\mu})(1 - m(\tilde{\mu}) g(y|\mu_2) dH(\tilde{\mu}|Z),$$

where $g(y|\mu_2)$ is a univariate normal density with parameters μ_2 and $c^2\mu_2^2$. This requires estimation of $m(\tilde{\mu})$ which may prove a difficult task. As an alternative approach, consider the following. It is known that, under G, the other "missing" band must be approximately of length x_1 since the single measurement y was close to x_2. Let,

m = the probability that the smaller band (of length approximately x_1) would have shown given G.

In the extreme, if m=1 then the suspect could not have committed the crime since if he/she had then another band would have been visible, which it was not. The quantity

m may be used as an estimate of $m(\tilde{\mu})$. This estimate will be very accurate for true allele lengths μ_1 that are similar to x_1, which are the lengths of most importance. For those lengths which are not close to x_1, $f(\tilde{x}|\tilde{\mu})$ will be small, reducing the adverse effects of poorly estimating $m(\tilde{\mu})$. The numerator of the likelihood becomes,

$$p(\tilde{x}, \tilde{y}|G, Z) = (1-m)\int\int f(\tilde{x}|\tilde{\mu})g(y|\mu_2)dH(\tilde{\mu}|Z).$$

Although Berry, Evett, and Pinchin do not explicitly state that they are using the quantity m as an estimate of some more complicated function which depends on the true allele lengths ($m(\tilde{\mu})$), it is suspected that they also followed the above line of reasoning.

Now consider the denominator of R, computed under the hypothesis of innocence.

$$\begin{aligned} p(\tilde{x}, y|I, Z) &= p(\tilde{x}|I, Z)p(y|I, Z) \\ &= p(y|I, Z)\int\int f(\tilde{x}|\tilde{\mu})dH(\tilde{\mu}|Z). \end{aligned} \tag{5.5}$$

The double integral may be evaluated using the estimated allele density $H(\tilde{\mu}|Z)$ and with $f(\cdot|\cdot)$ as a normal density function. It remains to estimate $p(y|I, Z)$. Under the hypothesis of innocence the observed fragment lengths could have arisen from two different scenarios. Either,

(a) the specimen is measurably homozygous (i.e. has fragments which are so close so as to be unresolvable by current techniques), or,

(b) the specimen is heterozygous with a second band which did not appear.

Define the following events.

$A =$ the event that the two bands are measurably homozygous,

$B =$ the event that a band lighter than y appears in the profile.

In terms of the events A and B there are three disjoint possibilities that explain the occurrence of a single measurement y. Letting A^c denote the complement of event A, these three possibilities may be written as, A^cB^c, AB^c, and AB. Let,

$m^* = p(B|I, Z)$, and

$h = p(A|I, Z)$.

Using this notation and the fact that $A^cB^c \cup AB^c \cup AB = A^cB^c \cup A$ one obtains,

$$\begin{aligned} p(y|I, Z) &= p(y|I, Z, A^cB^c)p(A^cB^c|I, Z) + p(y|I, Z, A)p(A|I, Z) \\ &= p(y|I, Z, A^cB^c)p(A^c|I, Z)p(B^c|I, Z) + p(y|I, Z, A)p(A|I, Z) \\ &= (1-m^*)(1-h)p(y|I, Z, A^cB^c) + hp(y|I, Z, A). \end{aligned} \tag{5.6}$$

Now,

$$\begin{aligned} p(y|I, Z, A^cB^c) &= \int\int p(y|I, Z, A^cB^c, (\nu, \mu))p((\nu, \mu)|I, Z, A^cB^c)d\nu d\mu \\ &= \int\int_{\nu<\mu} p(y|I, Z, A^cB^c, (\nu, \mu))p((\nu, \mu)|I, Z, A^cB^c)d\nu d\mu + \\ &\quad \int\int_{\nu>\mu} p(y|I, Z, A^cB^c, (\nu, \mu))p((\nu, \mu)|I, Z, A^cB^c)d\nu d\mu \\ &= \int\int_{\nu<\mu} p(y|I, Z, \mu)p((\nu, \mu)|I, Z, A^cB^c)d\nu d\mu + \\ &\quad \int\int_{\nu>\mu} p(y|I, Z, \nu)p((\nu, \mu)|I, Z, A^cB^c)d\nu d\mu \\ &= \int p(y|I, Z, \mu)p(\mu|I, Z)d\mu. \end{aligned} \tag{5.7}$$

Also,

$$
\begin{aligned}
& p(y|I,Z,A) \\
&= p(y|\ I,Z,|\nu-\mu|<\varepsilon) \\
&= \int\int p(y|\ I,Z,(\nu,\mu),|\nu-\mu|<\varepsilon)p((\nu,\mu)|\ |\nu-\mu|<\varepsilon,I,Z)d\nu d\mu \\
&= \int\int p(y|\ I,Z,(\nu,\mu),|\nu-\mu|<\varepsilon)\frac{p((\nu,\mu),|\nu-\mu|<\varepsilon|I,Z)p(I,Z)}{p(|\nu-\mu|<\varepsilon|\ I,Z)p(I,Z)}d\nu d\mu \\
&\approx \int\int p(y|I,Z,(\mu,\mu))\frac{p((\mu,\mu)|I,Z)}{p(|\nu-\mu|<\varepsilon|\ I,Z)}d\nu d\mu \\
&\approx \int p(y|I,Z,\mu)\frac{(p(\mu|I,Z))^2}{\int p((\mu,\mu)|I,Z)d\mu} \\
&= \int p(y|I,Z,\mu)\frac{(p(\mu|I,Z))^2}{\int(p(\mu|I,Z))^2 d\mu} \qquad (5.8)
\end{aligned}
$$

Now substituting 5.7 and 5.8 into 5.6 yields,

$$
\begin{aligned}
p(y|I,Z) &= (1-m^*)(1-h)\int p(y|I,Z,\mu)p(\mu|I,Z)d\mu\ + \\
&\qquad h\int p(y|I,Z,\mu)\frac{(p(\mu|I,Z))^2}{\int(p(\mu|I,Z))^2 d\mu}. \qquad (5.9)
\end{aligned}
$$

Substitution of 5.9 into 5.5 yields Berry's result for the denominator of R.

5.2.5 Remarks

Although many properties of the DNA profiling data are incorporated into this model, a phenomenon known as flanking region polymorphism has not been addressed. This is considered in the following section.

5.3 Considering Flanking Region Polymorphism

In reality, when the DNA is cut with a restriction enzyme, the resulting fragments have not only an integral number of repeat units, but also a stretch of DNA on

either side of these repeats which is called the flanking region. For certain restriction enzymes only one flanking region size is possible. For others, however, there are many possible flanking region sizes which may result. Devlin, Risch, and Roeder (1991, 1992) incorporate this flanking region polymorphism into their model. They take a mixture model approach, and the EM algorithm for mixtures is used to obtain maximum likelihood estimates of the model parameters.

5.3.1 Assumptions

Suppose that $x_1 = a_1 + \varepsilon_{a_1}$ and $x_2 = a_2 + \varepsilon_{a_2}$ are the two measurements obtained from an individual at a particular locus, where a_i represents the true allele length. The following assumptions are made,

(i) *Random Mating* : It is again assumed that individuals mate at random with respect to their genotypes and hence that there is independence between alleles within loci. For example, a_1 is independent of a_2.

(ii) It is assumed that there is independence across loci which implies linkage equilibrium (LE).

(iii) Measurement errors are assumed to be correlated within a locus but uncorrelated between different loci and different individuals. Let ε_{a_1} and ε_{a_2} have correlation coefficient ξ. One procedure for estimating ξ was outlined in Section 5.2.1. An alternative estimation technique may be found in Devlin, Risch, and Roeder, 1992.

(iv) The individual measurement errors are normally distributed with mean zero and standard deviation proportional to the size of the allele. That is,

$$\varepsilon_{a_i} \sim \mathcal{N}(0, ca_i).$$

From repeated measurements of the same allele Devlin, Risch, and Roeder estimate c to be $5\text{x}10^{-3}$.

(v) The joint distribution of two measurement errors (for an individual at a single locus) is bivariate normal. Let $g_{i,j}(x_1, x_2)$ denote the joint density of the pair (x_1, x_2) given that x_1 and x_2 are measurements of alleles of size A_i and A_j respectively. It follows that $g_{i,j}(x_1, x_2)$ is a bivariate normal density with mean vector $(A_i, A_j)^T$ and covariance matrix,

$$\Sigma = \begin{pmatrix} c^2 A_i^2 & \xi A_i A_j \\ \xi A_i A_j & c^2 A_j^2 \end{pmatrix}$$

(vi) It is assumed to be impossible to obtain $x_1 > x_2$ if $A_1 < A_2$. If A_1 and A_2 are very similar then the correlation between x_1 and x_2 will be close to one hence the measurements will be very tightly clustered with $x_1 < x_2$ at virtually all times. Conversely, if A_1 and A_2 are not very similar then observing $x_1 > x_2$ is very unlikely and is not allotted any significant amount of probability in the joint normal distribution. Henceforth $x_1 \leq x_2$ will imply $A_1 \leq A_2$.

5.3.2 Model

Let $\tilde{x} = (x_1, x_2)$ denote two measurements obtained from an individual at a particular locus. In the multiple flanking region scenario it is possible that two alleles with different numbers of repeats may still have the same length. If there are b possible numbers of repeats and L possible flanking region sizes then there are at *most* bL possible allelic types. Let the unique allele lengths be denoted by $a_{(1)}, \ldots, a_{(A)}$, where $A \leq bL$. If coalescence is ignored for the moment then the probability density of a pair of measurements is approximately given by,

$$f(x_1, x_2) \approx \begin{cases} \Sigma_{i=1}^{M} p_i^2 g_i(x_1) & x_1 = x_2 \\ \Sigma_{j=1}^{M-1} \Sigma_{i=j+1}^{M} 2 p_i p_j g_{i,j}(x_1, x_2) & x_1 < x_2 \end{cases} \tag{5.10}$$

This is only an approximation because the extremely rare case when $A_1 = A_2$ but $x_1 \neq x_2$ is ignored. This density may easily be modified to take the possibility of coalescence into account. The modified probability density for a single band pattern may be written as,

$$f^*(z,z) = \Sigma_{i=1}^{A} p_i^2 g_i(z) + 2\Sigma_{i<j} p_i p_j \int_0^\infty g_{ij}(z-t, z+t)\delta(t,z)dt \qquad x_1 = x_2 = z,$$

where $\delta(t,z)$ denotes the probability that two measurements of size $z+t$ and $z-t$ would coalesce. The first summation in $f^*(z,z)$ is the probability that the measurement actually represents a true homozygote. The second summation represents the probability that there were actually two distinct measurements which coalesced. Hence the probability density for a pair of measurements taking coalescence into account may be very accurately approximated by,

$$f(x_1,x_2) \approx \begin{cases} \Sigma_{i=1}^{A} p_i^2 g_i(z) + 2\Sigma_{i<j} p_i p_j \int_0^\infty g_{ij}(z-t, z+t)\delta(t,z)dt & x_1 = x_2 = z \\ \Sigma_{j=1}^{M-1}\Sigma_{i=j+1}^{M} 2p_i p_j g_{ij}(x_1,x_2) & x_1 < x_2 \end{cases} \tag{5.11}$$

5.3.3 Estimating the Likelihood Ratio

Let H_I denote the null hypothesis that the suspect is innocent, and let H_G denote the alternative hypothesis that the suspect is guilty. Let the likelihood ratio be denoted R,

$$R = \frac{L(\tilde{x}, \tilde{y}|H_G)}{L(\tilde{x}, \tilde{y}|H_I)}$$

where $\tilde{x} = (x_1, x_2)$ and $\tilde{y} = (y_1, y_2)$ are the measurements obtained from the suspect and specimen respectively. Further let p_i denote the probability of observing an allele of length $a_{(i)}$ in the population. Consider the numerator first.

$$L(\tilde{x}, \tilde{y}|H_I) = f(\tilde{x})f(\tilde{y}), \tag{5.12}$$

where $f(\cdot)$ is defined as in 5.11. If coalescence is ignored, use $f(\cdot)$ as defined in 5.10.

Now consider the denominator of R. If coalescence is ignored, this may be written as,

$$L(\tilde{x},\tilde{y}|H_G) = \begin{cases} \Sigma_{i=1}^{A} p_i^2 g_i(x_1) g_i(y_1) & x_1 = x_2,\ y_1 = y_2 \\ \Sigma_{i<j} 2p_i p_j g_{ij}(x_1,x_2) g_{ij}(y_1,y_2) & otherwise \end{cases} \tag{5.13}$$

where $g_i(\cdot)$ denotes the density function for a $\mathcal{N}(a_i,\sigma_i^2)$ distribution. This likelihood may be easily modified to account for coalescence. An important feature to note is that if either $x_1 \neq x_2$ or $y_1 \neq y_2$ then, under H_G, it is known that the suspect/specimen is heterozygous (since it is assumed to be impossible for a true homozygote to exhibit a double band pattern). Thus, 5.13 becomes,

$$L(\tilde{x},\tilde{y}|H_G) =$$

$$\begin{cases} \Sigma_{i=1}^{A} p_i^2 g_i(z_1) g_i(z_2) + \\ \Sigma_{i<j} 2p_i p_j \int_0^\infty g_{ij}(z_1 - t, z_1 + t)\delta(t,z_1)dt \cdot \\ \int_0^\infty g_{ij}(z_2 - t, z_2 + t)\delta(t,z_2)dt & x_1 = x_2 = z_1,\ y_1 = y_2 = z_2 \\ \Sigma_{i<j} 2p_i p_j g_{ij}(x_1,x_2) \int_0^\infty g_{ij}(z_2 - t, z_2 + t)\delta(t,z_2)dt & x_1 \neq x_2,\ y_1 = y_2 = z_2 \\ \Sigma_{i<j} 2p_i p_j g_{ij}(y_1,y_2) \int_0^\infty g_{ij}(z_1 - t, z_1 + t)\delta(t,z_1)dt & x_1 = x_2 = z_1,\ y_1 \neq y_2 \\ \Sigma_{i<j} 2p_i p_j g_{ij}(x_1,x_2) g_{ij}(y_1,y_2) & x_1 \neq x_2,\ y_1 \neq y_2 \end{cases}$$

Given an estimate of the measurement error correlation ξ (see Section 5.2.1), it remains only to estimate the allele distribution (i.e. to obtain estimates $\hat{p}_j$ for $j = 1,\ldots,A$) in order to estimate R.

5.3.4 Estimating the Allele Distribution

Suppose that in the reference set (of allele measurements for n randomly chosen individuals at a particular locus), n^* double band patterns were observed, and let them be denoted by $\tilde{x}_j = (x_{j_1}, x_{j_2})$, $j = 1,\ldots,n^*$. Under the assumption of random mating these $2n^*$ measurements are all independent. In addition, suppose that the

reference sample contains $(n - n^*)$ single band patterns, and let them be indexed by $j = n^* + 1, \ldots, n$. That is, the single band measurements in the reference sample may be written as $\tilde{z}_{n^*+1}, \ldots, \tilde{z}_n$, where $\tilde{z}_j = (z_j, z_j)$, $j = n^* + 1, \ldots, n$. Let $\tilde{X} = (\tilde{x}_1, \ldots, \tilde{x}_{n^*}, \tilde{z}_{n^*+1}, \ldots, \tilde{z}_n)$ denote the entire reference sample of measurements. In the multiple flanking region case different alleles may have the same length, whereas in the single flanking region case, alleles may be uniquely indexed by the number of repeats. The two cases will thus be considered separately.

Additional Assumption

The errors for two measurements taken at a single locus will be assumed to be uncorrelated. This greatly simplifies the calculations for this section and Devlin, Risch, and Roeder feel that it will not affect the results in a significant manner. This assumption is *not* made for the other parts of the analysis, only for the estimation of the allele distribution.

The Model for One Flanking Region

Suppose that cutting the DNA (at a particular locus) with a particular enzyme results in only one possible flanking region size, denoted by u. Let,

$$a_r = u + r\rho$$

where a_r is the length of the allele with r repeats, each repeat of length ρ, and u is the single flanking region size. Unless coalescence has occurred, the observed measurement of an allele of length a_r is $u+r\rho+\varepsilon$ where ε is approximately normally distributed with mean zero and standard deviation $\sigma_r=(5.75\text{x}10^{-3})a_r$. Let $g_r(\cdot)$ denote a normal density function with mean a_r and variance σ_r^2.

Let R^* denote the random variable for the number of repeats and let π_r denote the relative frequency of an allele of length a_r. That is,

$$\pi_r = p(R^* = r), \quad r = 1, \ldots, b.$$

Since only one flanking region size is possible, the different values of r index all the different allele sizes. That is,

$$p_i = \pi_i, \; i = 1, \ldots, A = b.$$

This will not hold true in the multiple flanking region case.

Letting $x = u + r\rho + \varepsilon$, the probabilty density function of a single measurement x may be written as,

$$\begin{aligned} f(x|\pi, u) &= \Sigma_{r=1}^{b} p(R^* = r) p(X = x|\pi, u, R^* = r) \\ &= \Sigma_{r=1}^{b} \pi_r g_r(x). \end{aligned}$$

Using the additional assumption that errors for a pair of measurements at a locus are independent, then for a given double band pair $(x_{j_1}, x_{j_2}) = (u + r_1\rho + \varepsilon_1, u + r_2\rho + \varepsilon_2)$, the joint probability density of (x_{j_1}, x_{j_2}) may be written as,

$$f(x_{j_1}, x_{j_2}|\tilde{\pi}, u) = f(x_{j_1}|\tilde{\pi}, u) f(x_{j_2}|\tilde{\pi}, u)$$

where $\tilde{\pi}$ denotes the vector of allele relative frequencies, $(\pi_1, \ldots, \pi_b)$.

The probability of coalescence is modelled as a function of average measurement length $(z = \frac{1}{2}(x_{j_1} + x_{j_2}))$ and the difference between the two measurements $(t = \frac{1}{2}|x_{j_1} - x_{j_2}|)$. This estimation procedure is outlined in this section. Letting $\delta(t, z)$ denote the conditional probability of coalescence given t and z, and letting C denote the event of such coalescence, the likelihood of the data if a double band pattern (x_{j_1}, x_{j_2}) is observed is,

$$\begin{aligned} L(x_{j_1}, x_{j_2}|\tilde{\pi}, u) &= f(x_{j_1}, x_{j_2}|\tilde{\pi}, u, C) p(C) + f(x_{j_1}, x_{j_2}|\tilde{\pi}, u, \bar{C}) p(\bar{C}) \\ &= f(x_{j_1}|\tilde{\pi}, u) f(x_{j_2}|\tilde{\pi}, u)[1 - \delta(t_j, z_j)], \end{aligned}$$

where $\bar{C}$ denotes the complement of C. Similarly, the likelihood for a single band measurement z_j is,

$$\begin{aligned} L(z_j|\tilde{\pi}, u) &= \int_0^\infty f(z_j - t|\tilde{\pi}, u) f(z_j + t|\tilde{\pi}, u)\delta(t, z_j)dt \\ &\approx \Sigma_{k=1}^{M} f(z_j - t_k|\tilde{\pi}, u) f(z_j + t_k|\tilde{\pi}, u)\delta(t_k, z_j)\Delta_k \end{aligned}$$

where $t_1, \ldots, t_M$ is an evenly spaced grid, and $\Delta_k = t_{k+1} - t_k$. Hence the likelihood of the entire set of data may be written,

$$\begin{aligned} L(data|\tilde{\pi}, u) \approx\ & \prod_{j=1}^{n^*}(f(x_j|\tilde{\pi}, u))(f(y_j|\tilde{\pi}, u))[1 - \delta(t_k, z_j)] \\ & \prod_{j=n^*+1}^{n} \Sigma_{k=1}^{M} f(z_j - t_k|\tilde{\pi}, u) f(z_j + t_k|\tilde{\pi}, u)\delta(t_k, z_j)\Delta_k, \end{aligned} \tag{5.14}$$

With prior estimates of b and $\delta(t, z)$ the quantities u and $\tilde{\pi}$ may be estimated by using the EM algorithm (see Appendix B for details).

The Model for Multiple Flanking Regions

Suppose that now an enzyme is used to cut the DNA which results in L different flanking region sizes. The allele lengths may be written as,

$$a_{rl} = u_l + r\rho \qquad r = 1, \ldots, b, \quad l = 1, \ldots, L$$

where u_l is the lth flanking region size. Let $\tilde{u} = (u_1, \ldots, u_L)$ denote the vector of ordered flanking region sizes. The observed measurement of a_{rl} may be written as $x = u_l + r\rho + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma_{rl}^2)$ with $\hat{\sigma}_{rl} = (5.75\text{x}10^{-3})a_{rl}$. Let $g_{rl}(\cdot)$ denote a normal density with mean a_{rl} and variance σ_{rl}^2. Let γ_{rl} denote the allele relative frequency of an allele of size a_{rl} and let ϕ_l denote the proportion of alleles with flanking region size u_l. Again let π_r denote the proportion of alleles with r repeats. Clearly,

$$\phi_l = \sum_{r=1}^{b} \gamma_{rl},$$

and,

$$\pi_r = \sum_{l=1}^{L} \gamma_{rl}.$$

Recall that in the multiple flanking region case the probabilities $\pi_r,\ r = 1, \ldots, b$ are not completely analogous to the p_i's, $i = 1, \ldots, A$ as they are in the single flanking region case. Instead,

$$p_i = \sum \gamma_{rl},$$

where the summation is over all r and l such that $a_{rl} = a_{(i)}$.

Analogous to the single flanking region case, the probability density of a fragment measurement $x = u_l + r\rho + \varepsilon$ may be written as,

$$f(x|\tilde{u}, \tilde{\pi}, \tilde{\phi}) = \sum_{l=1}^{L} \sum_{r=1}^{b} \gamma_{rl} g_{rl}(x - u_l - r\rho).$$

As a simplifying assumption it will be assumed that the variation in flanking region size and number of repeats are independent. That is,

$$\gamma_{rl} \ = \ \phi_l \pi_r \quad \forall \ l = 1, \ldots, L, \quad r = 1, \ldots, b.$$

The probability density of a single fragment may then be written as,

$$f(x|\tilde{\pi}, \tilde{\phi}, \tilde{u}) = \sum_{l=1}^{L} \phi_l \sum_{r=1}^{b} \pi_r g_{rl}(x - u_l - r\rho).$$

The likelihood of the data may be written exactly as in 5.14 but with $f(\cdot|\tilde{\pi}, u)$ replaced by $f(\cdot|\tilde{\pi}, \tilde{\phi}, \tilde{u})$. That is,

$$\begin{aligned} L(data|\tilde{\pi}, \tilde{\phi}, \tilde{u}) \ \approx \ & \prod_{j=1}^{n^*} (f(x_j|\tilde{\pi}, \tilde{\phi}, \tilde{u}))(f(y_j|\tilde{\pi}, \tilde{\phi}, \tilde{u}))[1 - \delta(t_k, z_j)] \cdot \\ & + \ \prod_{j=n^*+1}^{n} \Sigma_{k=1}^{M} f(z_j - t_k|\tilde{\pi}, \tilde{\phi}, \tilde{u}) f(z_j + t_k|\tilde{\pi}, \tilde{\phi}, \tilde{u}) \delta(t_k, z_j) \Delta_k. \end{aligned} \tag{5.15}$$

If prior estimates of L, b, and $\delta(t, z)$ are obtained (discussed in the following sections) then the EM algorithm may again be used to obtain estimates of $\tilde{\pi}$, $\tilde{u}$, and $\tilde{\phi}$ (see Appendix B for details).

Estimating the Number of Flanking Regions (L)

The number of possible flanking regions L may be estimated by cutting a sample of DNA fragments with two different enzymes and comparing the results (Devlin,

Risch, and Roeder, 1991). Let E_1 denote an enzyme which results in L different flanking region sizes at a particular locus, L unknown. Let E_2 denote an enzyme which results in only one flanking region size. Further let (X_{11}, X_{12}) and (X_{21}, X_{22}) denote the observed fragment lengths at a particular locus using enzymes E_1 and E_2 respectively. That is,

$$(X_{11}, X_{12}) = (u_l + r_1\rho + \varepsilon_1, u_l + r_2\rho + \varepsilon_2)$$

$$(X_{21}, X_{22}) = (\omega + r_1\rho + \nu_1, \omega + r_2\rho + \nu_2).$$

Taking differences yields,

$$X_{1j} - X_{2j} = u_l - \omega + \varepsilon_j - \nu_j$$

where this difference is normally distributed with mean $u_l - \omega$ and variance $\sigma^2_{(u_l + r_j\rho)} + \sigma^2_{(\omega + r_j\rho)}$. For each individual in the sample these differences are computed, and the result should yield L different clusters, since the effects of r_1 and r_2 have been removed

Estimating the Number of Possible Alleles (b)

As an estimate of the number of possible allele sizes, Devlin, Risch, and Roeder (1991) suggest,

$$\hat{b} = \frac{(L - S)}{\rho}$$

where L and S are the lengths of the longest and shortest observed fragments respectively, and ρ is the length of the repeat.

Estimating the Coalescence Probability

In order to estimate the probability of coalescence for two allele measurements x and y, Devlin, Risch, and Roeder (1990) model it as a function of the average distance (τ) between the two measurements, $\tau = \frac{|x-y|}{2}$. For any pair of measurements x_j and

y_j the probability of coalescence is reflected by $(1 - \frac{O_j}{H_j})$ where O_j and H_j are the observed and expected number of heterozygotes for the interval containing τ. Let c_j denote the midpoint of this interval. Devlin, Risch, and Roeder found that a plot of $(1 - \frac{O_j}{H_j})$ versus c_j for various pairs x_j and y_j was well approximated by a logistic model. For the exact model see Devlin, Risch, and Roeder (1990). The fitted model represents the estimated probabilities of coalescence ($\delta(\tau, z)$) for various values of τ.

This procedure does not take into account the fact that the probability of coalescence depends on the mean fragment pair size, $z = \frac{x_j + y_j}{2}$. Larger fragments will be more likely to coalesce than shorter fragments since they do not travel as far in the gel and hence do not have as much time to separate properly. Although they claim that this dependence on z will have little impact on the results, Devlin, Risch, and Roeder (1990) suggest a method of including it in the model, assuming there is a linear relationship between z and the probability of coalescence.

Identifiability of the Model

Let $H = \sum_{j=1}^{k} p_j F_j(\tilde{\theta})$ denote any finite mixture. For known k, H is identifiable if and only if the $F_j(\tilde{\theta})$'s are distinct for $j = 1, \ldots, k$. If the model is not identifiable then different values of $\tilde{\theta}$ may result in the same value for H.

Consider the probability density function of an allele measurement x given by,

$$f(x|\tilde{u}, \tilde{\pi}, \tilde{\phi}) = \sum_{l=1}^{L} \sum_{r=1}^{b} \gamma_{rl} g_{rl}(x - u_l - r\rho).$$

If either L or b is unknown, the model lacks identifiability. Procedures for estimating these quantities were outlined earlier in this section.

In addition, in the multiple flanking region case, there may be overlap in the allele distributions in which case there exist pairs $a_{r_1 l_1} = u_{l_1} + r_1\rho$ and $a_{r_2 l_2} = u_{l_2} + r_2\rho$ such that $a_{r_1 l_1} = a_{r_2 l_2}$ but $r_1 \neq r_2, l_1 \neq l_2$. In this case $g_{r_1 l_1}(x - u_{l_1} - r_1\rho) = g_{r_2 l_2}(x - u_{l_2} - r_2\rho)$

and the model is not identifiable. Two special cases exist in which the overlap in the allele distributions (and resulting lack of identifiability) may be ignored. These situations may be described as follows.

(i) If the amount of overlap between the allele distributions is small, Devlin, Risch, and Roeder recommend proceeding as though there were actually no overlap. Measurements of length $a_{r_1 l_1} = a_{r_2 l_2}$ would contribute to the estimates of both $\gamma_{r_1 l_1}$ and $\gamma_{r_2 l_2}$. This is effectively "double counting" these measurements but it is impossible to tell whether they are measurements of an allele with r_1 repeats and a flanking region of length u_{l_1}, or an allele with r_2 repeats and a flanking region of length u_{l_2}.

(ii) If for one of the overlapping distributions ϕ_l is small, then Devlin, Risch, and Roeder recommend ignoring the rare allele distribution, since it will result in very few observations. For example, if as before $a_{r_1 l_1} = a_{r_2 l_2}$ and ϕ_{l_1} is small, then all observations of length $a_{r_1 l_1}$ $(= a_{r_2 l_2})$ will contribute to the estimate of $\gamma_{r_2 l_2}$, with $\gamma_{r_1 l_1}$ taken to be zero.

Furthermore, there is a special case in which although the allele distributions overlap, the model is still identifiable (for known L and b). Let U and R denote the random variables for flanking region size and number of repeats respectively. If U and R are independent (i.e. $\gamma_{rl} = \pi_r \phi_l$) then the model is identifiable even if the allele distributions overlap. With this assumption the probability density of a measurement x may be written as,

$$\begin{aligned} f(x|\tilde{u}, \tilde{\pi}, \tilde{\phi}) &= \textstyle\sum_{l=1}^{L} \sum_{r=1}^{b} \gamma_{rl} g_{rl}(x - u_l - r\rho) \\ &= \textstyle\sum_{l=1}^{L} \phi_l \sum_{r=1}^{b} \pi_r g_{rl}(x - u_l - r\rho). \end{aligned}$$

The density $f(x|\tilde{u}, \tilde{\pi}, \tilde{\phi})$ is still a mixture, only now the densities being mixed are,

$$F_l = \sum_{r=1}^{b} \pi_r g_{rl}(x - u_l - r\rho) \quad for\ l = 1, \ldots, L.$$

To see that the model is identifiable consider the following. Suppose that $F_{l_1} = F_{l_2}$ for some $l_1 \neq l_2$. That is,

$$\sum_{r=1}^{b} \pi_r g_{rl_1}(x - u_{l_1} - r\rho) = \sum_{r=1}^{b} \pi_r g_{rl_2}(x - u_{l_2} - r\rho).$$

This implies,

$$\sum_{r=1}^{b} \pi_r [g_{rl_1}(x - u_{l_1} - r\rho) - g_{rl_2}(x - u_{l_2} - r\rho)] = 0,$$

with $\pi_r > 0 \ \ \forall\, r = 1, \ldots, b$. Hence $u_{l_1} = u_{l_2}$, or equivalently $l_1 = l_2$, a contradiction. All members F_l, $l = 1, \ldots, L$, must therefore be distinct, and the model is identifiable (for known L and b).

Smoothed Estimators

If there are many allele frequencies to be estimated at a particular locus the variance of these estimates may be high (due to sampling variability). Also, allele frequency estimates will tend to be negatively correlated with neighbouring estimates (due to measurement error). As an adjustment, Devlin, Risch, and Roeder (1992) suggest smoothing the allele frequency estimates with an empirical Bayes approach to shrinkage estimation, to produce estimators with a reduced average mean squared error.

If we assume that the π_r's, $r = 1, \ldots, b$, have prior distribution $\mathcal{N}(\nu, \tau^2)$ and that the conditional distribution of the estimate $\hat{\pi}_r$, given the true allele frequency π_r is $\mathcal{N}(\pi_r, \sigma^2)$, for $r = 1, \ldots, b$, then an improved estimate of π_r, call it π_r^*, may be obtained by shrinking towards the mean of the posterior distribution (i.e. towards the mean of the conditional distribution of π_r given the estimate $\hat{\pi}_r$). This yields an estimator of the form,

$$\hat{\pi}_r^* = (1 - B)\hat{\pi}_r + B\nu, \quad for\ r = 1, \ldots, b \tag{5.16}$$

where,

$$B = \frac{\sigma^2}{\sigma^2 + \tau^2} \tag{5.17}$$

is known as the shrinkage factor. The Stein shrinkage factor given by,

$$B_S = min[\frac{(b-2)\hat{\sigma}^2}{\sum_{r=1}^{b}(\hat{\pi}_r - \nu)^2}, 1]$$

is an estimate of 5.17. If ν is unknown an empirical Bayes approach could be used with ν estimated by $\frac{1}{b}\sum_{r=1}^{b}\hat{\pi}_r$. This is not recommended by Devlin, Risch, and Roeder however since many of the allele probability estimates are near zero (usually reflecting truly small allele relative frequencies), and shrinking towards the average will obscure this feature. The estimates should really be smoothed locally That is, for each π_r, choose a separate estimate $\hat{\nu}_r$, $r = 1, \ldots, b$. Devlin, Risch, and Roeder suggest obtaining the $\hat{\nu}_r$'s using nonparametric kernel regression, whereby $\hat{\nu}_r$ is a weighted average of $\hat{\pi}_r$ and neighbouring estimates $\hat{\pi}_i$, $i \neq r$, with weights decreasing as $|i - r|$ increases. Let,

$$\nu_r^* = \sum_{i=1}^{b} K_{ri}\hat{\pi}_i, \quad for\ r = 1, \ldots, b,$$

where K_{ri} denotes the weight allotted to $\hat{\pi}_i$. Since the allele probabilities are being estimated these ν_r^*'s should sum to one, hence the standardized estimates are used given by,

$$\nu_r^s = \frac{\nu_r^*}{\sum_{r=1}^{b}\nu_r^*}, \quad for\ r = 1, \ldots, b.$$

A reasonable choice for K_{ri} is a normal density function with mean $i\rho$ and variance σ_i^2, evaluated at $r\rho$. That is,

$$K_{ri}^* = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\frac{-1}{2\sigma_i^2}(r\rho - i\rho)^2,$$

for $r = 1, \ldots, b$, $i = 1, \ldots, b$. This choice of K_{ri} reflects the fact that the contribution of neighbouring estimates should depend on the standard deviation of the

measurement error, σ_i, $i = 1, \ldots, b$. That is, it should depend on how much false measurements of alleles of length a_i could have contributed to the estimate of the relative frequency of alleles of length a_r. Since kernel functions must sum to one, the K^*_{ri}'s are standardized to obtain,

$$K^s_{ri} = \frac{K^*_{ri}}{\Sigma^b_{r=1} K^*_{ri}}, \quad for\ r = 1, \ldots, b.$$

With these choices the improved estimates of the form 5.16 become,

$$\hat{\pi}^*_r = (1 - B_r)\hat{\pi}_r + B_r \nu_r,$$

where,

$$B_r = min[\frac{bs^2_r}{\Sigma^b_{r=1}(\hat{\pi}_r - \nu^s_r)^2}, 1],$$

and where s^2_r is the bootstrap estimate of the variance of $\hat{\pi}_r$. Since the π_r's are probabilities the improved estimates should be standardized and written as,

$$\hat{\pi}^s_r = \frac{\hat{\pi}^*_r}{\Sigma^b_{r=1}\hat{\pi}^*_r}, \quad for\ r = 1, \ldots, b.$$

Devlin, Risch, and Roeder perform several simulations which indicate that the empirical Bayes smoothing approach improves on the unadjusted maximum likelihood estimates obtained from the EM algorithm procedure.

5.3.5 Remarks

Although this method incorporates flanking region polymorphism and obtains maximum likelihood estimates of the allele relative frequencies, additional assumptions were required. It is not obvious which approach is the most appropriate. In order to decide which method is best suited to a particular set of data it is necessary to evaluate the goodness of fit of each of the models. The next section describes some potential goodness of fit tests.

5.4 Evaluating the Goodness of Fit of the Model

It is necessary to evaluate the goodness of fit of the estimated density function $f(\tilde{x}|\hat{u}, \hat{\pi}, \hat{\phi})$. If flanking region polymorphism is not being incorporated, $\hat{u}$ and $\hat{\phi}$ may be removed from the density. A standard chi-square goodness of fit analysis would proceed as follows. Partition the range of possible allele size measurements into K nonoverlapping subintervals $\Gamma_1, \ldots, \Gamma_K$. Let O_k denote the number of observed measurements falling in the interval Γ_k, and let E_k denote the number of measurements which would be expected to fall in interval Γ_k if $f(\tilde{x}|\hat{u}, \hat{\pi}, \hat{\phi})$ were the true density of the measurements. Consider the test statistic,

$$T = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k}.$$

In the usual scenario, under the null hypothesis that $f(\tilde{x}|\hat{u}, \hat{\pi}, \hat{\phi})$ is the true density, this test statistic would be asymptotically distributed as a chi-square random variable provided all parameters were estimated using the *counts*, O_k. This is usually very difficult and here the parameters have been estimated from the original data. Hence, as is well known, the asymptotic distribution of X^2 is not chi-square. Further, coalescence has occurred in the measurements which may be unaccounted for in the density. The null distribution of the test statistic is thus unknown. However, it may be estimated using Monte Carlo methods. A group of s data sets of size n are generated from $f(\tilde{x}|\hat{u}, \hat{\pi}, \hat{\phi})$. For each of the s data sets the n observations are paired at random and then coalesced according to the estimated coalescence density $\delta(t, z)$. For each data set, compute T_i, $i = 1, \ldots, s$. The empirical distribution of the T_i's may now be used to estimate the distribution of the test statistic under the null hypothesis. The p value for the goodness of fit test is the proportion of the T_i's which are larger than the test statistic for the original data set. Devlin, Risch, and Roeder suggest that any value of $s > 300$ should be sufficient.

As an additional method of judging the suitability of the model Devlin, Risch, and Roeder (1991) discuss a way to estimate the variance of the allele probability estimates in the single flanking region case ($var(\hat{\pi}_i)$ for $i = 1, \ldots, b$), using a parametric bootstrap. Generate B data sets of size n from $f(\cdot|\hat{\pi}, \hat{u})$. The n observations in each data set are paired at random and coalesced according to the estimated coalescence probability $\delta(t, z)$. The data sets are generated from the density $f(\cdot|\hat{\pi}, \hat{u})$ instead of from $f(\cdot|\hat{\pi}, \hat{u}, \hat{\phi})$ because, to simplify computations, the flanking region size is assumed to be constant (i.e. a single flanking region size u). This is reasonable since variation in the flanking region will not contribute significantly to the variance of the allele probability estimates. For each of the B data sets, the vector of allele probability estimates ($\hat{\Pi}_j = (\hat{\pi}_{j1}, \ldots, \hat{\pi}_{jb})$, $j = 1, \ldots, B$) is obtained and the vector of sample variances of these estimates is used to estimate $var(\hat{\pi}_i)$ for $i = 1, \ldots, b$.

As an important alternative approach to the parametric bootstrap, Devlin, Risch, and Roeder point out that an ordinary bootstrap resample could be performed on the original set of paired data. This would be a significantly better approach if the population is not in HWE, since the dependence between allele pairs would then be automatically included in the calculations.

Although these goodness of fit techniques give an indication of how well the models approximate the distribution of the DNA profiling data, concerns regarding the underlying assumptions of these models have not been addressed. Some of these issues will be discussed in the following chapter.

Chapter 6

Presence and Effects of Population Substructure

6.1 The Hardy-Weinberg Equilibrium Assumption

Deviations from HWE result primarily from population substructure. It is often claimed that humans do in fact tend to form groups based on religion, ethnicity, or geography, and to mate within those groups. They are inadvertantly mating nonrandomly with respect to their genes. Allele frequencies within a group will tend to be similar and follow HWE expectations, but treating these distinct groups as one larger population may invalidate the HWE assumption. It has been shown that there are significant differences between the allele relative frequencies in the Caucasian, Black, and Hispanic populations in the United States (see for example, Balazs et al., 1989). In forensic calculations the reference population is often taken to be one of these three subpopulations. It has been argued, however, that these groups are themselves composed of smaller subpopulations which differ with respect to their allele frequencies.

Various statistical methods have been proposed to examine this issue, but no consensus has been reached. Certain groups of population geneticists feel that the

issue can be resolved by applying statistical goodness of fit tests of HWE expectations, and tests for population substructure. Others feel that these tests lack the power to detect Hardy-Weinberg disequilibrium when it does exist, and that the issue can only be resolved by extensive sampling from all these smaller subpopulations. Clearly sampling from a wide variety of subpopulations could virtually resolve the debate but this is a very time consuming process and there are still issues of how far to go with this sampling. As Caskey (1991) puts it, "Do we use a Neapolitan or Sicilian data base on a fourth-generation Italian defendant?". Statistical testing, at least in the interim, provides valuable information in this regard.

In spite of extensive work in this area, no agreement has been reached about whether there is a significant amount of population substructure, and if there is, how much impact it actually has on genotype relative frequencies. For illustration some statistical tests are presented in this section which have been applied with varying results.

Hernández and Weir (1989), propose a test based on disequilibrium coefficients (D_{ij}), which measure the departure of the relative frequency of heterozygotes from the HWE expected frequency ($2\hat{p}_i\hat{p}_j$), where $\hat{p}_i$ is the estimated relative frequency of an allele of size a_i. The test does not consider departures of homozygotes from HWE expectations since, to be conservative, $2\hat{p}_i$ is usually used as the estimated probability of a homozygote instead of $\hat{p}_i^2$. The HWE assumption is thus invoked only for double band patterns, and departures from HWE expectations for homozygotes is not a concern. The disequilibrium coefficients may be written as,

$$\hat{D}_{ij} = \hat{p}_i\hat{p}_j - \frac{1}{2}\hat{P}_{ij} \quad i,j = 1,\ldots,b$$

where b is the number of possible alleles at the locus, and P_{ij} is the relative frequency of individuals heterozygous for alleles of size a_i and a_j. For two particular alleles with

estimated frequencies $\hat{p}_1$ and $\hat{p}_2$ the hypotheses to be tested are,

H_o: $P_{12} = 2p_1p_2$ (i.e. $D_{12} = 0$)

versus,

H_a: $P_{12} \neq 2p_1p_2$ (i.e. $D_{12} \neq 0$).

The test statistic is,

$$X^2 = \frac{\hat{D}_{12}^2}{\hat{var}(\hat{D}_{12})}$$

$$= \frac{2n\hat{D}_{12}^2}{\hat{p}_1\hat{p}_2[(1-\hat{p}_1)(1-\hat{p}_2)+\hat{p}_1\hat{p}_2]+(\hat{p}_1^2\hat{D}_{23}+\hat{p}_2^2\hat{D}_{13})}$$

where D_{13} and D_{23} are the disequilibrium coefficients based on $\hat{p}_3 = 1 - \hat{p}_1 - \hat{p}_2$. Under the null hypothesis, this statistic has asymptotically a chi-square distribution with one degree of freedom.

This test was applied to the data sets of the FBI (Weir, 1992, 1993), Cellmark Diagnostics (Weir, 1992), and Lifecodes (Weir, 1992). For the FBI data the fixed bin approach was used to obtain the allele relative frequency estimates. The Lifecodes and Cellmark Diagnostic data were analyzed with the floating bin approach. Each data set consisted of measurements for a Caucasian, Black, and Hispanic subpopulation at a variety of loci. The FBI data set was further subdivided into Texas, Florida, and California subpopulations. For each of the loci, the null hypothesis of no disequilibrium was rejected approximately 5% of the time at the 5% level of significance, which is what one would expect if the null hypothesis were true.

A criticism of this test is that it is a local test (i.e. tests only one pair of alleles at a time) with only one degree of freedom, and as a result may have low power. The noncentral chi-square (χ^2) distribution may be used to examine the power of this test. That is, the minimum level of disequilibrium (D_{ij}^{MIN}) that will be detected

by the above (5 percent level) test with a certain probability may be worked out. If there is disequilibrium then X^2 will have, asymptotically, a *noncentral* χ^2 distribution with noncentrality parameter $\mathcal{E}(X^2)$. For a specified significance level α, and desired power $100(1-\beta)\%$ it is necessary that,

$$p_{H_a}(X^2 < c) = \beta$$

where c defines the critical region such that $p_{H_o}(X^2 > c) \leq \alpha$. Equivalently,

$$p_{H_a}(\chi^2(\lambda) < c) = \beta \tag{6.1}$$

where $\lambda = \mathcal{E}(X^2)$ is the noncentrality parameter (which will depend on D_{ij}). Solving 6.1 for D_{ij} yields the minimum departure from HWE (D_{ij}^{MIN}) which will be detected with $100(1-\beta)\%$ probability. This may be calculated for a specific application of the test and if D_{ij}^{MIN} is small it may be used to address concerns about lack of power. Clearly for any specific significance level α, and level of disequilibrium D_{ij}, the power may be calculated by solving,

$$p_{H_a}(X^2 < c) = \beta$$

with respect to β.

The above disequilibria test may also be modified to be a test for an *overall* departure from HWE, by using the classical test statistic,

$$X^2 = \Sigma_{i=1}^{b}\Sigma_{j=1}^{b}\frac{n(\hat{P}_{ij} - \hat{p}_i\hat{p}_j)^2}{\hat{p}_i\hat{p}_j}.$$

Under the null hypothesis that all the disequilibrium coefficients are zero, X^2 has a χ^2 distribution with $\frac{b(b-1)}{2}$ degrees of freedom.

As an alternative to the above disequilibria tests, Hernández and Weir (1989) also suggest taking a likelihood ratio approach. The likelihood of the data is computed

under the hypothesis of HWE and HWD (Hardy-Weinberg disequilibrium). Denote these likelihoods by L_0 and L_1 respectively. The test statistic is,

$$G^2 = -2ln(\frac{L_1}{L_0}).$$

Under the null hypothesis of HWE, G^2 is approximately distributed as a χ^2 random variable with $\frac{b(b-1)}{2}$ degrees of freedom. This approximation is avoided by Weir (1992) in his analysis, and the empirical distribution of a set of bootstrapped data is used instead. For the FBI data set with probability estimates done by the fixed bin approach, in the results of Weir (1992) approximately one quarter of the tests showed significant departure from HWE.

Weir proposes and applies several other tests with varying results (see for example, Weir, 1991). Other examples of studies of this issue include Baird and Balazs (1986), Weir (1993), Devlin, Risch, and Roeder (1990), Cohen (1990), and Geisser and Johnson (1993). The issue of population substructure and its effect on allele relative frequencies for VNTR loci has clearly not been resolved as of yet. It may be possible, through these various statistical tests, to identify loci which consistently seem to conform to HWE expectations, and restrict attention to those for DNA profiling.

If a departure from HWE frequency is detected all is not lost. Other methods have been suggested which do not require the HWE assumption to be made (e.g. see Section 5.2). One frequently suggested approach is to compute the number of times, say x, that a particular single locus genotype was observed in a sample of S individuals. The quantity $max[\frac{1}{S}, \frac{x}{S}]$ may then be used as an estimate of the frequency of that genotype, rather than estimating the individual allele frequencies and multiplying. Nichols and Balding (1991) also provide an interesting alternative approach in which they attempt to find an upper bound for the effect of population substructure on the probability of obtaining a match between two randomly chosen individuals.

6.2 Linkage Equilibrium

The assumption of linkage equilibrium (LE) allows the multiplication of probabilities of events at different loci. If, for example, there is natural selection in the population which favours certain genotypic combinations over others, this assumption may be invalid. Much investigation has been carried out in this area and two tests for linkage disequilibrium (LD) with results are outlined in this section.

Risch and Devlin (1992) performed tests for LD on data obtained from the FBI and from Lifecodes. The Lifecodes data was divided into Caucasian, Black, and Hispanic subpopulations, with the Hispanic population further separated into Southeastern and Southwestern Hispanics. Two by two tables were constructed for each pair of loci, with the cells containing the observed number of matches and nonmatches for each of the two loci. For example, consider two loci D1S7 and D2S44, for the Caucasian subpopulation. The table constructed would be,

	Match at D1S7	No Match at D1S7
Match at D2S44	m_{11}	m_{12}
No match at D2S44	m_{21}	m_{22}

The expected value for each cell (under the null hypothesis of LE) is given by,

$$\begin{aligned}
\mathcal{E}(m_{11}) &= (N)p(\text{match at D2S44, match at D1S7}) \\
&= (N)p(\text{match at D2S44})p(\text{match at D1S7}) \\
&= (N)\tfrac{1}{N}(\#\text{matching at D2S44})\tfrac{1}{N}(\#\text{matching at D1S7}) \\
&= \tfrac{1}{N}(\#\text{matching at D2S44})(\#\text{matching at D1S7})
\end{aligned}$$

where N is the total number of comparisons. For each table, a chi-square goodness of fit test statistic was calculated and its distribution (under the null hypothesis of LE) was obtained by bootstrapping.

Since data were available on 5 loci for the FBI database, there were $\binom{5}{2}=10$ tables constructed for each of the four subpopulations, hence a total of forty tests were performed. Of these forty tests only three gave significant results. For the Lifecodes data set there were three loci used resulting in $\binom{3}{2}=3$ two by two tables for each of the three subpopulations. Of these nine tests, no results were significant (i.e. $p > 0.05$ for all tests).

Weir (1991) examines data obtained from the FBI database for which allele frequencies were determined by the fixed bin approach. A likelihood ratio test was used to test the null hypothesis of LE. Loci were analyzed in pairs and the likelihood of the observed set of two locus genotypes was calculated under the hypotheses of LE (L_0) and LD (L_1). The test statistic used was $T = -2(\ln L_1 - \ln L_0)$ with a bootstrapped distribution for T used to tabulate p-values. Data were available on six loci which resulted in $\binom{6}{2}=15$ paired comparisons for each of the subpopulations. The data were separated into Caucasian, Black, and Hispanic populations. These were further subdivided into Texas and Florida, as well as California in some cases. In total 141 tests were performed with approximately 60 significant results. Although this is far more than would be expected under the LE assumption, Weir shows that if only the double band patterns at each locus are considered then no significant values are obtained. This seems to suggest that the apparent departure from LE is due to the existence of pseudo-homozygotes and not to a lack of independence between loci.

Weir (1992) applies a disequilibrium test (Weir, 1979) to data obtained from Lifecodes and Cellmark Diagnostics. The database was divided into Caucasian and Black subpopulations and allele frequencies were estimated by the floating bin approach. Results were consistent with the hypothesis of linkage equilibrium.

Although the tests outlined here seem to strongly support the LE assumption,

there are many (e.g. Cohen, 1990) who feel that there is evidence to the contrary. Similar to the Hardy-Weinberg equilibrium debate the issue remains unresolved.

Chapter 7

Comparison of Approaches

The main advantage of the frequentist approach is the simplicity with which the results of the analysis may be presented in court. The statistic $\hat{P}(M|Y)$ described in Section 3.1 may be easily explained to the jury members whereas in a Bayesian analysis the relationship between prior and posterior odds must be carefully described. In fact, a "weak" Bayesian format may be used (Ellman and Kaye, 1979) in which the jury is not required to choose a specific prior distribution. Rather they are simply provided with the value of the likelihood ratio R, along with perhaps a few illustrations of how various choices of priors affect the posterior odds. Evett (see discussion of Berry, 1991) promotes a convention for which different values of the likelihood ratio R correspond to different statements about the strength of the evidence against the suspect. For example, a particular range of (large) values of R would correspond to "strong evidence", while a range of smaller values of R may correspond to "weak evidence" or even "no evidence".

An advantage of the Bayesian methods presented is that they make a serious attempt to model the observed properties of the data (e.g. measurement error). The match-binning approach merely tries to *overcompensate* for these properties in order to produce conservative estimates. Numerical comparisons between the frequentist and Bayesian methods are difficult since they differ fundamentally. In the frequentist

approach a formal match is declared (or the suspect is excluded) and the probability that such a match could have occurred between the specimen and a randomly selected individual is estimated. In the Bayesian approach however, no formal match is declared. Instead the relative weight of the DNA profile evidence in favour of G or I is given. It is clearly not appropriate to compare these two quantities. Berry (1991), however, describes a way in which the frequentist approach may be put in a Bayesian context as follows. Consider a single locus. Let "*match*" denote the event that the observed alleles x and y (for the suspect and specimen respectively) meet the matching criterion described in Section 4.3, and let "*exclusion*" denote the event that this criterion was not met. The DNA information obtained from the frequentist approach may be expressed as,

$$X_M = (match, y, z_1, \ldots, z_n)$$

$$X_N = (exclusion, y, z_1, \ldots, z_n),$$

where $z_1, \ldots, z_n$ is the reference sample of measurements for the locus. Further, suppose that $p(X_N|G) = 0$ and $p(X_M|G) = 1$ (i.e. no false exclusions). The quantity $p(X_M|I)$ is referred to as the match proportion and is the quantity calculated in the frequentist approach to estimate the proportion of people who could have contributed the incriminating sample. Examining the formula for the likelihood ratio $R = \frac{p(X|G,E)}{p(X|I,E)}$ it is clear that in the match binning context the DNA profile evidence X is either X_M or X_N. If an exclusion is declared, the likelihood ratio would thus be calculated as,

$$R = \frac{p(X_N|G,E)}{p(X_N|I)} = 0,$$

while if a match is declared,

$$R = \frac{p(X_M|G)}{p(X_M|I)} = \frac{1}{match\ proportion}.$$

Numerical comparisons between the two methods are now possible. One important feature is immediately obvious : if, for example, the DNA profiles are identical at all but one locus, then the frequentist method will be forced to declare an exclusion and the corresponding likelihood will be zero. On the other hand the Bayesian method yields a likelihood ratio which may still be quite large depending on how many other loci were examined. Since, as mentioned in Section 4.7, the one nonmatch could have been due to DNA degradation or other technical considerations it does not seem appropriate to have a likelihood ratio of zero. In this sense the Bayesian approach would seem superior.

Comparisons amongst the various Bayesian methods are also not straightforward. The method of Berry, Evett, and Pinchin (1992) (see Section 5.2) would seem to be an improvement over that of Berry (1991) (see Section 5.1) since the reasoning is similar but the assumption of no correlation between measurements taken at a single locus is no longer necessary. In addition, the controversial assumption of HWE is also no longer necessary. While Devlin, Risch, and Roeder (1991, 1992) extend their analysis even further to include flanking region polymorphism (see Section 5.3), they assume measurement errors are uncorrelated when estimating the allele distribution. Since it is possible to choose a restriction enzyme which results in little or no flanking region it may be preferable to use the approach of Berry, Evett and Pinchin. On the other hand, the mixture model approach of Devlin, Risch, and Roeder is intuitively more appealing than the *continuous* mixture model of Berry, Evett, and Pinchin. The true allele sizes do have a discrete distribution, and it is combined with the continuous measurement error distribution. They also obtain the maximum likelihood estimates of the allele probabilities rather than using a simple average of normal kernels to estimate the allele distribution. In addition, flanking region polymorphism results in a larger number of possible alleles at a locus, and thus yields greater discriminatory

power. The properties of the three Bayesian models are compared in Table 2.

The method chosen for analysis of the DNA profiling data should depend on which assumptions one is willing to accept. In general this will depend on the specific properties of the loci chosen to construct the profiles. Although the method considering flanking region polymorphism appears more appealing, some extra assumptions are necessary which may make it less attractive from a population genetics point of view.

Table 2 Comparison of Bayesian Methods

Method	Presentation of Evidence	Primary Assumptions	Method of Estimating the Allele Distribution
Berry (1991) (see Section 5 1)	Likelihood Ratio $=R$ $=P(X \mid G)/P(X \mid I)$ where X is the DNA profile evidence	(i) HWE (ii) LE (iii) measurement errors are independent (iv) measurements are lognormally distributed	Density averaging (normal kernels)
Berry, Evett, and Pinchin (1992) (see Section 5.2)	Likelihood Ratio $=R$ $=P(X \mid G)/P(X \mid I)$ where X is the DNA profile evidence	(i) LE (ii) measurement errors are correlated within a locus (iii) joint distribution of two measurement errors is bivariate normal	Density averaging (bivariate normal kernels)
Devlin, Risch, and Roeder (1991, 1992) (see Section 5 3)	Likelihood Ratio $=R$ $=P(X \mid G)/P(X \mid I)$ where X is the DNA profile evidence	(i) HWE (ii) LE (iii) measurement errors are correlated within a locus* (iv) measurement errors are normally distributed (v) joint distribution of two measurement errors is bivariate normal	EM Algorithm (with additional smoothing)

*For estimating the allele probability distribution the measurement errors are assumed to be uncorrelated within a locus

Chapter 8

Concluding Remarks

Theoretically, the approaches presented here seem more than adequate for dealing with the analysis of DNA profile data, and yet the debate is still raging over the admissibility of this type of evidence in court. The real controversy to be resolved is not one of "frequentist versus Bayesian" but whether either of these two approaches can adequately quantify the DNA profile evidence.

Resolution of this issue will require further research in a number of areas. An extensive investigation of the specific population characteristics of a few loci commonly used for DNA profiling would be highly beneficial. For example, an investigation of whether significant population substructuring does exist at these loci. If the HWE and LE assumptions could be established as reasonable for even these few loci then at the very least a base profile could be constructed for which the previously described methods would be reasonable. While one would not obtain the vanishingly small probabilities which are often reported for multiple locus profiles, a matching profile of only a few loci is still important evidence. It has also been suggested (see Risch and Devlin, 1992) that even if significant population substructuring does exist that it does not have much of an impact on the validity of the results. Further application of the proposed approaches to simulated datasets with significant substructuring could find evidence to support this claim. This would alleviate fears about the consequences

of violating the assumptions.

It is also an appealing possibility that further statistical models could be developed that would require less controversial assumptions. These models could be tested extensively on simulated datasets. To evaluate existing methods an empirical comparison based on realistic, simulated data should be performed. In addition to illustrating which method is superior, the study may also point towards shortcomings which are common to all the methods.

In light of the fact that many other forms of evidence (e.g. motive, character witnesses, etc.) are admissible which are not subject to the same sort of scientific scrutiny, it would seem reasonable that the DNA evidence, at least in some form, be admissible in court even while the statistical controversy remains unresolved.

Appendix A

Statistical Appendix

While analysis of DNA fingerprinting data encompasses a wide variety of statistical techniques, three concepts (Kullback-Leibler information, the EM algorithm, and bootstrapping) are given closer scrutiny.

A.1 Kullback-Leibler (K-L) Information

Suppose $\tilde{X} = (x_1, \ldots, x_n)$ are independent and identically distributed observations with density function $f(\tilde{X}, \tilde{\theta})$, $\tilde{\theta}$ unknown. The K-L information (Kullback and Leibler, 1951), denoted $K_{\tilde{X}}(H_0, H_1)$ is a measure of the average discrepancy between two hypothesized distributions $f_0(\tilde{X}, \tilde{\theta})$ and $f_1(\tilde{X}, \tilde{\theta})$. This quantity is used at the *design* stage of an experiment to decide where to draw the data $\tilde{X}$. It may be written as,

$$K_{\tilde{X}}(H_0, H_1) \;=\; \mathcal{E}[\log(\frac{f_0(\tilde{X}, \tilde{\theta})}{f_1(\tilde{X}, \tilde{\theta})})].$$

where $\mathcal{E}$ is the expected value calculated under the null hypothesis H_0. $K_{\tilde{X}}(H_0, H_1)$ represents the mean information per observation in $\tilde{X}$ for discriminating between H_0 and H_1 when H_0 is true. Some properties of K-L information are as follows.

Properties

(i) If $K_{\tilde{X}}(H_0, H_1)$ is large then the discrepancy between $f_0(\tilde{X}, \tilde{\theta})$ and $f_1(\tilde{X}, \tilde{\theta})$ is large and the probability of a Type I error (i.e. falsely rejecting H_0) will be small. That is, the data will be informative.

(ii) If $K_{\tilde{X}}(H_0, H_1)$ is very small (close to zero) the test will have low power since $f_0(\tilde{X}, \tilde{\theta})$ and $f_1(\tilde{X}, \tilde{\theta})$ will be virtually indistinguishable for data in this region. That is, the data will be uninformative.

(iii) $K_{\tilde{X}_i}(H_0, H_1)$ is additive for independent random vectors $\tilde{X}_i$, $i = 1, \ldots, N$. That is,

$$\begin{aligned} K_{\tilde{X}_1, \ldots, \tilde{X}_N}(H_0, H_1) &= \Sigma_{i=1}^{N} K_{\tilde{X}_i}(H_0, H_1) \\ &= \Sigma_{i=1}^{N} \mathcal{E}[\log(\tfrac{f_0^i(\tilde{X}_i, \tilde{\theta})}{f_1^i(\tilde{X}_i, \tilde{\theta})})] \end{aligned}$$

where $f_i^i(\tilde{X}_i, \tilde{\theta})$ is the density of $\tilde{X}_i$, $i = 1, \ldots, N$.

In terms of DNA profile analysis it has been suggested (Lange, 1991) that the K-L information may be used to decide between competing loci/restriction enzyme combinations (see Section 3.1.2).

A.2 The Expectation-Maximization (EM) Algorithm

The EM algorithm (Dempster, Laird, and Rubin, 1977) is a method of performing maximum likelihood estimation in the presence of missing data. Suppose it is necessary to make inference about a parameter ψ based on an observed data vector $\tilde{X}$ in which missing values are present. Since the likelihood function of the incomplete data, $L(\tilde{X}|\psi)$, is difficult to work with, it is expressed in terms of the likelihood function for a hypothetical complete set of data $\tilde{Y}$. For a given vector of observations $\tilde{X}$,

there are many possible associated complete data vectors $\tilde{Y}$. It is assumed however that once $\tilde{X}$ is observed, $\tilde{Y}$ is known to lie in some subspace $\mathcal{Y}(\tilde{Y})$ of the set of all possible complete data vectors.

The algorithm

Let $f(\tilde{Y}|\psi)$ denote the density of the (hypothetical) complete data $\tilde{Y}$, and let $g(\tilde{X}|\psi)$ denote the density of the incomplete observed data $\tilde{X}$.

$$g(\tilde{X}|\psi) = \int_{\mathcal{Y}(\tilde{X})} f(\tilde{Y}|\psi)d\tilde{Y}$$

For a given $g(\tilde{X}|\psi)$, $f(\tilde{Y}|\psi)$ is not unique and may be chosen for convenience.

The conditional density of $\tilde{Y}$ given $\tilde{X}$ is,

$$k(\tilde{Y}|\tilde{X},\psi) = \frac{f(\tilde{Y}|\psi)}{g(\tilde{X}|\psi)}. \tag{A.1}$$

Taking logarithms in A.1 and rearranging yields,

$$l(\tilde{Y}|\psi) = l(\tilde{X}|\psi) + l(\tilde{Y}|\tilde{X},\psi)$$

where $l(\cdot|\cdot)$ denotes the logarithm of the likelihood.

Now let ψ_A denote an arbitrary value of ψ and take expected values with respect to $k(\tilde{Y}|\tilde{X},\psi)$ to obtain,

$$Q(\psi|\psi_A) = l(\tilde{X}|\psi) + H(\psi|\psi_A) \tag{A.2}$$

where,

$$Q(\psi|\psi_A) = \mathcal{E}[l(\tilde{Y}|\psi)|\tilde{X},\psi_A]$$

and,

$$H(\psi|\psi_A) = \mathcal{E}[l(\tilde{Y}|\tilde{X},\psi)|\tilde{X},\psi_A].$$

For any given ψ_A the function $H(\psi|\psi_A)$ can be shown to be maximized when $\psi = \psi_A$ (Rao, 1965). The value of ψ for which $Q(\psi|\psi_A)$ is maximized will be a function of ψ_A, say $M(\psi_A)$.

If ψ^* denotes the maximum likelihood estimate of ψ (i.e. ψ^* maximizes $l(\tilde{X}|\psi)$ with respect to ψ), then the right hand side of A.2 will be maximized when $\psi = \psi^*$ (since this will maximize both $H(\psi|\psi_*)$ and $l(\tilde{X}|\psi)$). Hence $Q(\psi|\psi_A)$ must be maximized at $\psi = \psi^*$. This yields the fixed point equation,

$$\psi^* = M(\psi^*). \tag{A.3}$$

Equation A.3 suggests an algorithm of the form $\psi^{(m+1)} = M(\psi^{(m)})$ since if this algorithm converges (discussed below) then assuming continuity of $M(\cdot)$,

$$\lim_{m\to\infty} \psi^{(m+1)} = \lim_{m\to\infty} M(\psi^{(m)}.$$

That is, $\psi^* = M(\psi^*)$ where $\psi^* = lim\ \psi^{(m)}$ and hence ψ^* is a solution of the fixed point equation A.3.

Once an initial estimate $\psi^{(0)}$ of ψ^* is chosen, $Q(\psi|\psi_0)$ is computed (the expectation step) and then maximized with respect to ψ to obtain $\psi^{(1)}$ (the maximization step). This process is continued until hopefully some convergence criterion is met.

Convergence of the algorithm

A drawback of the algorithm is that it does not necessarily converge. In fact, when it does converge it is not guaranteed to be at a global maximum. Successive iterations of the algorithm will never reduce the likelihood however, so that it will never converge to a local minimum. For a discussion of conditions for convergence to local and global maxima see, for example, Wu (1983) or Coupal (1992).

It has been suggested (Devlin, Risch, and Roeder, 1991) that the EM algorithm may be used to estimate the allele relative frequency distribution (see Appendix B).

A.3 Bootstrapping

Consider a sample of observations $\tilde{X} = (x_1, \ldots, x_n)$ from a distribution $F(\tilde{X}, \theta)$, and suppose that $\hat{\theta}$ is an estimator of the unknown parameter θ (possibly a vector). Bootstrapping (Efron, 1982) provides a way to assess the sampling properties of $\hat{\theta}$ when the distribution of $\hat{\theta}$ is unknown.

Suppose it is desired to estimate some property of $\hat{\theta}$, say $g(\hat{\theta})$ (e.g. $g(\hat{\theta}) = var(\hat{\theta})$). Ideally, an estimation procedure based on simulation would be,

(i) draw B random samples of size n from $F(\tilde{X}, \theta)$

(ii) calculate $\hat{\theta}_i$, for each bootstrap sample, $i = 1, \ldots, B$

(iii) estimate $g(\hat{\theta})$ based on the observed distribution of $\hat{\theta}_1, \ldots, \hat{\theta}_B$ (e.g. if $g(\hat{\theta}) = var(\hat{\theta})$ then estimate $g(\hat{\theta})$ by the sample variance of the $\hat{\theta}_i$'s).

Unfortunately, this procedure is inadequate due to the fact that $F(\tilde{X}, \theta)$ is unknown. Instead $F(\tilde{X}, \theta)$ is replaced in the above steps by its nonparametric maximum likelihood estimator $\hat{F}(\tilde{X})$, the empirical distribution function of the data. $\hat{F}(\tilde{X})$ assigns a probability of $\frac{1}{n}$ to each observed data value. Each of the B bootstrap samples drawn from $\hat{F}(\tilde{X})$ is equivalent to n independent draws from $x_1, \ldots, x_n$ taken with replacement. For each bootstrap sample, let the corresponding value of the estimator (calculated in step (ii)) be denoted by $\hat{\theta}_i^*$, for $i = 1, \ldots, B$. Now $g(\hat{\theta})$ may be estimated based on the observed distribution of the $\hat{\theta}_i^*$'s, $i = 1, \ldots, B$. For example, if $g(\hat{\theta}) = var(\hat{\theta})$ then estimate $g(\hat{\theta})$ by the sample variance of the $\hat{\theta}_i^*$'s.

The bootstrap technique can also be performed parametrically. For example, if it is suspected that the data are normally distributed then the same steps as above are followed except that the B bootstrap samples are drawn from a normal population

with mean vector ($\tilde{\mu}$) and covariance matrix (Σ) estimated from the original set of data (i.e. the sample mean and covariance).

Convergence

As $B \to \infty$, $\hat{g}^*(\hat{\theta})$ would approach $g(\hat{\theta})$ if $\hat{F}(\tilde{X})$ were in fact the true distribution of the data. Usually $\hat{F}(\tilde{X})$ will estimate $F(\tilde{X}, \theta)$ imperfectly, but it is the nonparametric maximum likelihood estimator (mle) of $F(\tilde{X}, \theta)$. Thus asymptotically $\hat{g}^*(\hat{\theta})$ is the nonparametric mle of $g(\hat{\theta})$. In fact, Efron and Tibshirani (1986) illustrate how even fairly small values of B (e.g. $B \leq 100$) give quite accurate results.

Confidence Intervals

The following are three of the methods discussed by Efron and Tibshirani (1986), for constructing $100(1-2\alpha)\%$ confidence intervals for θ using bootstrapping techniques. Let $G(s)$ be the cumulative distribution function of the bootstrap estimates $\hat{\theta}_i^*$, $i = 1, \ldots, B$.

(i) the standard method

A $100(1-2\alpha)\%$ confidence interval for θ is given by,

$$[\hat{\theta} - \hat{\sigma}^* z^{(\alpha)},\ \hat{\theta} + \hat{\sigma}^* z^{(\alpha)}],$$

where $\hat{\sigma}^*$ is the bootstrap estimate of the standard deviation of $\hat{\theta}$ and $z^{(\alpha)}$ is the $100(1-\alpha)$ percentile point of the standard normal distribution.

(ii) the percentile method

A $100(1-2\alpha)\%$ confidence interval for θ is given by,

$$[G^{-1}(\alpha),\ G^{-1}(1-\alpha)],$$

where $G^{-1}(\alpha)$ and $G^{-1}(1-\alpha)$ are the 100α and $100(1-\alpha)$ percentile points of $G(s)$ respectively.

(iii) the bias-corrected percentile method (BC method)

A $100(1-2\alpha)\%$ confidence interval for θ is given by,

$$[G^{-1}(\Phi\{2z_0 + z^{(\alpha)}\}),\ G^{-1}(\Phi\{2z_0 + z^{(1-\alpha)}\})],$$

where,

$$\Phi(s) \ = \ \int_{-\infty}^{s} h(\tilde{y})d\tilde{y},$$

with $h(\tilde{y})$ a standard normal density function, and where,

$$z_0 \ = \ \Phi^{-1}(G(\hat{\theta})).$$

If $G(s)$ is a normal distribution function then the standard and percentile methods are exactly equivalent, otherwise they may give very different results. The percentile intervals are transformation invariant while the standard intervals are not. That is, if $\tilde{\phi} = g(\theta)$ then the corresponding percentile interval for $\tilde{\phi}$ will simply be,

$$[g(G^{-1}(\alpha)),\ g(G^{-1}(1-\alpha))].$$

Thus, provided $\hat{\theta}$ is approximately normally distributed, the percentile confidence intervals for θ will be transformation invariant.

If $\hat{\theta}$ is a biased estimator of θ then the percentile intervals can be misleading. The BC intervals attempt to compensate for this. Suppose for example that $G(s)$ is perfectly symmetric about $\hat{\theta}$. Then $G(\hat{\theta}) = 0.5$ and $z_0 = \Phi^{-1}\{G(\hat{\theta})\} = 0$. In this case the BC intervals will be equivalent to the percentile intervals. Otherwise the BC intervals are adjusted accordingly to account for the skewness of the distribution $G(s)$.

Several other methods were discussed by Efron (1982a), and a full evaluation and comparison of numerous methods was given by Hall (1988).

In the course of modelling the DNA fingerprinting data, many parameters are estimated and technical problems (e.g. coalescence) may have occurred. Goodness of fit test statistics may thus not have their standard distributions. Bootstrapping may be used to estimate the distribution of these test statistics either under the null hypothesis or a specified alternative (as the empirical distribution of the B bootstrap test statistics, see Section 5.4), or it may be used to construct confidence intervals.

Appendix B

Estimating $\tilde{u}$ and $\tilde{\pi}$ Using the EM Algorithm

The DNA profiling data lend themselves to an "incomplete data" interpretation. The information available is a set of allele measurements, but it is not known (because of measurement error) which true allele sizes correspond to these measurements. In addition, for single band measurements, it is not known which measurements were actually obtained, only that they coalesced to an intermediate value, z_j, $j = n^* + 1, \ldots, n$. The EM algorithm provides maximum likelihood estimates of $\tilde{u}$ and $\tilde{\pi}$ by using the likelihood function of a hypothetical "complete" set of data ($\tilde{Y}$).

B.1 Single Flanking Region Case

Using the independence of measurements from different individuals, the likelihood of the incomplete data may be written as,

$$\begin{aligned} & L(\tilde{X}|\Psi) \\ \approx\ & \prod_{j=1}^{n^*} f(x_{j_1}|\Psi) f(x_{j_2}|\Psi)[1-\delta(l,z)] \cdot \\ & \prod_{j=n^*+1}^{n} \sum_{k=1}^{M} f(z_j - l_k|\Psi)\delta(l_k, z_j)\Delta_k \\ =\ & \prod_{j=1}^{n^*} (\sum_{r=1}^{b} \pi_r g_r(x_{j_1}))(\sum_{r=1}^{b} \pi_r g_r(x_{j_2}))[1-\delta(l,z)] \cdot \\ & \prod_{j=n^*+1}^{n} \sum_{k=1}^{M} (\sum_{r=1}^{b} \pi_r g_r(z_j + l_k))(\sum_{r=1}^{b} \pi_r g_r(z_j - l_k))\delta(l_k, z_j)\Delta_k \end{aligned}$$

where $\Psi = (u, \tilde{\pi})^T$.

The complete data may be written as,

$$y_j = \begin{cases} (x_{j_1}, x_{j_2}, I^r_{j_1}, I^r_{j_2}) & for\ j = 1, \ldots, n^*, \\ (z_j, I^r_{j_1}, I^r_{j_2}, I^k_j) & for\ j = n^* + 1, \ldots, n \end{cases}$$

where $I^r_{j_i}$ is an indicator variable equal to one if the measurement (x_{j_i} for $j = 1, \ldots, n^*$, z_j for $j = n^* + 1, \ldots, n$) corresponds to a true allele length $a_r = u + r\rho$, and zero otherwise. Similarly I^k_j is an indicator variable equal to one if the true measurements which coalesced to z_j were $z_j + t_k$ and $z_j - t_k$. With this notation, the likelihood of the complete data may be written as,

$$\begin{aligned} & L(\tilde{Y}|\psi) \\ = \quad & \textstyle\prod_{j=1}^{n^*}(\prod_{r=1}^{b}[\pi_r g_r(x_{j_1})]^{I^r_{j_1}})(\prod_{r=1}^{b}[\pi_r g_r(x_{j_2})]^{I^r_{j_2}})[1 - \delta(t,z)] \cdot \\ & \textstyle\prod_{j=n^*+1}^{n}\prod_{k=1}^{M}(\prod_{r=1}^{b}[\pi_r g_r(z_j + t_k)]^{I^r_{j_1} I^k_j})(\prod_{r=1}^{b}[\pi_r g_r(z_j - t_k)]^{I^r_{j_2} I^k_j}) \cdot \\ & [\delta(t_k, z_j)\Delta_k]^{I^k_j}. \end{aligned}$$

Taking logarithms and letting $l(\tilde{Y}|\psi)$ denote the log likelihood of the complete data yields,

$$\begin{aligned} & l(\tilde{Y}|\psi) \\ = \quad & \textstyle\sum_{j=1}^{n^*}\sum_{r=1}^{b}\{I^r_{j_1}\log\pi_r g_r(x_{j_1}) + I^r_{j_2}\log\pi_r g_r(x_{j_2}) + \log[1 - \delta(t,z)]\} + \\ & \textstyle\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b}\{I^k_j I^r_{j_1}\log\pi_r g_r(z_j + t_k) + I^k_j I^r_{j_2}\log\pi_r g_r(z_j - t_k) + \\ & I^k_j\log[\delta(t_k, z_j)\Delta_k]\}. \end{aligned}$$

Since the quantities $\log[1 - \delta(t,z)]$ and $I^k_j\log[\delta(t_k, z_j)\Delta_k]$ do not contain any of the relevant parameters (i.e. u or $\tilde{\pi}$) they will disappear during the maximization process and will not affect the maximum likelihood results. For simplicity they will henceforth be suppressed from the notation.

Taking the expected value of the log likelihood conditional on the incomplete data $\tilde{X}$ and current parameter estimates $\psi^{(m)}$ yields,

$$
\begin{aligned}
&Q(\psi|\psi^{(m)}) \\
&= \sum_{j=1}^{n^*}\sum_{r=1}^{b}\{\mathcal{E}(I^r_{j_1}|x_{j_1},\psi^{(m)})\log\pi_r g_r(x_{j_1}) + \mathcal{E}(I^r_{j_2}|x_{j_2},\psi^{(m)})\log\pi_r g_r(x_{j_1})\} + \\
&\quad \sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b}\{\mathcal{E}(I^k_j I^r_{j_1}|\psi^{(m)},z_j)\log\pi_r g_r(z_j+t_k) + \\
&\quad \mathcal{E}(I^k_j I^r_{j_2}|\psi^{(m)},z_j)\log\pi_r g_r(z_j+t_k)\} \\
&= \sum_{j=1}^{n^*}\sum_{r=1}^{b}\{W^r_{j_1}(\psi^{(m)})\log\pi_r g_r(x_{j_1}) + W^r_{j_2}(\psi^{(m)})\log\pi_r g_r(x_{j_2})\} + \\
&\quad \sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b}\{W^{rk}_{j_1}(\psi^{(m)})\log\pi_r g_r(z_j+t_k) + \\
&\quad W^{rk}_{j_2}(\psi^{(m)})\log\pi_r g_r(z_j+t_k)\} \qquad \text{(B.1)}
\end{aligned}
$$

where,

$$
\begin{aligned}
W^r_{j_s}(\psi^{(m)}) &= \mathcal{E}(I^r_{j_s}|x_{j_s},\psi^{(m)}), \quad for \quad j=1,\ldots,n^* \\
W^{rk}_{j_s}(\psi^{(m)}) &= \mathcal{E}(I^r_{j_s}I^k_j|z_j,\psi^{(m)}) \quad for \quad j=n^*+1,\ldots,n.
\end{aligned}
$$

These expected values may be evaluated as follows. For $j=1,\ldots,n^*$,

$$
\begin{aligned}
W^r_{j_s} &= \mathcal{E}(I^r_{j_s}|\psi^{(m)},x_{j_s}) \\
&= p(I^r_{j_s}=1|\psi^{(m)},x_{j_s}) \\
&= p(true\ allele\ size\ is\ a_r|\psi^{(m)},x_{j_s}) \\
&= p(a_r,x_{j_s}|\psi^{(m)})\frac{1}{p(x_{j_s}|\psi^{(m)})} \\
&= \frac{p(x_{j_s}|a_r,\psi^{(m)})p(a_r|\psi^{(m)})}{p(x_{j_s}|\psi^{(m)})} \\
&= \frac{p(x_{j_s}|a_r,\psi^{(m)})p(a_r|\psi^{(m)})}{\sum_{r=1}^{b}p(x_{j_s}|a_r,\psi^{(m)})p(a_r|\psi^{(m)})} \\
&= \frac{\pi_r^{(m)}g_r^{(m)}(x_{j_s})}{\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(x_{j_s})}. \qquad \text{(B.2)}
\end{aligned}
$$

For $j = n^* + 1, \ldots, n$,

$$\begin{aligned} W_{j_1}^{rk}(\psi^{(m)}) &= \mathcal{E}(I_{j_1}^r I_j^k | z_j, \psi^{(m)}) \\ &= p(I_{j_1}^r = 1, I_j^k = 1 | z_j, \psi^{(m)}) \\ &= p(I_{j_1}^r = 1 | z_j, \psi^{(m)}, I_j^k = 1) p(I_j^k = 1 | z_j, \psi^{(m)}) \end{aligned}$$

Now, given $I_j^k = 1$ and z_j only two possibilities exist for the measurement : either it is $z_j + t_k$ or it is $z_j - t_k$. These two events are (conditionally) equiprobable. That is,

$$p(z_j + t_k | z_j, \psi^{(m)}, I_j^k = 1) = p(z_j - t_k | z_j, \psi^{(m)}, I_j^k = 1) = \frac{1}{2}.$$

Hence by the law of total probability,

$$\begin{aligned} W_{j_1}^{rk}(\psi^{(m)}) &= \{p(I_{j_1}^r = 1 | z_j, \psi^{(m)}, I_j^k = 1, z_j + t_k) p(z_j + t_k | z_j, \psi^{(m)}, I_j^k = 1) + \\ &\quad p(I_{j_1}^r = 1 | z_j, \psi^{(m)}, I_j^k = 1, z_j - t_k) p(z_j - t_k | z_j, \psi^{(m)}, I_j^k = 1)\} \cdot \\ &\quad p(I_j^k = 1 | z_j, \psi^{(m)}) \\ &= \frac{1}{2} \frac{p(I_{j_1}^r = 1, z_j, I_j^k = 1, z_j + t_k | \psi^{(m)}) p(I_j^k = 1 | z_j, \psi^{(m)})}{p(z_j, I_j^k = 1, z_j + t_k | \psi^{(m)})} + \\ &\quad \frac{1}{2} \frac{p(I_{j_1}^r = 1, z_j, I_j^k = 1, z_j - t_k | \psi^{(m)}) p(I_j^k = 1 | z_j, \psi^{(m)})}{p(z_j, I_j^k = 1, z_j - t_k | \psi^{(m)})}. \end{aligned} \tag{B.3}$$

Now,

$$\begin{aligned} & p(z_j, I_j^k = 1, z_j + t_k | \psi^{(m)}) \\ &= [\sum_{r=1}^{b} \pi_r^{(m)} g_r^{(m)}(z_j + t_k)][\sum_{r=1}^{b} \pi_r^{(m)} g_r^{(m)}(z_j - t_k)] \delta(t_k, z_j) \Delta_k \\ &= p(z_j, I_j^k = 1, z_j + t_k | \psi^{(m)}), \end{aligned}$$

$$\begin{aligned} & p(I_j^k = 1 | z_j, \psi^{(m)}) \\ &= \frac{p(I_j^k = 1, z_j | \psi^{(m)})}{p(z_j | \psi^{(m)})} \\ &= \frac{[\sum_{r=1}^{b} \pi_r^{(m)} g_r^{(m)}(z_j + t_k)][\sum_{r=1}^{b} \pi_r^{(m)} g_r^{(m)}(z_j - t_k)] \delta(t_k, z_j) \Delta_k}{\sum_{k=1}^{M} [[\sum_{r=1}^{b} \pi_r^{(m)} g_r^{(m)}(z_j + t_k)][\sum_{r=1}^{b} \pi_r^{(m)} g_r^{(m)}(z_j - t_k)] \delta(t_k, z_j) \Delta_k}, \end{aligned}$$

and,

$$P(I^r_{j_i} = 1, z_j, I^k_j = 1, z_j \pm t_k | \psi^{(m)})$$

$$= [\pi^{(m)}_r g^{(m)}_r(z_j \pm t_k)][\sum_{r=1}^{b} \pi^{(m)}_r g^{(m)}_r(z_j - t_k)]\delta(t_k, z_j)\Delta_k.$$

Substituting these quantities into B.3 yields after simplification,

$$W^{rk}_{j_i}(\psi^{(m)}) = \frac{1}{2}\Big[\frac{\pi_r g_r(z_j + t_k)][\sum_{r=1}^{b} \pi_r g_r(z_j - t_k)]\delta(t_k, z_j)\Delta_k}{\pi_r g_r(z_j + t_k)][\sum_{r=1}^{b} \pi_r g_r(z_j - t_k)]\delta(t_k, z_j)\Delta_k} +$$

$$\frac{t_k)][\sum_{r=1}^{b} \pi_r g_r(z_j + t_k)]\delta(t_k, z_j)\Delta_k}{z_j - t_k)][\sum_{r=1}^{b} \pi_r g_r(z_j + t_k)]\delta(t_k, z_j)\Delta_k}. \quad \text{(B.1)}$$

Three important features of B.2 and B.4 are,

(i) $\sum_{r=1}^{b} W^r_{j_i}(\psi^{(m)}) = 1$,

(ii) $\sum_{k=1}^{M}\sum_{r=1}^{b} W^{rk}_{j_i}(\psi^{(m)}) = 1$, and,

(iii) $W^{rk}_{j1}(\psi^{(m)}) = W^{rk}_{j2}(\psi^{(m)})$ The common value of $W^{rk}_{j_i}(\psi^{(m)})$ for $i = 1, 2$ will be denoted $W^{rk}_j(\psi^{(m)})$.

Estimating $\tilde{\pi}$

Differentiating $Q(\psi|\psi^{(m)})$ (see equation B.1) with respect to π_r subject to the constraint that $\sum_{r=1}^{b} \pi_r = 1$ yields,

$$\frac{\partial Q(\psi|\psi^{(m)})}{\partial \pi_r} = \sum_{j=1}^{n^*}\{\frac{W^r_{j1}(\psi^{(m)})}{\pi_r} + \frac{W^r_{j2}(\psi^{(m)})}{\pi_r}\} +$$

$$2\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\frac{W^{rk}_j(\psi^{(m)})}{\pi_r} - \lambda. \quad \text{(B.5)}$$

Setting B.5 equal to zero and solving for π_r using the fact that $\sum_{r=1}^{b} W_{j_1}^{r}(\psi^{(m)}) = 1$ and $\sum_{k=1}^{M}\sum_{r=1}^{b} W_{j}^{rk}(\psi^{(m)}) = 1$ yields, after simplification,

$$\pi_r^{(m+1)} = \frac{1}{2n}\Big[\sum_{j=1}^{n^*}\frac{\pi_r^m g_r^m(x_{j_1})}{\sum_{r=1}^{b}\pi_r^m g_r^m(x_{j_1})} + \frac{\pi_r^m g_r^m(x_{j_2})}{\sum_{r=1}^{b}\pi_r^m g_r^m(x_{j_2})}\Big] +$$

$$\frac{1}{2n}\sum_{j=n^*+1}^{n}\Big[\frac{\sum_{k=1}^{M}\pi_r^m g_r^m(z_j+t_k)\sum_{r=1}^{b}\pi_r^m g_r^m(z_j-t_k)\delta(t_k,z_j)\Delta_k}{\sum_{k=1}^{M}(\sum_{r=1}^{b}\pi_r^m g_r^m(z_j+t_k))(\sum_{r=1}^{b}\pi_r^m g_r^m(z_j-t_k))} +$$

$$\frac{\sum_{k=1}^{M}\pi_r^m g_r^m(z_j-t_k)\sum_{r=1}^{b}\pi_r^m g_r^m(z_j+t_k)\delta(t_k,z_j)\Delta_k}{\sum_{k=1}^{M}(\sum_{r=1}^{b}\pi_r^m g_r^m(z_j+t_k))(\sum_{r=1}^{b}\pi_r^m g_r^m(z_j-t_k))}\Big]. \qquad \text{(B.6)}$$

This is identical to the iterative equation given in Devlin, Risch, and Roeder (1991).

Estimating u

Differentiating $Q(\psi|\psi^{(m)})$ (see equation B.1) with respect to u yields,

$$\frac{\partial Q(\psi|\psi^{(m)})}{\partial u} = \sum_{j=1}^{n^*}\sum_{r=1}^{b}\{W_{j_1}^{r}(\psi^{(m)})\frac{\partial \log g_r(x_{j_1})}{\partial u} + W_{j_2}^{r}(\psi^{(m)})\frac{\partial \log g_r(x_{j_2})}{\partial u}\} +$$

$$\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b}\{W_j^{rk}(\psi^{(m)})\frac{\partial \log g_r(z_j+t_k)}{\partial u} +$$

$$W_j^{rk}(\psi^{(m)})\frac{\partial \log g_r(z_j-t_k)}{\partial u}\}.$$

This is a complicated expression since u appears in both the mean $(u + r\rho)$ and standard deviation $(\sigma_r = (5\text{x}10^{-3})(u + r\rho))$ of $g_r(\cdot)$. The procedure was carried out using two different simplifying assumptions. Either,

(i) assume that the standard deviation of the measurement error is a constant (i.e. $\sigma_r = \sigma$, $r = 1, \ldots, b$), or,

(ii) assume that σ_r may be replaced by $\sigma_r^{(m)} = (5\text{x}10^{-3})(u^{(m)} + r\rho)$ at each step of the algorithm.

The first assumption, although much stronger, results in a simple iterative scheme for estimating u. Under this assumption, the derivative of $Q(\psi|\psi^{(m)})$ with respect to u becomes,

$$\frac{\partial Q(\psi|\psi^{(m)})}{\partial u} = \sum_{j=1}^{n^*}\sum_{r=1}^{b}\{W_{j_1}^{r}(\psi^{(m)})\frac{(x_{j_1}-u-r\rho)}{\sigma^2}+W_{j_2}^{r}(\psi^{(m)})\frac{(x_{j_2}-u-r\rho)}{\sigma^2}\}+$$

$$\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b}\{W_{j}^{rk}(\psi^{(m)})\frac{(z_j+t_k-u-r\rho)}{\sigma^2}+$$

$$W_{j}^{rk}(\psi^{(m)})\frac{(z_j-t_k-u-r\rho)}{\sigma^2}\}. \qquad (B.7)$$

Setting B.7 equal to zero and solving for u using the fact that $\sum_{r=1}^{b} W_{j_i}^{r}(\psi^{(m)}) = 1$ and $\sum_{k=1}^{M}\sum_{r=1}^{b} W_{j}^{rk}(\psi^{(m)}) = 1$ yields, after simplification,

$$u^{(m+1)} =$$

$$\frac{1}{2n}\sum_{j=1}^{n^*}\sum_{r=1}^{b}[\frac{\pi_r^{(m)}g_r^{(m)}(x_{j_1})}{\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(x_{j_1})}(x_{j_1}-r\rho)+\frac{\pi_r^{(m)}g_r^{(m)}(x_{j_2})}{\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(x_{j_2})}(x_{j_2}-r\rho)]+$$

$$\frac{1}{2n}\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b}[\frac{[\pi_r^{(m)}g_r^{(m)}(z_j+t_k)][\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(z_j-t_k)]\delta(t_k,z_j)\Delta_k(z_j-r\rho)}{\sum_{k=1}^{M}[[\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(z_j+t_k)][\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(z_j-t_k)]\delta(t_k,z_j)\Delta_k}+$$

$$\frac{[\pi_r^{(m)}g_r^{(m)}(z_j-t_k)][\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(z_j+t_k)]\delta(t_k,z_j)\Delta_k(z_j-r\rho)}{\sum_{k=1}^{M}[[\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(z_j+t_k)][\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(z_j-t_k)]\delta(t_k,z_j)\Delta_k}].$$

Using assumption (ii), although more realistic, yields a more complicated iterative equation for u of the form,

$$u^{(m+1)} =$$

$$\frac{1}{D(\tilde{X})}\sum_{j=1}^{n^*}\sum_{r=1}^{b}[\frac{\pi_r^{(m)}g_r^{(m)}(x_{j_1})}{\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(x_{j_1})}\frac{(x_{j_1}-r\rho)}{[\sigma_r^{(m)}]^2}+\frac{\pi_r^{(m)}g_r^{(m)}(x_{j_2})}{\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(x_{j_2})}\frac{(x_{j_2}-r\rho)}{[\sigma_r^{(m)}]^2}]+$$

$$\frac{1}{D(\tilde{X})}\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b}[\frac{[\pi_r^{(m)}g_r^{(m)}(z_j+t_k)][\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(z_j-t_k)]\delta(t_k,z_j)\Delta_k\frac{(z_j-r\rho)}{[\sigma_r^{(m)}]^2}}{\sum_{k=1}^{M}[[\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(z_j+t_k)][\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(z_j-t_k)]\delta(t_k,z_j)\Delta_k}+$$

$$\frac{[\pi_r^{(m)}g_r^{(m)}(z_j-t_k)][\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(z_j+t_k)]\delta(t_k,z_j)\Delta_k\frac{(z_j-r\rho)}{[\sigma_r^{(m)}]^2}}{\sum_{k=1}^{M}[[\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(z_j+t_k)][\sum_{r=1}^{b}\pi_r^{(m)}g_r^{(m)}(z_j-t_k)]\delta(t_k,z_j)\Delta_k}]$$

where,

$$D(\tilde{X}) = \sum_{j=1}^{n^*} \sum_{r=1}^{b} \frac{W_{j_1}^r(\psi^{(m)})}{[\sigma_r^{(m)}]^2} + \frac{W_{j_2}^r(\psi^{(m)})}{[\sigma_r^{(m)}]^2} + 2\sum_{j=n^*+1}^{n} \sum_{r=1}^{b} \sum_{k=1}^{M} \frac{W_j^{rk}(\psi^{(m)})}{[\sigma_r^{(m)}]^2}.$$

These results for u are similar, but not identical, to those of Devlin, Risch, and Roeder (1991) which may have involved further approximations or a slightly different approach.

B.2 Multiple Flanking Region Case

In the case of multiple flanking regions, the incomplete data $\tilde{X}$ needs to be augmented by three indicator variables. The complete data is given by,

$$y_j = \begin{cases} (x_{j_1}, x_{j_2}, I_{j_1}^r, I_{j_2}^r, I_{j_1}^l, I_{j_2}^l) & for\ j = 1, \ldots, n^*, \\ (z_j, I_{j_1}^r, I_{j_2}^r, I_j^k, I_{j_1}^l, I_{j_2}^l) & for\ j = n^* + 1, \ldots, n \end{cases}$$

where $I_{j_i}^r$ is an indicator variable equal to one if the measurement (x_{j_i} for $j = 1, \ldots, n^*$, z_j for $j = n^* + 1, \ldots, n$) corresponds to a true allele length $a_r = u + r\rho$, and zero otherwise. I_j^k is an indicator variable equal to one if the true measurements which coalesced to z_j were $z_j + t_k$ and $z_j - t_k$. $I_{j_i}^l$ is an indicator variable equal to one if the flanking region size is u_l and zero otherwise.

From 5.15 the likelihood of the incomplete data may be written as,

$$\begin{aligned} & L(\tilde{X}|\Psi) \\ \approx\ & \prod_{j=1}^{n^*} (\sum_{l=1}^{L} \phi_l \sum_{r=1}^{b} \pi_r g_r(x_{j_1}))(\sum_{l=1}^{L} \phi_l \sum_{r=1}^{b} \pi_r g_r(x_{j_2}))[1 - \delta(t,z)] \cdot \\ & \prod_{j=n^*+1}^{n} \sum_{k=1}^{M} \{(\sum_{l=1}^{L} \phi_l \sum_{r=1}^{b} \pi_r g_r(z_j + t_k)) \cdot \\ & (\sum_{l=1}^{L} \phi_l \sum_{r=1}^{b} \pi_r g_r(z_j - t_k))\delta(t_k, z_j)\Delta_k \end{aligned}$$

where $\Psi = (\tilde{u}, \tilde{\pi}, \tilde{\phi})^T$. The likelihood of the complete data is thus,

$$\begin{aligned} & L(\tilde{Y}|\psi) \\ =\ & \prod_{j=1}^{n^*} (\prod_{l=1}^{L} \prod_{r=1}^{b} [\phi_l \pi_r g_r(x_{j_1})]^{I_{j_1}^r I_{j_1}^l})(\prod_{l=1}^{L} \prod_{r=1}^{b} [\phi_l \pi_r g_r(x_{j_2})]^{I_{j_2}^r I_{j_2}^l})[1 - \delta(t,z)] \cdot \\ & \prod_{j=n^*+1}^{n} \prod_{k=1}^{M} (\prod_{l=1}^{L} \prod_{r=1}^{b} [\phi_l \pi_r g_r(z_j + t_k)]^{I_{j_1}^r I_j^k I_{j_1}^l}) \cdot \\ & (\prod_{l=1}^{L} \prod_{r=1}^{b} [\phi_l \pi_r g_r(z_j - t_k)]^{I_{j_2}^r I_j^k I_{j_2}^l})[\delta(t_k, z_j)\Delta_k]^{I_j^k}. \end{aligned}$$

Taking logarithms and letting $l(\tilde{Y}|\psi)$ denote the log likelihood of the complete data yields,

$$
\begin{aligned}
& l(\tilde{Y}|\psi) \\
&= \sum_{j=1}^{n^*}\sum_{l=1}^{L}\sum_{r=1}^{b}\{I_{j_1}^r I_{j_1}^l \log\phi_l\pi_r g_r(x_{j_1}) + I_{j_2}^r I_{j_2}^l \log\phi_l\pi_r g_r(x_{j_2}) + \log[1-\delta(t,z)]\} + \\
&\quad \sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{l=1}^{L}\sum_{r=1}^{b}\{I_j^k I_{j_1}^r I_{j_1}^l \log\phi_l\pi_r g_r(z_j+t_k) + I_j^k I_{j_2}^r I_{j_2}^l \log\phi_l\pi_r g_r(z_j-t_k) + \\
&\quad I_j^k \log[\delta(t_k,z_j)\Delta_k]\}.
\end{aligned}
$$

Since the quantities $\log[1-\delta(t,z)]$ and $I_j^k \log[\delta(t_k,z_j)\Delta_k]$ do not contain any of the relevant parameters (i.e. $\tilde{u}$, $\tilde{\pi}$ or $\tilde{\phi}$) they will disappear during the maximization process and will not affect the maximum likelihood results. For simplicity they will henceforth be suppressed from the notation.

Taking the expected value of the log likelihood conditional on the incomplete data $\tilde{X}$ and current parameter estimates $\psi^{(m)}$ yields,

$$
\begin{aligned}
& Q(\psi|\psi^{(m)}) \\
&= \sum_{j=1}^{n^*}\sum_{l=1}^{L}\sum_{r=1}^{b}\{\mathcal{E}(I_{j_1}^r I_{j_1}^l|x_{j_1},\psi^{(m)})\log\phi_l\pi_r g_r(x_{j_1}) + \\
&\quad \mathcal{E}(I_{j_2}^r I_{j_1}^l|x_{j_2},\psi^{(m)})\log\phi_l\pi_r g_r(x_{j_1})\} + \\
&\quad \sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{l=1}^{L}\sum_{r=1}^{b}\{\mathcal{E}(I_j^k I_{j_1}^r I_{j_1}^l|\psi^{(m)},z_j)\log\phi_l\pi_r g_r(z_j+t_k) + \\
&\quad \mathcal{E}(I_j^k I_{j_2}^r I_{j_1}^l|\psi^{(m)},z_j)\log\phi_l\pi_r g_r(z_j+t_k)\} \\
&= \sum_{j=1}^{n^*}\sum_{l=1}^{L}\sum_{r=1}^{b}\{W_{j_1}^{rl}(\psi^{(m)})\log\phi_l\pi_r g_r(x_{j_1}) + W_{j_2}^{rl}(\psi^{(m)})\log\phi_l\pi_r g_r(x_{j_2})\} + \\
&\quad \sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{l=1}^{L}\sum_{r=1}^{b}\{W_{j_1}^{rlk}(\psi^{(m)})\log\phi_l\pi_r g_r(z_j+t_k) + \\
&\quad W_{j_2}^{rlk}(\psi^{(m)})\log\phi_l\pi_r g_r(z_j+t_k)\}. \qquad \text{(B.8)}
\end{aligned}
$$

Analogous to the single flanking region case the expected values may be evaluated

quite easily. For $j = 1, \ldots, n^*$,

$$\begin{aligned} W_{j_1}^{rl}(\psi^{(m)}) &= \mathcal{E}(I_{j_1}^r I_{j_1}^l | \psi^{(m)}, x_{j_1}) \\ &= p(I_{j_1}^r = 1, I_{j_1}^l = 1 | \psi^{(m)}, x_{j_1}) \\ &= \frac{p(I_{j_1}^r = 1, I_{j_1}^l = 1, x_{j_1} | \psi^{(m)})}{p(x_{j_1} | \psi^{(m)})} \\ &= \frac{\phi_l \pi_r^{(m)} g_r^{(m)}(x_{j_1})}{\sum_{l=1}^{L} \sum_{r=1}^{b} \phi_l \pi_r^{(m)} g_r^{(m)}(x_{j_1})}. \end{aligned} \tag{B.9}$$

Similarly, for $j = n^* + 1, \ldots, n$,

$$\begin{aligned} & W_{j_1}^{rlk}(\psi^{(m)}) \\ &= \mathcal{E}(I_{j_1}^r I_{j_1}^l I_{j_1}^k | \psi^{(m)}, z_j) \\ &= \frac{p(I_{j_1}^r = 1, I_{j_1}^l = 1, I_{j_1}^k, z_j | \psi^{(m)})}{p(z_j | \psi^{(m)})} \\ &= \frac{\frac{1}{2}[\phi_l^{(m)} \pi_r^{(m)} g_r^{(m)}(z_j + t_k)][\sum_{l=1}^{L} \sum_{r=1}^{b} \phi_l^{(m)} \pi_r^{(m)} g_r^{(m)}(z_j - t_k)] \delta(t_k, z_j) \Delta_k}{\sum_{k=1}^{M} [[\sum_{l=1}^{L} \sum_{r=1}^{b} \phi_l^{(m)} \pi_r^{(m)} g_r^{(m)}(z_j + t_k)][\sum_{l=1}^{L} \sum_{r=1}^{b} \phi_l^{(m)} \pi_r^{(m)} g_r^{(m)}(z_j - t_k)] \delta(t_k, z_j) \Delta_k} + \\ & \frac{\frac{1}{2}[\phi_l^{(m)} \pi_r^{(m)} g_r^{(m)}(z_j - t_k)][\sum_{l=1}^{L} \sum_{r=1}^{b} \phi_l^{(m)} \pi_r^{(m)} g_r^{(m)}(z_j + t_k)] \delta(t_k, z_j) \Delta_k}{\sum_{k=1}^{M} [[\sum_{l=1}^{L} \sum_{r=1}^{b} \phi_l^{(m)} \pi_r^{(m)} g_r^{(m)}(z_j - t_k)][\sum_{l=1}^{L} \sum_{r=1}^{b} \phi_l^{(m)} \pi_r^{(m)} g_r^{(m)}(z_j + t_k)] \delta(t_k, z_j) \Delta_k}. \end{aligned}$$

Three important features are,

(i) $\sum_{l=1}^{L} \sum_{r=1}^{b} W_{j_1}^{rl}(\psi^{(m)}) = 1$,

(ii) $W_{j_1}^{rlk}(\psi^{(m)}) = W_{j_2}^{rlk}(\psi^{(m)})$, and,

(iii) $\sum_{k=1}^{M} \sum_{l=1}^{L} \sum_{r=1}^{b} W_j^{rlk}(\psi^{(m)}) = 1$.

Estimating $\tilde{\phi}$ and $\tilde{\pi}$

Differentiating $Q(\psi|\psi^{(m)})$ (see equation B.8) with respect to ϕ_l subject to the constraint that $\sum_{l=1}^{L} = 1$ yields,

$$\frac{\partial Q(\psi|\psi^{(m)})}{\partial \phi_l} = \sum_{j=1}^{n^*}\sum_{r=1}^{b}\{\frac{W_{j_1}^{rl}(\psi^{(m)})}{\phi_l} + \frac{W_{j_2}^{rl}(\psi^{(m)})}{\phi_l}\} + 2\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b}\frac{W_j^{rlk}(\psi^{(m)})}{\phi_l} - \lambda. \quad (B.10)$$

Setting B.10 equal to zero and solving for ϕ_l yields after simplification,

$$\phi_l^{(m+1)} = \frac{1}{2n}\sum_{j=1}^{n^*}\sum_{r=1}^{b} W_{j_1}^{rl}(\psi^{(m)}) + W_{j_2}^{rl}(\psi^{(m)}) + \frac{1}{n}\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b} W_j^{rlk}(\psi^{(m)}).$$

In an analogous way one obtains for π_r the iterative equation,

$$\pi_r^{(m+1)} = \frac{1}{2n}\sum_{j=1}^{n^*}\sum_{l=1}^{L} W_{j_1}^{rl}(\psi^{(m)}) + W_{j_2}^{rl}(\psi^{(m)}) + \frac{1}{n}\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{l=1}^{L} W_j^{rlk}(\psi^{(m)}).$$

Estimating $\tilde{u}$

In order to simplify the problem (as in the single flanking region case) it will be assumed that σ_r may be replaced by $\sigma_r^{(m)}$. The stronger assumption that σ_r is a constant for all r does not result in as much simplification in the multiple flanking region case as it did in the single flanking region case, so it will be avoided. With this convention, differentiating $Q(\psi|\psi^{(m)})$ (see equation B.8) with respect to u_l one obtains,

$$\frac{\partial Q(\psi|\psi^{(m)})}{\partial u_l} = \sum_{j=1}^{n^*}\sum_{r=1}^{b}\{\frac{W_{j1}^{rl}(\psi^{(m)})}{[\sigma_r^{(m)}]^2}(x_{j1} - u_l - r\rho) + \frac{W_{j2}^{rl}(\psi^{(m)})}{[\sigma_r^{(m)}]^2}((x_{j2} - u_l - r\rho)\} +$$

$$2\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b}\frac{W_j^{rlk}(\psi^{(m)})}{[\sigma_r^{(m)}]^2}(z_j - u_l - r\rho). \qquad \text{(B.11)}$$

Setting B.11 equal to zero and solving for u_l yields,

$$u_l^{(m+1)} =$$

$$\frac{\sum_{j=1}^{n^*}\sum_{r=1}^{b}\{\frac{W_{j1}^{rl}(\psi^{(m)})}{[\sigma_r^{(m)}]^2}(x_{j1} - r\rho) + \frac{W_{j2}^{rl}(\psi^{(m)})}{[\sigma_r^{(m)}]^2}((x_{j2} - r\rho)\}}{\sum_{j=1}^{n^*}\sum_{r=1}^{b}\{\frac{W_{j1}^{rl}(\psi^{(m)})}{[\sigma_r^{(m)}]^2} + \frac{W_{j2}^{rl}(\psi^{(m)})}{[\sigma_r^{(m)}]^2}\} + 2\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b}\frac{W_j^{rlk}(\psi^{(m)})}{[\sigma_r^{(m)}]^2}} +$$

$$\frac{2\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b}\frac{W_j^{rlk}(\psi^{(m)})}{[\sigma_r^{(m)}]^2}(z_j - r\rho)}{\sum_{j=1}^{n^*}\sum_{r=1}^{b}\{\frac{W_{j1}^{rl}(\psi^{(m)})}{[\sigma_r^{(m)}]^2} + \frac{W_{j2}^{rl}(\psi^{(m)})}{[\sigma_r^{(m)}]^2}\} + 2\sum_{j=n^*+1}^{n}\sum_{k=1}^{M}\sum_{r=1}^{b}\frac{W_j^{rlk}(\psi^{(m)})}{[\sigma_r^{(m)}]^2}}.$$

Values for $W_j^{rlk}(\psi^{(m)})$ and $W_{j1}^{rl}(\psi^{(m)})$ may be substituted in. Again these equations are similar but not identical to those obtained in Devlin, Risch, and Roeder (1991).

Bibliography

[1] Baird, M., Balazs, I., Giusti, A., Miyazaki, L., Nicholas, L., Wexler, K., Kanter, E., Glassberg, J., Allen, F., Rubinstein, P., Sussman, L. Allele Frequency Distribution of Two Highly Polymorphic DNA Sequences in Three Ethnic Groups and Its Application to the Determination of Paternity. *American Journal of Human Genetics* 39: 489-501, 1986.

[2] Balazs, I., Baird, M., Clyne, M., Meade, E. Human Population Genetic Studies of Five Hypervariable DNA Loci. *American Journal of Human Genetics* 44: 182-190, 1989.

[3] Berry, D.A. Inferences Using DNA Profiling in Forensic Identification and Paternity Cases. *Statistical Science* 6: 175-205, 1991.

[4] Berry, D.A., Evett, I.W., Pinchin, R. Statistical Inference in Crime Investigations using Deoxyribonucleic Acid Profiling. *Applied Statistics* 41: 499-531, 1992.

[5] Budowle, B., Giusti, A.M., Waye, J.S., Baechtel, F.S., Fourney, R.M., Adams, D.E., Presley, L.A., Deadman, H.A., Monson, K.L. Fixed-Bin Analysis for Statistical Evaluation of Continuous Distributions of Allelic Data from VNTR Loci, for Use in Forensic Comparisons. *American Journal of Human Genetics* 48: 841-855, 1991.

[6] Caskey, C.T. Comments on DNA-based Forensic Analysis. *American Journal of Human Genetics* 49: 893-895, 1991.

[7] Chakraborty, R., Kidd. K. The Utility of DNA Typing in Forensic Work. *Science* 254: 1735-1739, 1991.

[8] Cohen, J.E. DNA Fingerprinting for Forensic Identification : Potential Effects on Data Interpretation of Subpopulation Heterogeneity and Band Number Variability. *American Journal of Human Genetics* 46: 358-368, 1990.

[9] Debenham, P.G. DNA Fingerprinting. *Journal of Pathology* 164: 101-106, 1991.

[10] Dempster, A.P., Laird, N.M., Rubin, D.B. Maximum Likelihood From Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B* 39: 1-38, 1977.

[11] Devlin, B., Risch, N., Roeder, K. No Excess of Homozygosity at Loci Used for DNA Fingerprinting. *Science* 249: 1416-1419, 1990.

[12] Devlin, B., Risch, N., Roeder, K. Estimation of Allele Frequencies for VNTR Loci. *American Journal of Human Genetics* 48: 662-676, 1991.

[13] Devlin, B., Risch, N., Roeder, K. Forensic Inference From DNA Fingerprints. *Journal of the American Statistical Association* 87: 337-350, 1992.

[14] Devlin, B., Risch, N., Roeder, K. Statistical Evaluation of DNA Fingerprinting : A Critique of the NRC's Report. *Science* 259: 1993.

[15] Efron, B. Nonparametric estimates of standard error : The jackknife, the bootstrap and other methods. *Biometrika* 68: 589-599, 1981.

[16] Efron, B., Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science* 1: 54-77, 1986.

[17] Ellman, I.M., Kaye, D.H. Probabilities and Proof: Can HLA and blood group testing prove paternity? *New York University Law Review* 54: 1131-1162, 1979.

[18] Geisser, S., Johnson, W. Testing Independence of Fragment Lengths Within VNTR Loci. *American Journal of Human Genetics* 53: 1103-1106, 1993.

[19] Goodman, L.A. On Simultaneous Confidence Intervals for Multinomial Proportions. *Technometrics* 7: 247-254, 1965.

[20] Hall, P. Theoretical Comparison of Bootstrap Confidence Intervals. *The Annals of Statistics* 16: 927-953, 1988.

[21] Hernández, J.L., Weir, B.S. A Disequilibrium Coefficient Approach to Hardy-Weinberg Testing. *Biometrics* 45: 53-70, 1989.

[22] Kirby, L.T. DNA Fingerprinting An Introduction. New York, NY : Stockton Press ; 1990.

[23] Kullback, S., Leibler, R.A. On Information and Sufficiency. *Annals of Mathematical Statistics* 22: 79-86, 1951.

[24] Lander, E. S. Lander Reply. *American Journal of Human Genetics* 49: 899-903, 1991.

[25] Lehman, E.L. Theory of Point Estimation. John Wiley and Sons, 1983.

[26] Lewontin R., Hartl, D. Population Genetics in Forensic DNA Typing. *Science* 254: 1745-1750, 1991.

[27] Mendell, N.R., Simon, G.A. A general expression for the variance-covariance matrix of estimates of gene frequency : the effects of departures from Hardy-Weinberg Equilibrium. *Annals of Human Genetics* 48: 283-286, 1984.

[28] National Research Council. DNA Technology in Forensic Science. Washington, D.C. : National Academy Press ; 1992.

[29] Nichols, R.A., Balding, D.J. Effects of Population Structure on DNA Fingerprint Analysis in Forensic Science. *Heredity* 66: 297-302, 1991.

[30] Quesenberry, C.P., Hurst, D.C. Large Sample Simultaneous Confidence Intervals for Multinomial Proportions. *Technometrics* 6: 191-195, 1964.

[31] Rao, C.R. Linear Statistical Inference and its Applications. Wiley, 1965.

[32] Risch, N., Devlin, B. On the Probability of Matching DNA Fingerprints. *Science* 255: 717-720, 1992.

[33] Robbins, H.E. Estimating the Total Probability of the Unobserved Outcome of an Experiment. *Annals of Mathematical Statistics* 39: 256-257, 1968.

[34] Rothwell, N.V., Understanding Genetics. New York, NY : Oxford University Press ; 1988.

[35] Weir, B.S. Inferences About Linkage Disequilibrium. *Biometrics* 35: 235-254, 1979.

[36] Weir, B.S. Independence of VNTR Alleles Defined as Fixed Bins. *Genetics* 130: 873-887, 1992.

[37] Weir, B.S. Independence of VNTR Alleles Defined as Floating Bins. *American Journal of Human Genetics* 51: 992-997, 1992.

[38] Weir, B.S. Independence Tests for VNTR Alleles Defined as Quantile Bins. *American Journal of Human Genetics* 53: 1107-1113, 1993.